**Análisis de Base de Datos**

**Objetivo:**  
Cada grupo deberá seleccionar una base de datos, realizar un análisis exploratorio básico, comprender su estructura y contenido, y presentar un informe con los hallazgos más relevantes.

Cada grupo debe realizar las siguientes tareas con la base de datos seleccionada:

1. **Selección de la Base de Datos**

* Nombre de la base de datos: Smart\_Farming\_Crop\_Yield\_2024
* Fuente (URL, repositorio, etc.): <https://www.kaggle.com/datasets/atharvasoundankar/smart-farming-sensor-data-for-yield-prediction>
* Formato del archivo (.csv, .xlsx.): .csv

1. **Estructura de la Base de Datos**

* Número total de filas (registros): 501
* Número total de columnas (atributos o variables): 22

1. **Descripción de las Columnas**

Por cada columna se debe especificar:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nombre de la columna** | **Tipo de dato** | **Descripción** | **Cantidad de datos** | **Valores únicos (si aplica)** |
| farm\_id | IDENTIFICADOR | Identificador de la granja | 500 | 500 |
| region | STRING | Ubicación | 500 | 5 |
| crop\_type | STRING | Tipo de cultivo | 500 | 5 |
| soil\_moisture\_% | FLOAT | Humedad del suelo | 500 | 475 |
| soil\_pH | FLOAT | PH suelo | 500 | 184 |
| temperature\_C | FLOAT | Temperatura | 500 | 448 |
| rainfall\_mm | FLOAT | Precipitacion | 500 | 496 |
| humidity\_% | FLOAT | Porcentaje de Humedad | 500 | 481 |
| sunlight\_hours | FLOAT | Horas de luz solar | 500 | 332 |
| irrigation\_type | STRING | Tipo de Irrigacion | 500 | 4 |
| fertilizer\_type | STRING | Tipo de fertilizante | 500 | 3 |
| pesticide\_usage\_ml | FLOAT | uso de pesticidas | 500 | 469 |
| sowing\_date | TIME | Fecha de siembra | 500 | 84 |
| harvest\_date | TIME | Fecha de cosecha | 500 | 123 |
| total\_days | INT | Dias totales | 500 | 61 |
| yield\_kg\_per\_hectare | FLOAT | Rendimiento\_kg\_por\_hectárea | 500 | 500 |
| sensor\_id | IDENTIFICADOR | Identificador del sensor | 500 | 500 |
| timestamp | TIME | Marca de tiempo | 500 | 172 |
| latitude | INT | Latitud | 500 | 500 |
| longitude | INT | Longitud | 500 | 500 |
| NDVI\_index | FLOAT | Indice NDVI | 500 | 61 |
| crop\_disease\_status | STRING | Estado de enfermedad de los cultivos | 500 | 4 |

1. **Calidad de los Datos**

* ¿Hay valores faltantes en alguna columna? ¿Cuántos?

No

* ¿Hay columnas con valores repetidos o inconsistentes?

No

* ¿Existen datos fuera de rango o errores evidentes?

Si, en el formato de las fechas

* ¿Conoce la última fecha de actualización de los datos?

Ultima fecha Agosto 2024

1. **Análisis Cuantitativo Básico**

Para columnas numéricas:

* Mínimo, máximo, media, mediana, desviación estándar.

**soil\_moisture\_%**

Mínimo 10.16

Máximo 44.98

**soil\_pH**

media: 6.42

mediana: 6.355

**temperature\_C**

desviación estándar

Desviación estándar poblacional: 5.99

Desviación estándar muestral: 6.10

Para columnas categóricas:

* Frecuencia de cada categoría.

0

* Número de categorías únicas.

22

1. **Conclusiones Iniciales del Grupo**

* ¿Qué información relevante puede extraerse de esta base de datos?

Localidad, Fechas de cosecha y siembra, Códigos de identificación, tiempos (longitud – latitud) Tipo de fertilizante-pesticida.

* ¿Qué problemas de calidad o estructura podrían afectar un análisis más profundo?

Formato de fechas

* ¿Qué preguntas podría ayudar a responder esta base de datos?

Que cantidad de pesticida se requiere para cada siembra

Diferentes calidades de las cosechas

Tiempo que se tarda desde la siembra hasta la cosecha

* ¿Qué mejoras o limpiezas serían necesarias?

Estandarizar formatos de fechas

Estandarizar por rangos la cantidad de pesticidas y fertilizante utilizada.

1. **Entrega**

Cada grupo debe entregar este documento diligenciado y, si es posible, el archivo de base de datos con sus observaciones o anotaciones.

Juan Diego Arias Grisales

Jacobo Giraldo

Paolo Andrés Patiño