**Report on Data Preprocessing**

The data collected for the research titled “AI-powered Integrated Crop Management System for Nigerian Farmers” has been prepared for onward usage in building the proposed System.

The data preparation activities encompasses cleaning, normalization, encoding, rebalancing, augmentation, and segmentation as detailed in this document.

**1.0 Experimental Approach**

The experimental analysis was conducted using the Keras framework (version 2.3) with TensorFlow (version 1.14) on a desktop computer equipped with an Intel Core i9-7900X Skylake processor, an NVIDIA GTX 1080 Ti 11GB graphics card, and 64GB of RAM.

The methodology for this analysis began with Exploratory Data Analysis (EDA), which was performed using advanced statistical tools from the scikit-learn library in Python. This initial step involved a thorough examination and analysis of the dataset, focusing on understanding the relationships between soil properties, crop treatments, and crop yield. Following the EDA, statistical tests such as ANOVA and t-tests were conducted to evaluate the significance of mean differences between various groups within the data.

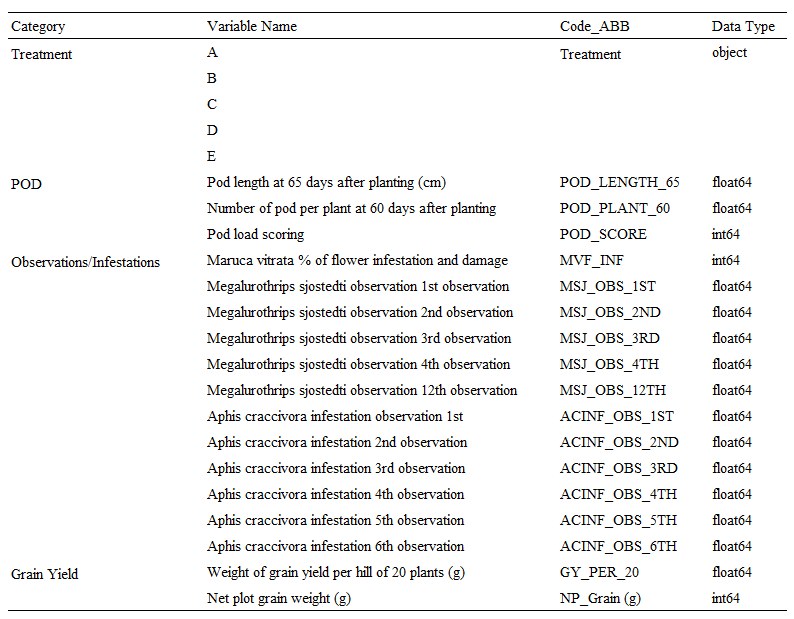
* 1. **Description of Dataset**

The dataset for this research is structured into four main segments:

1. Treatment: This segment details the different crop treatments applied.
2. Yield Data: This includes measurements of grain yield weight.
3. Plant Growth Data: This encompasses metrics such as pod length and the number of pods.
4. Pest Infestation Data: This covers the presence of specific pests, namely **Megalurothrips sjostedti** and **Aphis craccivora**.

Table 1 summarizes the crop yield dataset, outlining key parameters and variables crucial for subsequent analysis and modeling. To enhance clarity and facilitate interpretation, abbreviated codes were used for longer variable names. Additionally, the dataset was supplemented with meteorological data obtained from the local weather station's website for the farm region.

**Table 1 Description of dataset variable**



* 1. **Pre-processing Experiment (Cleaning, Normalization, Encoding, Rebalancing and Segmentation)**

Preprocessing is essential for ensuring data quality and preventing errors during analysis and model training. The process began by merging the soil nutrient data with the soil particle data using the `SampleID` as a key. During data cleaning, it was discovered that the soil nutrient data had missing values. These were addressed using the method proposed by Newgard and Lewis (2015), which involved replacing missing values with the average value from each column. Additionally, the cowpea yield dataset had 10 missing data points, which were removed to prevent anomalies and bias in model deployment.

The dataset includes several categorical variables, such as Treatment, Soil Classes, and Sample Specimen, as outlined in Table 1. To prepare these categorical variables for analysis and machine learning training, numerical values were assigned using the `LabelEncoder` class from scikit-learn. This step facilitates the encoding of categorical data for further processing. Key features that significantly influence yield outcomes were identified from the merged and cleaned data. Spearman’s rank correlation coefficient (Eq. 2) was used to ensure data consistency during feature selection. Features with absolute correlation coefficients greater than 0.35 were selected as crucial for model training. To facilitate analysis and comparison, the dataset was segmented by treatments and replications. Data normalization was then performed to improve data quality and consistency. The `MinMaxScaler` function from scikit-learn was employed for normalization, scaling values to a range between 0 and 1. This step is critical for accurate data evaluation and interpretation, as detailed in Eq. (1).

