Team 3: Advanced Soybean Agricultural Dataset

1 Introduction

The `3.csv` a **Advanced Soybean Agricultural** dataset that is a meticulously curated dataset designed to support agricultural research and machine learning applications. Compiled in 2025 as part of a collaborative research effort at the College of Agriculture, University of Tikrit, this dataset provides valuable insights into soybean plant growth, productivity, and nutritional characteristics under varying environmental conditions.

2 Dataset Description

This dataset consists of **55,450 rows** and **13 columns**, capturing essential agricultural parameters related to soybean plants. The dataset includes the following key attributes:

- **Plant Height** – Measures the growth of the soybean plant.
- **Number of Pods** – The count of soybean pods per plant.
- **Biological Weight** – The total biomass of the plant.
- **Chlorophyll Content** – Indicates the plant's photosynthetic efficiency.
- **Protein Percentage** – The percentage of protein in the soybean seeds.
- **Seed Yield** – The total soybean seed production.
- **Relative Water Content** – The water retention capacity of leaves.

A key feature of this dataset is the **Parameters** column, which encodes experimental conditions affecting soybean growth. The parameters include:

- **G (Genotype):** Six different soybean genotypes.
- **C (Salicylic Acid):** Three levels – 250 mg, 450 mg, and a control level.
- **S (Water Stress):** Two levels –
  - Water stress at 5% of field capacity.
  - Water stress at 70% of field capacity.

3 Tasks and Requirements

To analyze and extract meaningful insights from the dataset, the following tasks are required:

3.1 Data Exploration and Preprocessing

- Load and review the dataset.
- Handle missing values and clean inconsistent data.
- Perform exploratory data analysis (EDA) to identify trends and anomalies.
- Normalize numerical features and encode categorical variables.
- Split the dataset into training and test sets for predictive modeling.

3.2 Data Analysis and Modeling

- Analyze the effect of genotype, salicylic acid levels, and water stress on soybean growth and yield.
- Develop predictive models for soybean yield based on key agricultural parameters.
- Identify optimal conditions for maximizing soybean productivity.
- Evaluate correlations between plant physiological traits and seed yield.

3.3 Visualization and Reporting

- Use appropriate visualizations (scatter plots, bar charts, heatmaps) to present insights.

4 Submission Requirements

- A well-structured report detailing the methodology, results, and analysis in a given report format.
- Python code is used for implementation.
- A presentation summarizing key findings and recommendations in a given presentation format.