Data Preparation/Feature Engineering

Data Preparation/Feature Engineering

1. Overview
   The dataset used in this project focuses on crop recommendations, aiming to predict irrigation frequency and other crop-related factors. Data preparation involved exploratory analysis, and feature encoding, which are vital for improving model accuracy.
2. Data Collection
   The dataset was imported using Pandas from a CSV file (Crop_recommendationV2.csv). Initial exploration provided information on the dataset's structure, showing the number of rows, columns, and unique labels in the label column.

Code Snippet:

```
[2] df = pd.read_csv('Crop_recommendationV2.csv')
    df.head()
```

	N	P	K	temperature	humidity	ph	rainfall	label	soil_moisture	soil_type	...	organic_matter	irrigation_frequency	crop_densi
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice	29.446064	2	...	3.121395	4	11.7439
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice	12.851183	3	...	2.142021	4	16.7971
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice	29.363913	2	...	1.474974	1	12.6543
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice	26.207732	3	...	8.393907	1	10.8643
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice	28.236236	2	...	5.202285	3	13.8529

5 rows × 23 columns

3. Data Cleaning
   The dataset had no missing values as confirmed by df.isnull().sum(). Outliers were detected and managed using the IQR method to adjust the values.

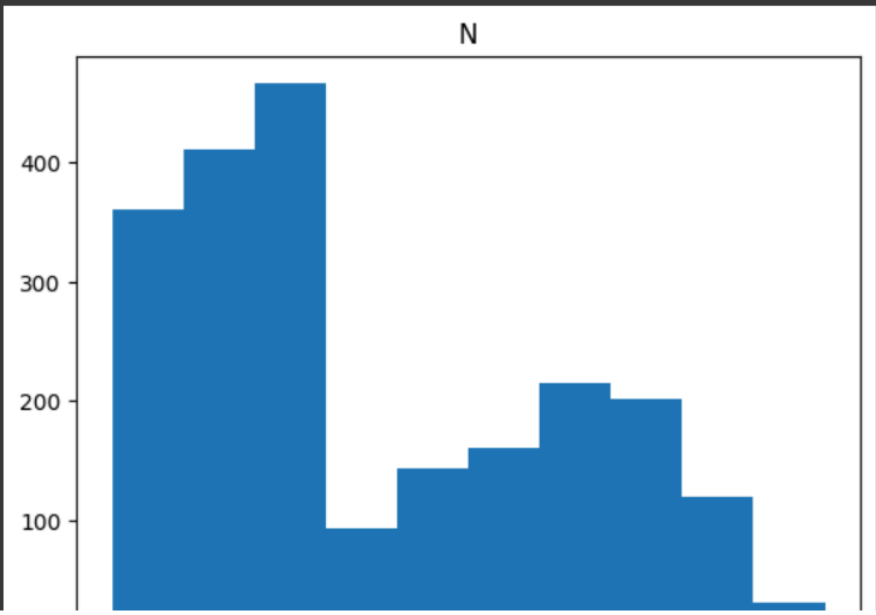
```
[8] df.isnull().sum()
```

	0
N	0
P	0
K	0
temperature	0
humidity	0
ph	0
rainfall	0
label	0
soil_moisture	0
soil_type	0
sunlight_exposure	0
wind_speed	0
co2_concentration	0
organic_matter	0
irrigation_frequency	0

4. Exploratory Data Analysis (EDA)
   Several histograms and boxplots were created to understand the distribution and outliers for different features. A heatmap was also generated to analyze correlations between features.

