**TEMA 1: Introducción a los Sistemas de Big Data.**

**Diferencia entre probabilidad y estadística:**

**- Probabilidad:** Se basa en modelos teóricos y se usa para predecir resultados futuros.

**- Estadística:** Se basa en datos reales y se usa para analizar y describir datos pasados.

La probabilidad es el estudio de la incertidumbre y se utiliza para modelar eventos futuros,

mientras que la estadística se enfoca en la recopilación, análisis e interpretación de datos existentes.

**Tipos de variables:**

**- Variables cuantitativas:** Numéricas y pueden ser discretas (valores enteros) o continuas (valores decimales).

**- Variables cualitativas:** Categóricas y pueden ser nominales (sin orden) u ordinales (con orden).

**- Variables continuas:** Pueden tomar cualquier valor dentro de un rango.

**- Variables discretas:** Sólo pueden tomar valores específicos y separados.

- **Variables datos de panel:** Se incorporan datos en la dimensión temporal, permitiendo analizar cambios a lo largo del tiempo.

**- Variables corte transversal**: Datos recopilados en un solo punto en el tiempo, proporcionando una instantánea de una población o fenómeno.

**- Variables métricas:** Datos que pueden ser medidos y cuantificados, permitiendo operaciones matemáticas.

**- Variables no métricas:** Datos que no pueden ser medidos numéricamente, como categorías o cualidades.

**Fases de un estudio estadístico:**

**1. Delimitación del objetivo del estudio:** Definir claramente el propósito y las preguntas de investigación.

**2. Especificaciones de las características:** Determinar las variables y datos necesarios para el estudio.

**3. Identificación del colectivo:** Seleccionar la población o muestra a estudiar.

**4. Recogida de datos:** Recolectar la información necesaria mediante encuestas, experimentos u otras técnicas.

**5. Revisión de los datos, análisis e interpretación:** Verificar la calidad de los datos, analizarlos estadísticamente e interpretar los resultados.

**Data Lake:** Es un repositorio que permite almacenar grandes cantidades de datos en su formato original.

**Data Warehouse:** Es un sistema utilizado para el almacenamiento y análisis de datos estructurados, optimizado para consultas y generación de informes.

**ETL:** Proceso de Extracción, Transformación y Carga de datos.

**Tema 2: Matemáticas discretas.**

Las matemáticas discretas estudian estructuras que son discontinuas o formadas por elementos separados, a diferencia de las matemáticas continuas. Las matemáticas discretas son la base de la computación, ya que permiten representar y razonar sobre procesos digitales, datos y algoritmos.

MATLAB: Es un lenguaje de programación especializado en el cálculo numérico, análisis de datos y simulación.

Se usa principalmente para:

* Operaciones con matrices y vectores.
* Modelado y simulación de sistemas físicos.
* Análisis de datos y visualización gráfica.

Comandos básicos de MATLAB:

**Creación de vectores y matrices.**

| **Comando** | **Descripción** | **Ejemplo** |
| --- | --- | --- |
| v = [1 2 3 4] | Crear vector fila manualmente | v = [1 2 3 4] |
| v = [1;2;3;4] | Crear vector columna | v = [1;2;3;4] |
| zeros(m,n) | Matriz de ceros de tamaño m×n | Z = zeros(3,4) |
| ones(m,n) | Matriz de unos de tamaño m×n | O = ones(2,5) |
| eye(m,n) | Matriz identidad m×n | I = eye(3) |
| rand(m,n) | Matriz m×n con números aleatorios entre 0 y 1 | R = rand(2,3) |
| randi([a,b],m,n) | Matriz m×n con enteros aleatorios entre a y b | A = randi([0,100],1,10) |
| linspace(a,b,n) | Vector con n valores igualmente espaciados entre a y b | v = linspace(0,10,5) |
| a:b:c | Vector desde a hasta c con paso b | v = 1:2:10 |

**Operaciones básicas sobre arrays y matrices.**

| Comando | Descripción | Ejemplo |
| --- | --- | --- |
| A + B | Suma elemento a elemento | C = A + B |
| A - B | Resta elemento a elemento | C = A - B |
| A .\* B | Multiplicación elemento a elemento | C = A .\* B |
| A ./ B | División elemento a elemento | C = A ./ B |
| A .^ n | Potencia elemento a elemento | C = A .^ 2 |
| A \* B | Multiplicación matricial | C = A \* B |

**Indexación y extracción de elementos.**

| Comando | Descripción | Ejemplo |
| --- | --- | --- |
| A(i,j) | Elemento de la fila i y columna j | x = A(2,3) |
| A(i,:) | Fila i completa | fila = A(2,:) |
| A(:,j) | Columna j completa | col = A(:,3) |
| A(start:end) | Subvector desde start hasta el final | v = A(3:end) |
| A([1 3],:) | Seleccionar varias filas | B = A([1 3],:) |

**Operaciones estadísticas y funciones útiles.**

| Comando | Descripción | Ejemplo |
| --- | --- | --- |
| sum(A) | Suma de todos los elementos (fila por defecto) | s = sum(A) |
| prod(A) | Producto de los elementos | p = prod(A) |
| mean(A) | Promedio | m = mean(A) |
| median(A) | Mediana | med = median(A) |
| max(A) | Máximo valor | m = max(A) |
| min(A) | Mínimo valor | m = min(A) |
| sort(A) | Ordenar los elementos | B = sort(A) |
| length(A) | Número de elementos en un vector | n = length(A) |
| size(A) | Tamaño de la matriz (filas y columnas) | [m,n] = size(A) |

**Modificación de arrays y matrices.**

| Comando | Descripción | Ejemplo |
| --- | --- | --- |
| A(:) | Convertir toda la matriz en un vector columna | v = A(:) |
| reshape(A,m,n) | Cambiar dimensiones de la matriz | B = reshape(A,2,5) |
| A(:,end:-1:1) | Invertir columnas | B = A(:, end:-1:1) |
| A(end:-1:1,:) | Invertir filas | B = A(end:-1:1,:) |
| A(A>50)=50 | Reemplazar elementos según condición | A(A>50)=50 |
| [A B] | Concatenar matrices horizontalmente | C = [A B] |

**Comprobaciones lógicas.**

| Comando | Descripción | Ejemplo |
| --- | --- | --- |
| A > 5 | Devuelve matriz lógica (true/false) | L = A > 5 |
| any(A) | Comprueba si algún elemento es true | any(A>5) |
| all(A) | Comprueba si todos los elementos son true | all(A>0) |