(1)

Where A is the attribute data considered, and the respective minimum and maximum value for each attribute data considered.

(2)

where '' symbolizes the coefficient outcomes. '' represents the actual occupancy values pertinent to the prediction tasks, while '' corresponds to the values of each feature. Further clarification stems from and , signifying the average values of '' and '', respectively. It is noteworthy that the proximity of the absolute value of to 1 or -1 indicates a heightened degree of correlation.

* 1. **Descriptive Statistics and Data Interpretation**

The yield statistics presented in Table 2 indicate that Treatment C produced the highest average yield (417.5 units), although with significant variability (Std = 112.27 units), suggesting a trade-off between high yield and consistency. Treatment B, while yielding lower on average (327.5 units), exhibited the least variability (Std = 49.76 units), making it the most consistent and reliable option. Treatment D showed relatively high yields (Mean = 356.0 units) but with considerable variability and a skew towards lower yields. Treatment A presented moderate yields with low variability, while Treatment E had the lowest yield (136.5 units) and moderate variability, indicating it was the least effective among the treatments in terms of yield output.

**Table 2 Yield statistics by treatment**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Treatments** | **Mean** | **Median** | **Std** | **Min** | **Max** |
| A | 296.5 | 304.0 | 56.7 | 224 | 354 |
| B | 327.5 | 342.0 | 49.76 | 260 | 366 |
| C | 417.5 | 426.0 | 112.27 | 278 | 540 |
| D | 356.0 | 393.0 | 130.61 | 176 | 462 |
| E | 136.5 | 119.0 | 41.29 | 110 | 198 |

The soil class distribution presented in Table 3, shows that most of the samples are Loamy Sand (13 counts), followed by Sand (4 counts) and Sandy Loam (3 counts). The mixed class of Sand/Loamy Sand and Sandy Clay Loam were less common, each with 3 and 2 counts respectively. This distribution indicates a predominance of loamy sand, with fewer instances of pure sand and other sandy soil types.

**Table 3 Soil class distribution**

|  |  |
| --- | --- |
| Soil Class | count |
| LOAMY SAND | 13 |
| SAND | 4 |
| SANDY LOAM | 3 |
| SAND/LOAMY SAND | 3 |
| SANDY CLAY LOAM | 2 |

The one-way ANOVA result presented in Table 4 indicate a statistically significant difference between the treatment group means, as evidenced by an F-statistic of 5.9653 and a p-value of 0.0044. Since the p-value is below the conventional threshold of 0.05, we can reject the null hypothesis, suggesting that at least one group's mean is significantly different from the others.

**Table 4 One-way ANOVA results**

|  |  |
| --- | --- |
| Metrics | Value |
| F-statistic | 5.9653 |
| p-value | 0.0044 |

The pairwise t-test result presented in Table 5 reveal that most comparisons between treatments do not show statistically significant differences, as indicated by p-values greater than 0.05. Specifically, comparisons between Treatments A and B, A and C, A and D, B and C, B and D, and C and D all have p-values above 0.05, suggesting no significant differences in mean yields between these pairs. However, significant differences were found in the comparisons of Treatments A vs. E (p = 0.0038), B vs. E (p = 0.001), C vs. E (p = 0.0033), and D vs. E (p = 0.0185), indicating that Treatment E's mean yield is significantly different from all other treatments. This suggests that Treatment E is statistically distinct in its effect on yield compared to the other treatments.

**Table 5 Pairwise t-test results**

|  |  |  |  |
| --- | --- | --- | --- |
| Treatment 1 | Treatment 2 | t-statistic | p-value |
| A | B | -0.8219 | 0.4425 |
| A | C | -1.9242 | 0.1027 |
| A | D | -0.8358 | 0.4353 |
| A | E | 4.5624 | 0.0038 |
| B | C | -1.4658 | 0.1931 |
| B | D | -0.4078 | 0.6976 |
| B | E | 5.908 | 0.001 |
| C | D | 0.7142 | 0.5019 |
| C | E | 4.6983 | 0.0033 |
| D | E | 3.2048 | 0.0185 |

The result of this analysis will be used in the next stage of this research which include activities such as selection of features based on the above identified factors influencing crop production, developing the AI model using machine learning algorithm and designing the user interface and data input channels for the AI system.