*Department of Computer Science and Engineering (CSE)*
*Semester: (Fall, Year: 2025), B.Sc. in CSE (Day)*

# Feature Engineering & Data Preprocessing Statistical Analysis and & Visualization

**Lab Report 01**
*Course Title: Data Mining Lab*
*Course Code: CSE 436*
*Section: 221 D3*

Students Details

| Name | ID |
|---|---|
| Arman Hossain | 221002624 |

*Lab Date: 18 Oct 2025*
*Submission Date: 24 Oct 2025*
*Course Teacher's Name: Mayeesha Farjana*

[For teachers use only: Don't write anything inside this box]

| Lab Project Status | |
|---|---|
| **Marks:** | **Signature:** |
| **Comments:** | **Date:** |

# 1 Title of Lab Report

- Implement data processing and visualization on a dataset using Python.

# 2 Objective

- To clean and prepare the dataset by handling missing values and removing duplicates.
- To visualize data distributions and relationships using Seaborn and Matplotlib.
- To normalize data using scaling techniques for better model performance.
- To split the dataset for training and testing the models.
- To calculate correlations between two features.

# 3 Dataset Description

The dataset **Fertilizer Prediction.csv** contains environmental and soil features such as temperature, humidity, moisture, soil type, crop type, and key nutrients (nitrogen, phosphorus, potassium). The target variable is the fertilizer name. This dataset helps recommend the optimal fertilizer based on soil and climate conditions to improve crop yield.

## Dataset Source

- **File Name:** Fertilizer Prediction.csv
- **Origin:** This dataset was downloaded from www.kaggle.com

## Dataset Sample (First 5 Rows)

| Temperature | Humidity | Moisture | Soil Type | Crop Type | N | P | K | Fertilizer Name |
|---|---|---|---|---|---|---|---|---|
| 32 | 51 | 41 | Red | Ground Nuts | 7 | 3 | 19 | 14-35-14 |
| 35 | 58 | 35 | Black | Cotton | 4 | 14 | 16 | Urea |
| 27 | 55 | 43 | Sandy | Sugarcane | 28 | 0 | 17 | 20-20 |
| 33 | 56 | 56 | Loamy | Ground Nuts | 37 | 5 | 24 | 28-28 |
| 32 | 70 | 60 | Red | Ground Nuts | 4 | 6 | 9 | 14-35-14 |

Table 1: First Five Rows of the Fertilizer Prediction Dataset

# 4 Implementation

### Cell 1: Import library and read data set

```
import pandas as pd
import numpy as np
df = pd.read_csv('/kaggle/input/fertilizer-prediction/Fertilizer
   ↪ Prediction.csv')
df.head()
```

### Cell 2: Drop Unwanted Columns & Remove Duplicates

```
df.drop("Crop Type", axis=1, inplace=True)
df = df.drop_duplicates ()
df.head()
```

### Cell 3: Handle Missing Values

```
df['Temparature']=df['Temparature'].fillna(df['Temparature'].mean())
df['Humidity']=df['Humidity'].fillna(df['Humidity'].mean())
df['Nitrogen']=df['Nitrogen'].fillna(df['Nitrogen'].mean())
df['Potassium']=df['Potassium'].fillna(df['Potassium'].mean())
df['Phosphorous']=df['Phosphorous'].fillna(df['Phosphorous'].mean())
```

### Cell 4: Check for Missing Values

```
df.isnull ().sum ()
```

### Cell 5: Detect Outliers (Humidity column)

```
Q1 = new_dataset['Humidity'].quantile(0.25)
Q3 = new_dataset['Humidity'].quantile(0.75)
IQR = Q3 - Q1
lower = Q1 - 1.5 * IQR
upper = Q3 + 1.5 * IQR

outliers_mask = (df['Humidity'] < lower) | (df['Humidity'] > upper)

print("Outlier count:", outliers_mask.sum())
print("Lower limit:", lower)
print("Upper limit:", upper)
```

### Cell 6: Visualizing the distribution of Temperature values using a histogram

```
import matplotlib.pyplot as plt
plt.hist(df['Temparature'], bins=10, edgecolor='black')
plt.title('Histogram of Temparature')
plt.xlabel('Temparature Value')
plt.ylabel('Frequency')
plt.show()
```

