Taller 2: Modelos Estocásticos, Análisis de Datos y Ajuste de Modelos de Probabilidad

Objetivo:

Desarrollar la comprensión y aplicación de modelos estocásticos y probabilísticos, mediante el estudio de Cadenas de Markov y simulaciones Monte Carlo, así como fortalecer habilidades en el pre-procesamiento y análisis exploratorio de datos. Además, identificar y ajustar distribuciones de probabilidad apropiadas para datos reales mediante métodos gráficos, pruebas de hipótesis y análisis de bondad de ajuste.

Instrucciones:

1. Lea cada una de las partes del taller y realice los ejercicios planteados.
2. Utilice Python para implementar las soluciones numéricas y analice los resultados.
3. Responda las preguntas conceptuales de manera argumentada y soportado en soluciones con modelos computacionales como en Python, en el caso de requerirse.

Parte 1: Pre-procesamiento y análisis exploratorio de datos (Cuarta Sesión)

1. **Análisis y Pre-procesamiento de Datos Meteorológicos de la RMCAB (2016–2023)**:

   El objetivo de este punto es aplicar técnicas de **pre-procesamiento de datos**, **análisis exploratorio** y **ajuste de tendencias** para preparar variables meteorológicas que serán utilizadas posteriormente en un ejercicio de simulación mediante el **método de Monte Carlo**.

   Se ha puesto a disposición de los estudiantes el histórico de datos meteorológicos horarios de la **Red de Monitoreo de Calidad del Aire de Bogotá (RMCAB)**, correspondiente al período **2016 a 2023**, recopilado en las **19 estaciones** de monitoreo.

   La información se encuentra distribuida en tres archivos en formato `Excel`, cada uno compuesto por una única hoja que contiene los datos de todas las estaciones. Debido al gran tamaño de los archivos, no se recomienda abrirlos directamente con Excel. Para conocer la estructura de los datos, se sugiere revisar el archivo `4 Datos RMCAB Sep 2023.xlsx`, que contiene únicamente un mes de información y presenta el mismo formato que los archivos principales.

   Los archivos proporcionados incluyen, entre otras variables, registros de **precipitación** y **temperatura**.

   a. Carga y Exploración Inicial de Datos

   - Cargar los tres archivos de datos horarios disponibles, siguiendo las instrucciones del cuaderno de Colab `pre-procesamiento_RMCABdata`.
   - Realizar una primera exploración:
     - ¿Cuántos registros conforman el conjunto total?

  - ¿Qué variables meteorológicas están disponibles?
  - ¿Todas las estaciones cuentan con registros completos para el periodo 2016–2023?

b. Selección de Variables de Análisis

- Focalizarse en las variables de **Precipitación** y **Temperatura**.
- Identificar los nombres de las columnas correspondientes en los archivos.
- Calcular el promedio de las mediciones para toda la ciudad, utilizando los datos de las 19 estaciones.

c. Pre-procesamiento de Datos

- Identificar y gestionar valores faltantes:
  - Celdas vacías, datos nulos o errores evidentes.
  - La limpieza inicial de los datos será realizada automáticamente mediante el script del cuaderno de Colab adjunto.
- Detectar y justificar el tratamiento de valores atípicos (*outliers*) en las series de precipitación y temperatura.

d. Análisis Exploratorio de Datos

- Generar visualizaciones para cada variable:
  - Histogramas.
  - Boxplots.
  - Series de tiempo.
- Calcular estadísticos descriptivos relevantes:
  - Media, mediana, percentiles, desviación estándar, entre otros.
  - Identificar periodos atípicos de precipitación, correspondientes a niveles muy altos o muy bajos.

e. Preparación de Variable para Simulación Monte Carlo

- Proponer una forma de modelar la variable **precipitación** para su posterior uso en simulaciones Monte Carlo:
  - Ajuste a una distribución de probabilidad (normal, gamma, lognormal, etc.).
  - Definición de parámetros estadísticos como la media y la varianza.
- Justificar la elección de la distribución o del método propuesto.

Notas Importantes

- Se deben utilizar los datos de las **19 estaciones**.
- Es obligatorio documentar claramente todas las decisiones de pre-procesamiento y modelamiento adoptadas.

Parte 2: Aplicación de Modelos Estocásticos — Monte Carlo y Cadenas de Markov (Tercera Sesión)

1. **Simulación Método Monte Carlo - Pronóstico de Sostenibilidad Hídrica Aplicación del modelo con datos reales de Bogotá:** Utilizando información pública sobre la oferta y demanda de agua para la ciudad de Bogotá, adapte el modelo de simulación Monte Carlo para evaluar la sostenibilidad del recurso hídrico durante los próximos 20 años.

   Datos sugeridos:

   - Área de la cuenca abastecedora (Chingaza, Sumapaz, Tunjuelo): aproximadamente **1.200 km²**.
   - Eficiencia hídrica (escorrentía estimada): valor promedio de **0.25**.
   - Precipitación media anual: entre **1.000 y 1.200 mm/año**.
   - Demanda actual: alrededor de **18 m³/s**, equivalente a **567 millones de m³/año**.
   - Crecimiento de demanda: suponer un **1.5% anual**.

