**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 15 July 2024 |
| Team ID | 740067 |
| Project Title | Crop Prediction using machine learning |
| Maximum Marks | 6 Marks |

**Data Exploration and Preprocessing Template**

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

|  |  |
| --- | --- |
| **Section** | **Description** |
| Data Overview | #Structure of the data:        #Descriptive Statistical:  Descriptive analysis is to study the basic features of data with the statistical process. Here pandas has a worthy function called describe. With this describe function we can understand the unique, top and frequent values of categorical features. And we can find mean, std, min, max and percentile values of continuous features. |
| Univariate Analysis | Visual analysis is the process of using visual representations, such as charts, plots, and graphs, to explore and understand data. It is a way to quickly identify patterns, trends, and outliers in the data, which can help to gain insights and make informed decisions.  Univariate Analysis:  In simple words, univariate analysis is understanding the data with a single feature. We have displayed three different types of graphs and plots.  For simple visualizations we can use the matplotlib. pyplot library. Here the plt. figure() command is used to determine the size of the plot.  We have histogram for all features of the dataset which include phosphorus, humidity, temperature as well . The histogram shows the distribution of nitrogen fertilizers for crop.                  Here we have plotted Boxplot using seaborn library. These boxplots can be plotted without using any external library. We have plotted the boxplot using the inbuilt plot function in python.          Here we have plotted distribution plot using seaborn library. These plots can be plotted without using any external library. A Distplot or distribution plot, depicts the variation in the data distribution. Seaborn Distplot represents the overall distribution of continuous data variables. |
| Bivariate Analysis | Bivariate Analysis:  To find the relation between two features we use bivariate analysis. Here we are visualising the relationship between predicted crop and temperature.  #Scatter plot:  Scatterplot can be used with several semantic groupings which can help to understand well in a graph. They can plot two-dimensional graphics that can be enhanced by mapping up to three additional variables while using the semantics of hue, size, and style parameters. All the parameter control visual semantic which are used to identify the different subsets. Using redundant semantics can be helpful for making graphics more accessible.  We have depicted the relationship between temperature and predicted crop using Scatterplot.      #FacetGrid:  FacetGrid class helps in visualizing distribution of one variable as well as the relationship between multiple variables separately within subsets of your dataset using multiple panels. |
| Multivariate Analysis | Multivariate Analysis:  Multivariate analysis is a statistical technique used to analyse data that involves more than two variables. It aims to understand the relationships between multiple variables in a dataset by examining how they are related to each other and how they contribute to a particular outcome or phenomenon.  In multivariate analysis we try to find the relation between multiple features. This can be done primarily with the help of Correlation matrix.      For multivariate analysis we will also plot a Heatmap  This code creates a heatmap that shows how much each column in a given data frame is related to each other column. It does this by first creating a new data frame that has all the columns of the original data frame except for ‘Label’ column. It then calculates the correlation between all the remaining columns and creates a matrix that shows these correlations.  Finally, it generates a heatmap of this matrix using a library called Seaborn.      We also have plot seaborn.pairplot() :  To plot multiple pairwise bivariate distributions in a dataset, you can use the .pairplot() function. |
| Outliers and Anomalies | There is no Outliers in our project. |
| **Data Preprocessing Code Screenshots** | |
| Loading Data | #Loading the data |
| Handling Missing Data | For checking the null values . isnull() function is used. To sum those null values we use. sum() function. From the below image we found that there are no null values present in our dataset. So we can skip handling the missing values step. |
| Data Transformation | ---------- |
| Feature Engineering | --------- |
| Save Processed Data | ---------- |