**Cell 7: Visualization of the humidity distribution between different soil types using a boxplot**

```
import seaborn as sns

sns.boxplot(x='Soil Type', y='Humidity', data=df)
plt.title('Humidity distribution by Soil Type')
plt.show()
```

**Cell 8: Visualizing the relationship between Temperature and Humidity using a boxplot**

```
import seaborn as sns
sns.catplot(x='Temparature',y='Humidity', kind='box', data=df)
```

**Cell 9: Feature Scaling (Normalization & Standardization)**

```
from sklearn.preprocessing import MinMaxScaler, StandardScaler

mm = MinMaxScaler()
st = StandardScaler()

df['Temparature'] = mm.fit_transform(df[['Temparature']])
df['Humidity'] = mm.fit_transform(df[['Humidity']])
df['Moisture'] = st.fit_transform(df[['Moisture']])
df['Nitrogen'] = mm.fit_transform(df[['Nitrogen']])
df['Potassium'] = st.fit_transform(df[['Potassium']])
df['Phosphorous'] = st.fit_transform(df[['Phosphorous']])
df.head()
```

**Cell 10: Encode Categorical Columns**

```
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()
df['Soil Type'] = le.fit_transform(df['Soil Type'])
df['Fertilizer Name'] = le.fit_transform(df['Fertilizer Name'])
df.head()
```

**Cell 11: Splitting the dataset into training and testing sets (80% train, 20% test)**

```
from sklearn.model_selection import train_test_split

x = df.iloc[:, :-1].values
y = df.iloc[:, 6].values

X_train, X_test, y_train, y_test = train_test_split(x, y, test_size
    ↪ =0.2, random_state=0)

print("X_train:", X_train.shape[0])
print("X_test:", X_test.shape[0])
```

**Cell 12: Checking the correlation between Fertilizer Name and other features (Temperature Humidity)**

```
print(df["Temparature"].corr(df["Fertilizer Name"]))
print(df["Humidity"].corr(df["Fertilizer Name"], method='spearman'))
```

# 5 Output

[19…

| | Temparature | Humidity | Moisture | Soil Type | Crop Type | Nitrogen | Potassium | Phosphorous | Fertilizer Name |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 32 | 51 | 41 | Red | Ground Nuts | 7 | 3 | 19 | 14-35-14 |
| 1 | 35 | 58 | 35 | Black | Cotton | 4 | 14 | 16 | Urea |
| 2 | 27 | 55 | 43 | Sandy | Sugarcane | 28 | 0 | 17 | 20-20 |
| 3 | 33 | 56 | 56 | Loamy | Ground Nuts | 37 | 5 | 24 | 28-28 |
| 4 | 32 | 70 | 60 | Red | Ground Nuts | 4 | 6 | 9 | 14-35-14 |

Figure 1: Cell 1: Read Dataset

[20…

| | Temparature | Humidity | Moisture | Soil Type | Nitrogen | Potassium | Phosphorous | Fertilizer Name |
|---|---|---|---|---|---|---|---|---|
| 0 | 32 | 51 | 41 | Red | 7 | 3 | 19 | 14-35-14 |
| 1 | 35 | 58 | 35 | Black | 4 | 14 | 16 | Urea |
| 2 | 27 | 55 | 43 | Sandy | 28 | 0 | 17 | 20-20 |
| 3 | 33 | 56 | 56 | Loamy | 37 | 5 | 24 | 28-28 |
| 4 | 32 | 70 | 60 | Red | 4 | 6 | 9 | 14-35-14 |

Figure 2: Cell 2: After removing 'Crop Type' and duplicates

[22…

```
Temparature      0
Humidity         0
Moisture         0
Soil Type        0
Nitrogen         0
Potassium        0
Phosphorous      0
Fertilizer Name  0
dtype: int64
```

Figure 3: Cell 3 & Cell 4: After filling missing values with mean Null value count for each column

```
Outlier count: 0
Lower limit: 37.0
Upper limit: 85.0
```

Figure 4: Cell 5: Outlier detection results for Humidity (IQR method)

