Data Collection and Preprocessing Phase

Date	25 December 2025
Project Title	Predicting Plant Growth Stages with Environmental and Management Data Using Power BI.
Maximum Marks	10 Marks

Data Exploration and Preprocessing:

Dataset Name	Column Name	Data Type	Description	Sample Values
plant_growth_data	Soil_Type	Categorical (Text)	The classification of soil used for the plant sample.	Loam, Sandy, Clay
plant_growth_data	Sunlight_Hours	Numerical (Decimal)	The average duration of sunlight received daily.	5.19, 9.83, 4.03
plant_growth_data	Water_Frequency	Categorical (Text)	How often the plant was watered during the growth cycle.	Daily, Weekly, Monthly
plant_growth_data	Fertilizer_Type	Categorical (Text)	The category of fertilizer applied to the soil.	Chemical, Organic, None
plant_growth_data	Temperature	Numerical (Decimal)	The average ambient temperature recorded in degrees Celsius.	31.7, 18.5, 25.0
plant_growth_data	Humidity	Numerical (Decimal)	The average relative humidity percentage recorded.	61.5, 52.4, 44.6
plant_growth_data	Growth_Milestone	Binary (Integer)	The target variable indicating success (1) or failure (0).	0, 1

Data Exploration (The Variable Dictionary):

- **Soil_Type (Categorical):** We identified values like Loam, Sandy, Clay. This is crucial because your analysis later shows Clay has a high failure rate.
- **Sunlight_Hours (Numerical):** We noted values like 5.19 and 9.83. Recognizing this as a decimal (Float) is important so you don't accidentally treat it as text in Power BI.
- **Growth_Milestone (Binary Target):** We identified this as "0" (Failure) or "1" (Success). This is the **most important column** because this is what you are trying to predict.

Preprocessing Steps:

- **Step 1:**
  - **Duplicate Removal:**
    - Action: Scanned the 193 rows.
    - Result: Found 0 duplicates.
    - Why this matters: If you had duplicates, your success rates would be fake (inflated). Confirming "No Duplicates" proves your data is trustworthy.
- **Step 2:**
  - **Missing Value Imputation:**
    - Action: Checked for blank cells.
    - Result: Found 0 missing values.
    - Why this matters: In Power BI, blank values can break calculations. Confirming the data is full means you don't need to delete rows or guess values (imputation).
- **Step 3:**
  - **Datatype Verification:**
    - Action: Checked for columns, whether it is stored as decimals and categorical columns.
    - Result: Verified.
    - Why this matters: Checked that numerical columns (Temperature) are stored as decimals and categorical columns (Soil_Type) as text. In Power BI, Growth_Milestone is confirmed as an Integer (0/1) suitable for classification.
- **Step 4:**
  - **Outlier Inspection:**
    - Action: Checked for impossible values.
    - Result: Found no critical outliers.
    - Why this matters: It is reviewed to check whether Temperature and Humidity have any impossible values (e.g., >100°C). In Power BI, the values are within realistic agricultural ranges (Temp: 15-35°C).