**CHAPTER THREE**

**METHODOLOGY**

The system uses machine learning to make predictions of the crop and Python as the programming language since Python is accepted widely as a language for experimenting in the machine learning area. Machine learning uses historical data and information to gain experiences and generate a trained model by training it with the data. This model then makes output predictions. The better the collection of dataset, the better will be the accuracy of the classifier. I observed that machine-learning methods such as regression and classification perform better than various statistical models.

Crop production is completely dependent upon geographical factors such as soil chemical composition, rainfall, terrain, soil moisture, sunlight intensity, temperature etc. These factors play a major role in increasing crop yield. In addition, market conditions affect the crop(s) to be grown to gain maximum benefit. We need to consider all the factors altogether to predict the yield.

Hence, using Machine Learning techniques in the agricultural field, we build a system that uses machine learning to make predictions of the production of crops by studying the factors such as rainfall, temperature, area, season, etc.

**MACHINE LEARNING**

Machine Learning is undeniably one of the most influential and powerful technologies in today‘s world. Machine learning is a tool for turning information into knowledge. In the past 50 years, there has been an explosion of data. This mass of data is useless; we analyzed it and found the patterns hidden within. Machine learning techniques are used automatically to find the valuable underlying patterns within complex data that we would otherwise struggle to discover.

The hidden patterns and knowledge about a problem can be used to predict future events and perform all kinds of complex decision-making. To learn the rules governing a phenomenon, machines have to go through a learning process**,** trying different rules and learning from how well they perform. Hence, why it‘s known as Machine Learning.

**BASIC TERMINOLOGY**

● Dataset: A set of data examples, which contain features important to solving the problem.

● Features: Important pieces of data that help us understand a problem. These are fed into a

Machine Learning algorithm to help it learn.

● Model: The representation (internal model) of a phenomenon that a Machine Learning algorithm has learnt. It learns this from the data it is shown during training. The model is the output you get after training an algorithm. For example, a decision tree algorithm would be trained and produce a decision tree model.

**BASIC PROCESS**

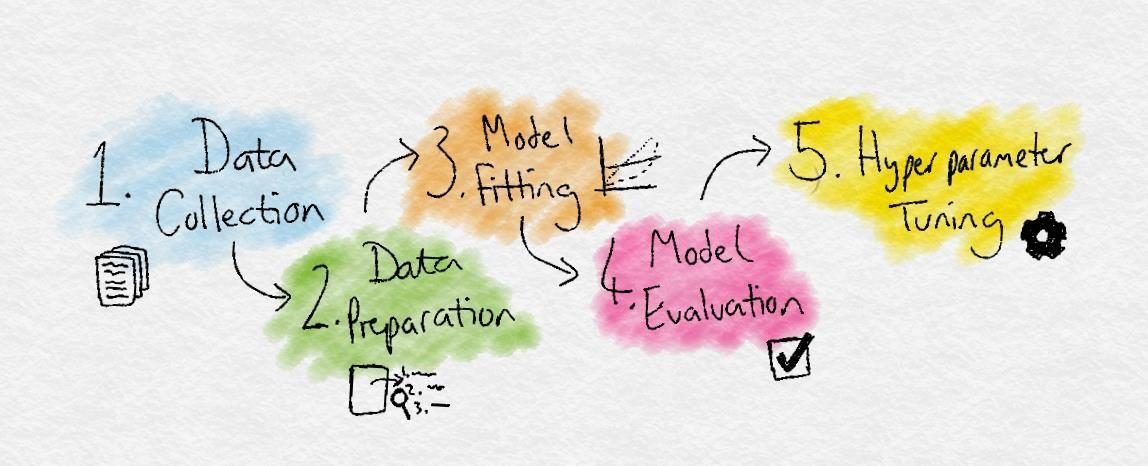
i. Data Collection**:** Collect the data that the algorithm will learn from.

ii. Data Preparation**:** Format and engineer the data into the optimal format, extracting important features and performing dimensionality reduction.

iii. Training**:** Also known as the fitting stage, this is where the Machine Learning algorithm actually learns by showing it the data that has been collected and prepared.

iv. Evaluation**:** Test the model to see how well it performs.

v. Tuning**:** Fine-tune the model to maximize its performance.