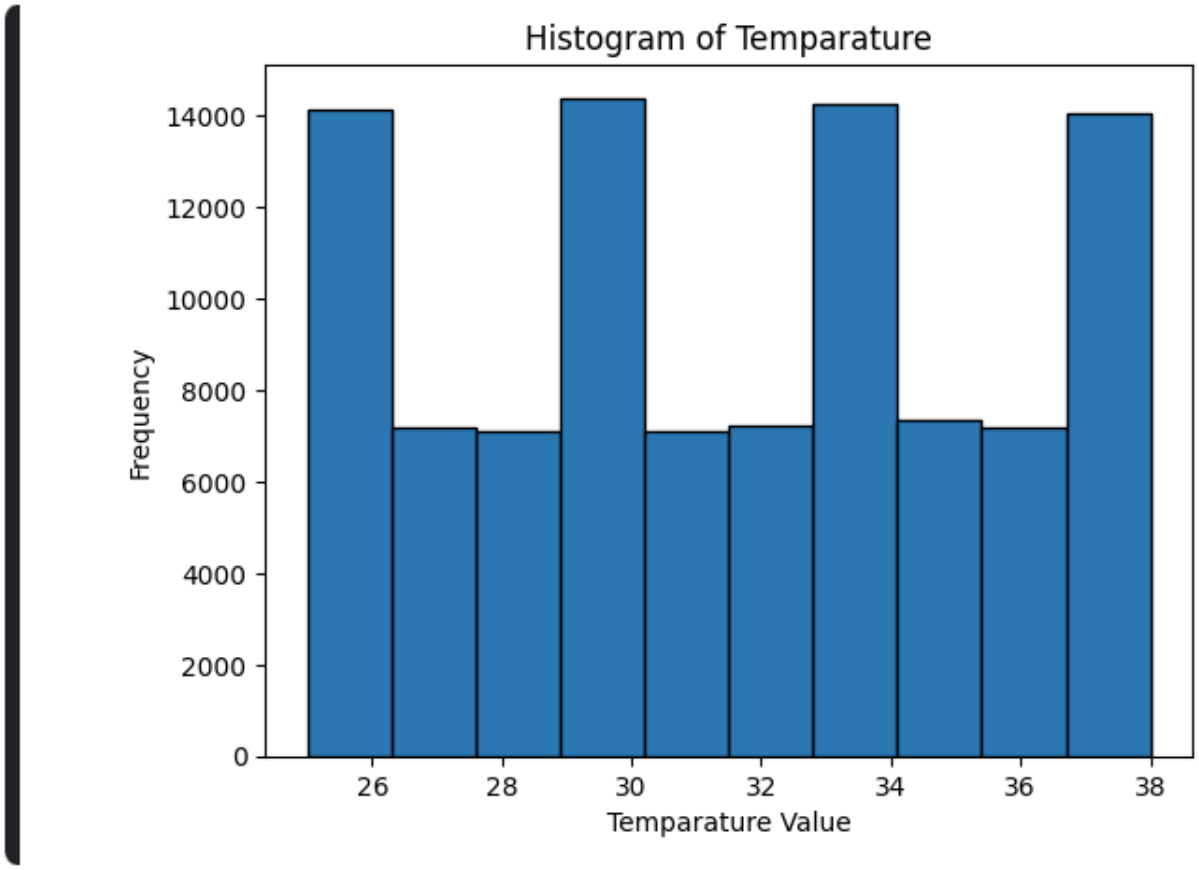


Figure 5: Cell 6: Histogram showing the frequency distribution of Temperature values