   Preguntas:

   (a) Ajusta los parámetros del modelo con los valores anteriores.

   (b) Introduce una tendencia de **crecimiento de la demanda** año a año.

   (c) Simula al menos **1.000 trayectorias de 20 años**.

   (d) Grafica el comportamiento del almacenamiento y la **probabilidad de agotamiento** del recurso.

   (e) Discute los resultados y plantea estrategias de gestión como:
   - Reforestación en cuencas (incrementar eficiencia).
   - Reducción del consumo.
   - Mayor captación de aguas lluvias.

2. **Simulación del Cadenas de Markov - Pronosticos de Sismos:** En este ejercicio aplicaremos conceptos de cadenas de Markov para analizar el comportamiento espacial de los sismos históricos en Colombia. A partir de la ubicación geográfica (latitud y longitud) de cada sismo, construiremos una matriz de frecuencias de transición entre zonas, la cual permitirá estimar probabilidades de ocurrencia a largo plazo. Este análisis permitirá reforzar el entendimiento de procesos estocásticos aplicados a fenómenos naturales.

   - Acceda desde este link al archivo de datos de sismicidad histórica de Colombia ubicado en el GitHub

   - **Clasificación geográfica de los sismos:**
     Cada sismo registrado debe ser clasificado en uno de los **cuatro cuadrantes** establecidos, usando como referencia el siguiente **punto de intersección**:
     - Latitud de corte: `4.028369`
     - Longitud de corte: `-73.610852`

     La clasificación de cuadrantes se realiza aplicando las siguientes condiciones:

Cuadrante	Condición lógica
NO (Noroeste)	Latitud ¿ 4.028369 y Longitud ¡ -73.610852
NE (Noreste)	Latitud ¿ 4.028369 y Longitud $\geq$ -73.610852
SO (Suroeste)	Latitud $\leq$ 4.028369 y Longitud ¡ -73.610852
SE (Sureste)	Latitud $\leq$ 4.028369 y Longitud $\geq$ -73.610852

Esta clasificación permite organizar espacialmente los sismos para identificar patrones de transición entre zonas.

Nota: En Excel esta lógica puede implementarse mediante una fórmula anidada utilizando `SI` y `Y`:

```
=SI(Y([@Latitud]>4.028369,[@Longitud]<-73.610852),"NO",
  SI(Y([@Latitud]>4.028369,[@Longitud]>=-73.610852),"NE",
  SI(Y([@Latitud]<=4.028369,[@Longitud]<-73.610852),"SO",
  "SE")))
```

- **Construcción de la matriz de frecuencia**
  - Ordene los sismos cronológicamente (por fecha y hora).
  - Clasifique cada sismo en su cuadrante correspondiente.
  - Registre la secuencia de cuadrantes de eventos consecutivos.
  - Construya una matriz de frecuencia que indique cuántas veces se transita de un cuadrante a otro.

  Esta matriz permitirá modelar el comportamiento de los sismos como una **cadena de Markov**.

- **Análisis de la matriz de transición y del estado estable**

  Preguntas:

  (a) ¿Qué cuadrante tiene mayor probabilidad de ser el siguiente en la secuencia de sismos según la matriz de transición?

  (b) ¿Existen zonas que "retienen" los sismos, es decir, donde hay mayor probabilidad de que el siguiente sismo ocurra en la misma zona?

  (c) ¿Qué se puede interpretar del hecho de que el cuadrante SE tenga probabilidad límite cero?

  (d) ¿Las probabilidades límite se reparten de forma equilibrada? ¿Qué indica esto sobre el patrón espacial a largo plazo?

  (e) Si ocurriera un sismo hoy en el cuadrante NO, ¿cuál sería la distribución esperada para el próximo evento? ¿Y en 5 pasos?

  (f) ¿Cómo podría usarse este tipo de análisis para apoyar estrategias de preparación ante emergencias sísmicas en Colombia?

  (g) ¿Qué limitaciones tiene este enfoque?

  (h) ¿Qué otras variables deberían considerarse para mejorar la interpretación (por ejemplo, magnitud, profundidad, densidad poblacional)?

  (i) ¿Cómo cambiarían los resultados si en vez de cuadrantes se usaran divisiones políticas o zonas sísmicamente activas conocidas (como el Eje Cafetero o la zona de subducción del Pacífico)?

Instrucciones de entrega:

- La entrega del taller debe realizarse en un documento en formato PDF, con redacción clara, organizada, concisa y estructurada por secciones.
- Todas las respuestas deben incluir la explicación detallada del razonamiento, no solo el resultado final.
- Se espera que los estudiantes analicen, interpreten y argumenten sus respuestas, especialmente en la formulación de modelos y justificación de decisiones.
- El uso de gráficas, ecuaciones, esquemas y tablas es obligatorio cuando estos elementos apoyen o clarifiquen el análisis.
- Los códigos en Python pueden utilizarse como soporte técnico para resolver los modelos, pero no reemplazan la explicación matemática ni conceptual. Es decir: el desarrollo en Python debe servir como base de análisis, no como único medio de respuesta.
- La entrega final debe realizarse de acuerdo con los grupos previamente definidos y ser subida a la plataforma **AVATA**.