*General Processes involved*

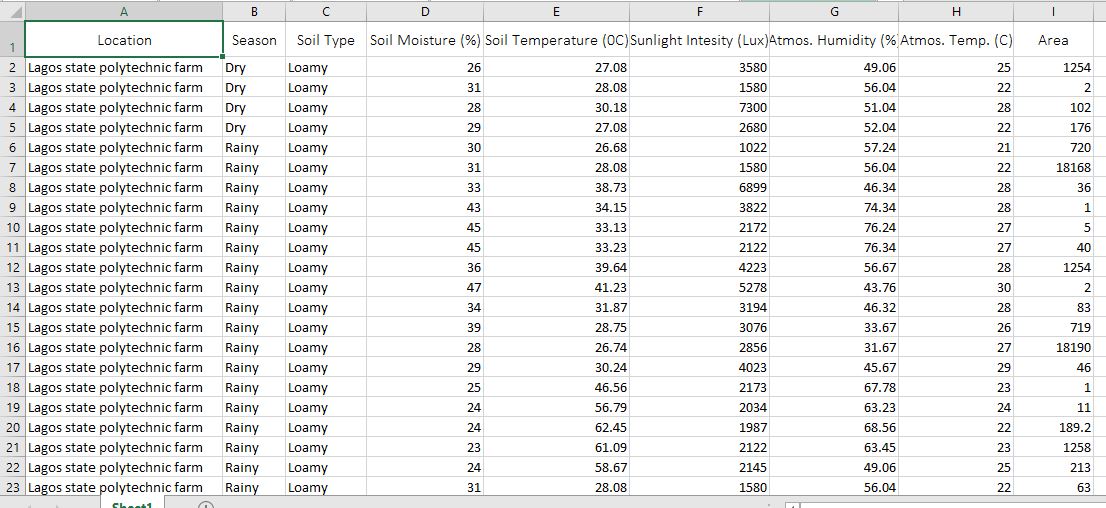
**DATASETS**

Machine Learning depends heavily on data. It‘s the most crucial aspect that makes algorithm training possible. It uses historical data and information to gain experiences. The better the collection of the dataset, the better will be the accuracy.

The first step is Data Collection. For this project, we collected two datasets that were used to form a complete dataset for modelling the yield prediction algorithm.

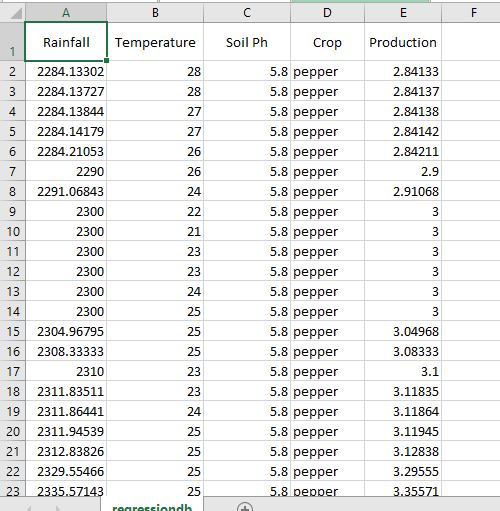
These two parameters were used as inputs for predicting the crop yield. The sources of our datasets are from Lagos State Polytechnic Farm.

The first module dataset has the following columns: Location, Season, Soil Type, Soil Moisture, Soil Temperature, Sunlight Intensity, Atmospheric Humidity, Atmospheric Temperature, and Area.