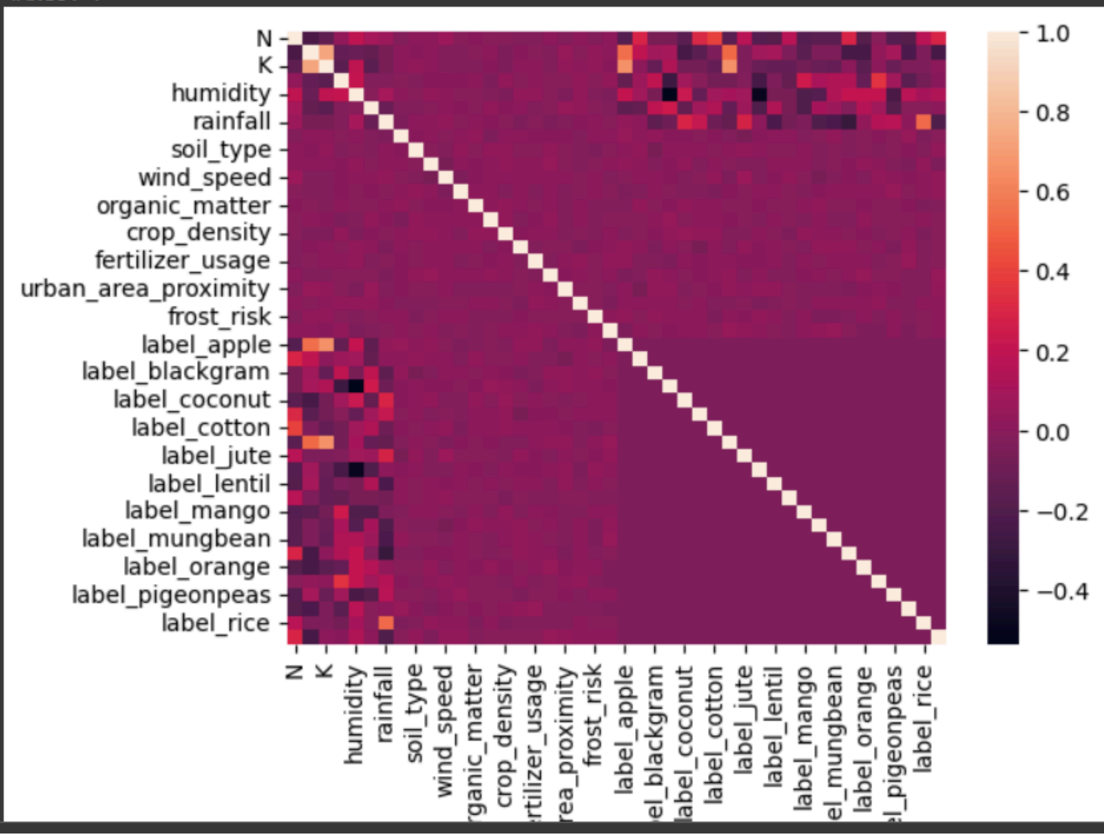
Code Snippet:

```python
import matplotlib.pyplot as plt
for col in df.columns:
  plt.hist(df[col])
  plt.title(col)
  plt.show()
```



```
[14] import seaborn as sns
     sns.heatmap(df.corr())
```

```
<Axes: >
```



5. Key insights included:
    - Close to zero relationship
6. Feature Engineering

   The label column was one-hot encoded to prepare for modeling. This process transformed categorical labels into numeric format, allowing machine learning algorithms to process them.

Code Snippet:

```python
from sklearn.preprocessing import OneHotEncoder

# Create an instance of OneHotEncoder
encoder = OneHotEncoder(handle_unknown='ignore', sparse_output=False)

# Fit and transform the 'label' column
encoded_labels = encoder.fit_transform(df[['label']])

# Create a DataFrame from the encoded labels
encoded_df = pd.DataFrame(encoded_labels, columns=encoder.get_feature_names_out(['label']))

# Concatenate the encoded DataFrame with the original DataFrame
df = pd.concat([df, encoded_df], axis=1)

# Optionally, drop the original 'label' column
df = df.drop('label', axis=1)
```

○

7. Data Transformation
   The dataset was transformed with outlier handling and encoding, ensuring that numerical data was ready for model input.

Model Exploration

1. Model Selection
   Multiple models were tested, including Naive Bayes, Decision Trees, and a Neural Network built using TensorFlow.
2. Model Training
   Different models were trained but i didnt achieve the desired results.