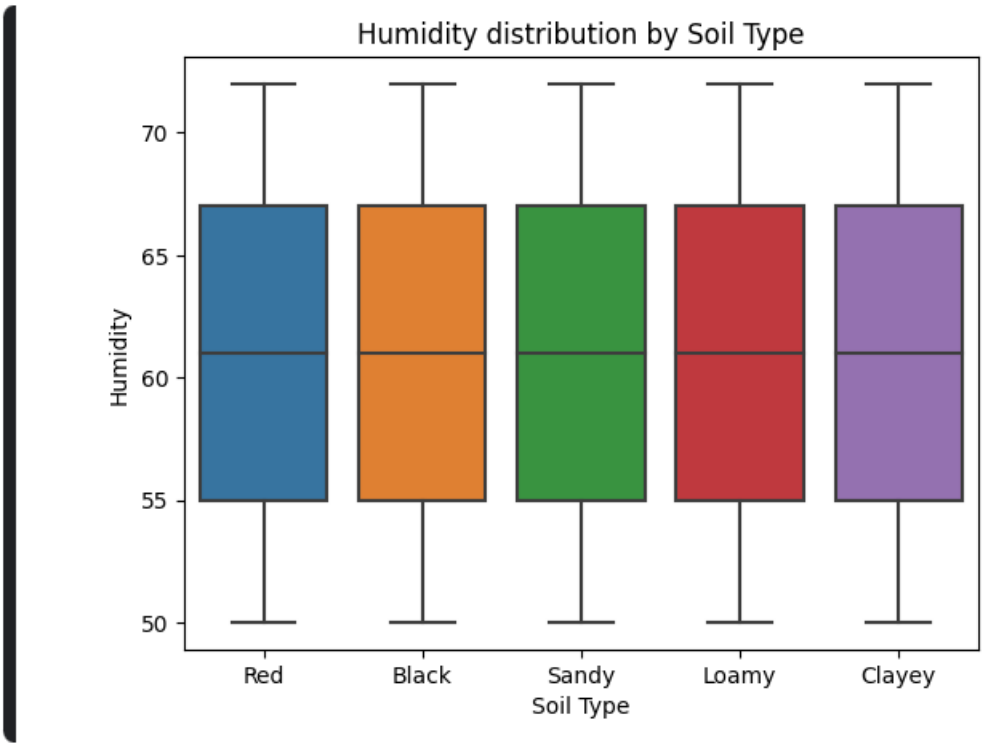


Figure 6: Cell 7: Boxplot showing Humidity distribution by Soil Type

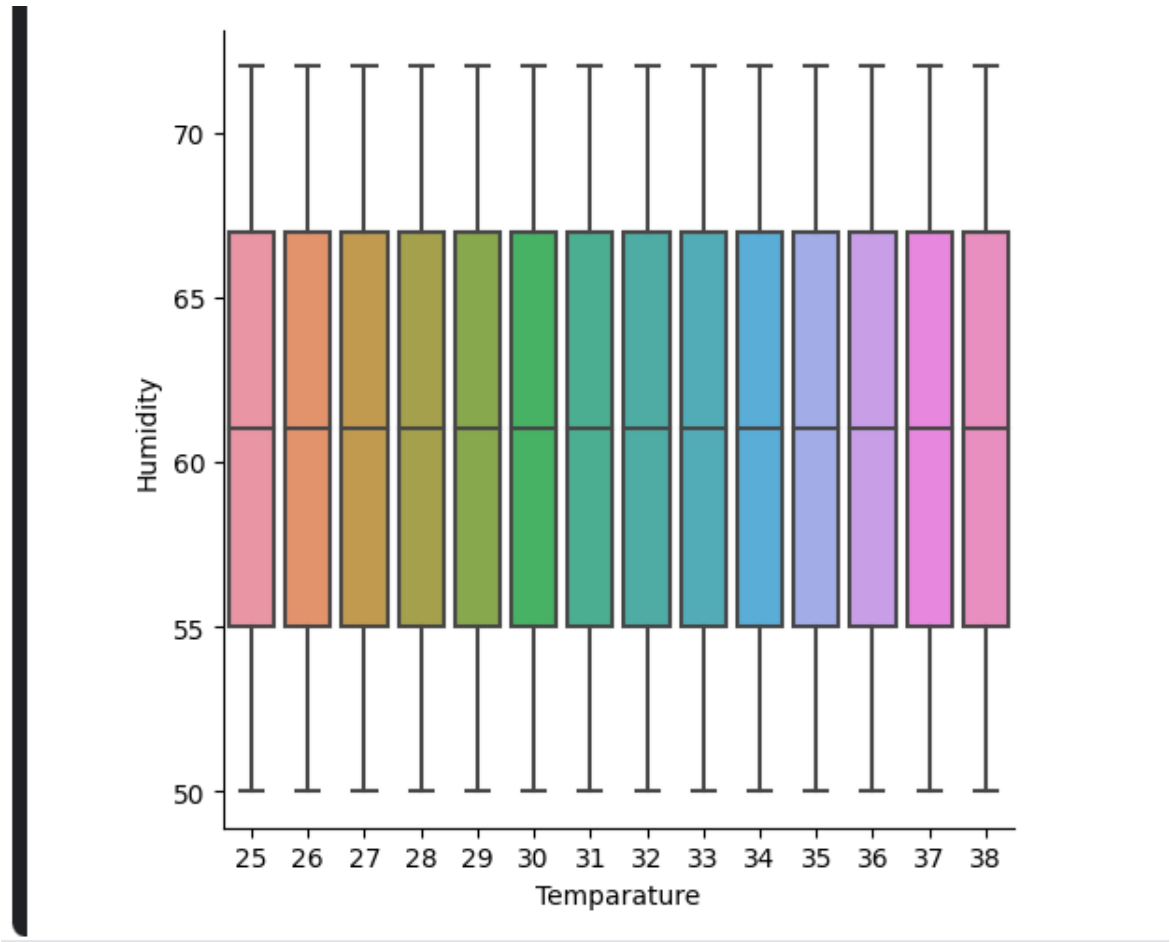


Figure 7: Cell 8: Boxplot showing the relationship between Temperature and Humidity

[27...

| | Temparature | Humidity | Moisture | Soil Type | Nitrogen | Potassium | Phosphorous | Fertilizer Name |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.538462 | 0.045455 | -0.338167 | Red | 0.078947 | -1.121987 | -0.162494 | 14-35-14 |
| 1 | 0.769231 | 0.363636 | -0.844981 | Black | 0.000000 | 0.784910 | -0.404603 | Urea |
| 2 | 0.153846 | 0.227273 | -0.169229 | Sandy | 0.631579 | -1.642049 | -0.323900 | 20-20 |
| 3 | 0.615385 | 0.272727 | 0.928869 | Loamy | 0.868421 | -0.775278 | 0.241021 | 28-28 |
| 4 | 0.538462 | 0.909091 | 1.266746 | Red | 0.000000 | -0.601924 | -0.969523 | 14-35-14 |

Figure 8: Cell 9: Scaled dataset after normalization and standardization

[28...

| | Temparature | Humidity | Moisture | Soil Type | Nitrogen | Potassium | Phosphorous | Fertilizer Name |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.538462 | 0.045455 | -0.338167 | 3 | 0.078947 | -1.121987 | -0.162494 | 1 |
| 1 | 0.769231 | 0.363636 | -0.844981 | 0 | 0.000000 | 0.784910 | -0.404603 | 6 |
| 2 | 0.153846 | 0.227273 | -0.169229 | 4 | 0.631579 | -1.642049 | -0.323900 | 3 |
| 3 | 0.615385 | 0.272727 | 0.928869 | 2 | 0.868421 | -0.775278 | 0.241021 | 4 |