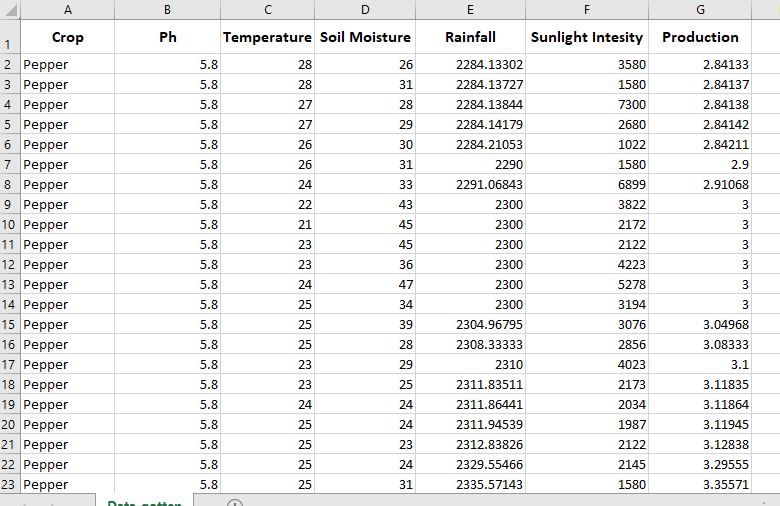
*First module dataset*

The second data set contains the following columns: Rainfall, Temperature, Soil Ph., Crop, and Production, which are the major factors that crops depend on. Here, Production is the dependent variable or the class variable.



*Second module dataset*

We achieved this by merging the datasets. The merged datasets took location and Temperature as the common attribute in both. There are six independent variables and one dependent variable. We achieved this by merging the datasets. We considered only one here, Lagos State Polytechnic Farm as the suicide rates of farmers in these area found to be very high.

**

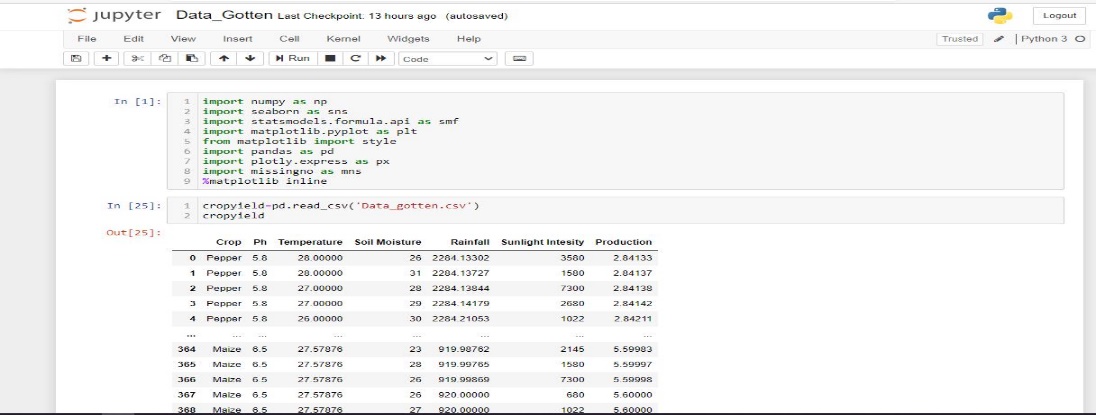
*The Merged Dataset*

**CHAPTER FOUR**

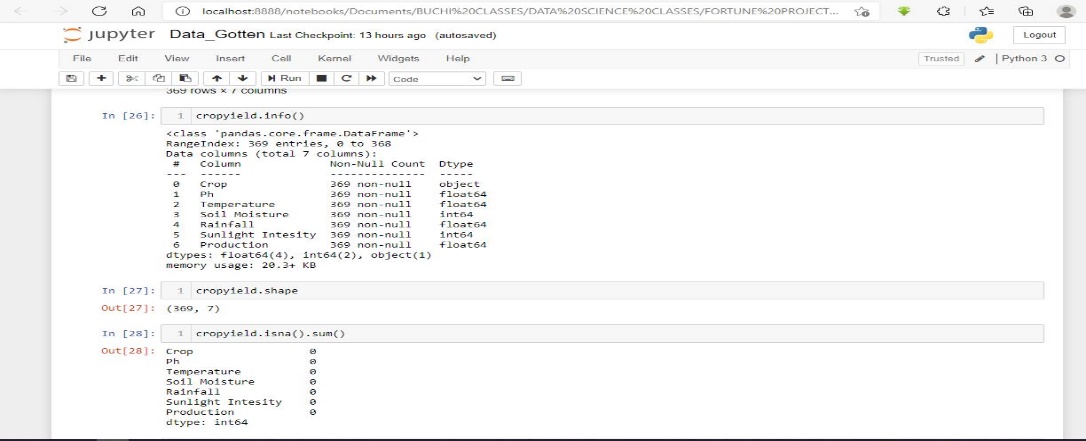
**EXPLORATORY DATA ANALYSIS (EDA)**

It is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. It is also about knowing your data, gaining a certain amount of familiarity with the data, before one starts to extract insights from it. The idea is to spend less time coding and focus more on the analysis of data itself. After the data collection, it undergoes some processing before cleaning and EDA is performed. Subsequently, use the cleaned dataset and knowledge from EDA to perform modelling and reporting. Exploratory data analysis is generally cross- classified in two ways. First, each method is either non-graphical or graphical. Second, each method is either univariate or multivariate (usually just bivariate). It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand. EDA can give us the following:

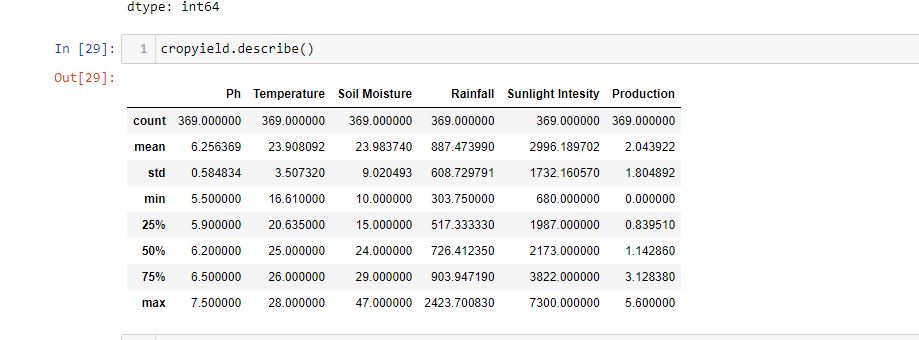
* Importing the necessary tools required in data analysis.
* Data preview.



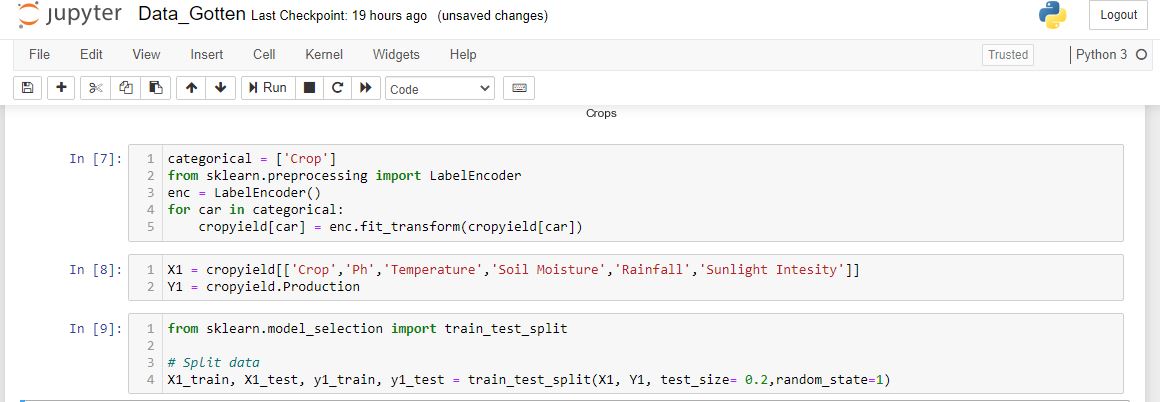
* Check total number of entries and column types using in built functions. It is a good practice to know the columns and their corresponding data types.
* We can get an insight of the number of values in each column, which can give us information about the null values or the duplicate data.