| 4 | 0.538462 | 0.909091 | 1.266746 | 3 | 0.000000 | -0.601924 | -0.969523 | 1 |

Figure 9: Cell 10: Encoded dataset (Soil Type and Fertilizer Name)

```
X_train: 80000
X_test: 20000
```

Figure 10: Cell 11: Dataset split into training and testing sets showing sample counts

```
0.0010671037208614966
0.002917616582511524
```

Figure 11: Cell 12: Dataset split into training and testing sets showing sample counts

# 6 Analysis and Discussion

In this project, I cleaned the Fertilizer dataset by handling missing values and removing duplicates. I detected outliers in Humidity and visualized data distributions using histograms and boxplots. Feature scaling and label encoding prepared the data for modeling. Finally, I split the data into training and testing sets and analyzed correlations between features and fertilizer type to understand their relationships.

# 7 Summary

This lab helped me practice data cleaning, visualization, and basic statistical analysis. I learned to preprocess data, detect outliers, scale features, encode categories, and prepare data for model building. Correlation analysis gave insight into important features for fertilizer prediction.

# 8 References


- https://www.kaggle.com/datasets/irakozekelly/fertilizer-prediction?select=Fertilizer+Prediction.csv
- https://www.kaggle.com/code/ahbijoy121/data-mining-lab-report-01