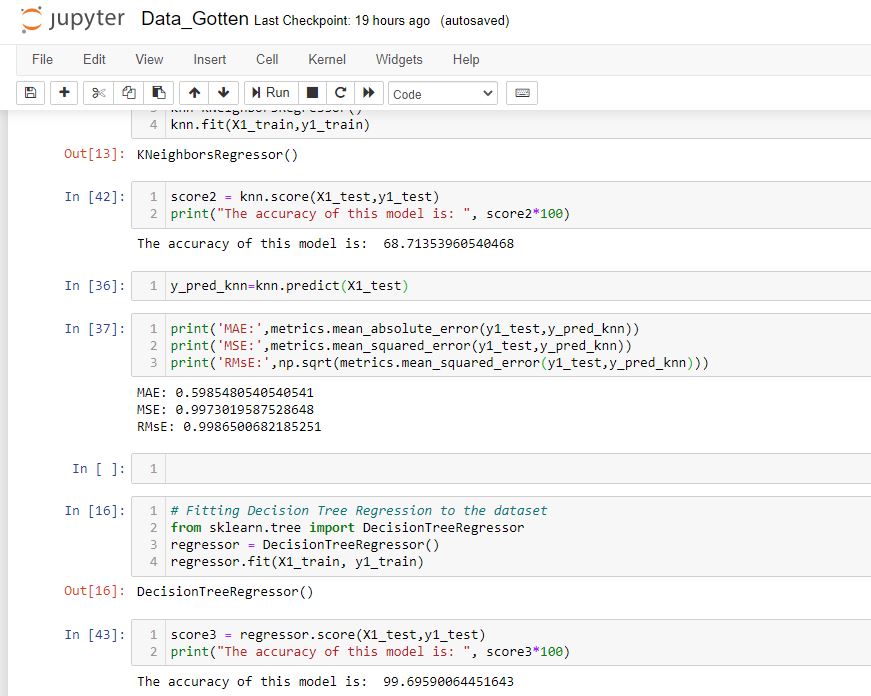
* We can also find the mean, standard deviation, minimum value and the maximum value.



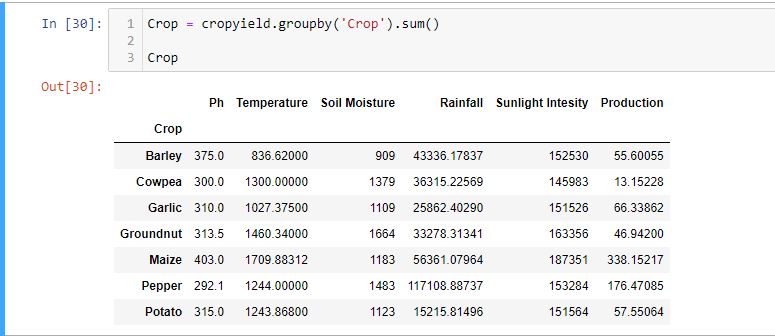
* Training and splitting the model based on parameters

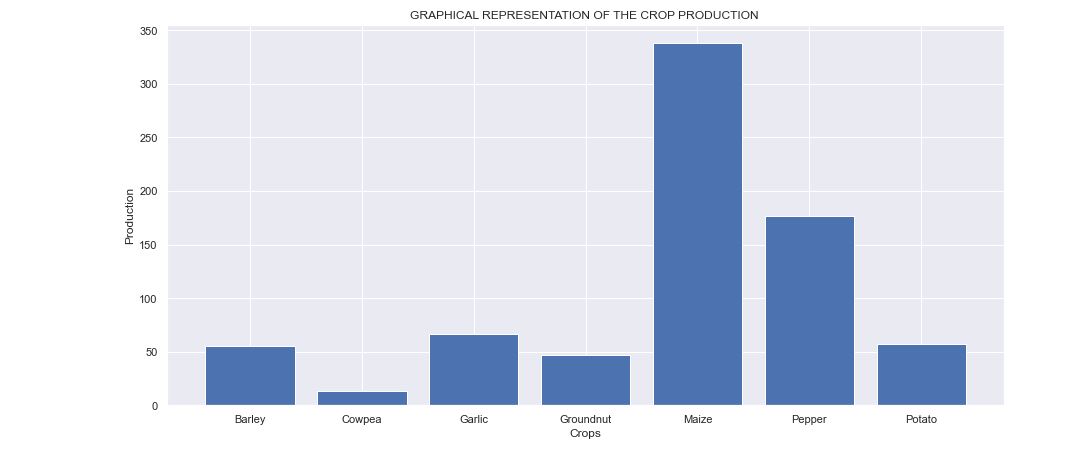


* Prediction of production (Output model) based on the accuracy of the given dataset. This is the basic procedure of EDA.



* To get a better insight of the used data, we can plot graphs to know the crop with a good yield following the given dataset, which gives us a lot of information about how variables (columns) are related to each other and the impact each of them have on the other.





**ALGORITHMS USED**

Machine Learning offers a wide range of algorithms to choose. They are divided usually into classification, regression, clustering and association. Classification and regression algorithms come under supervised learning while clustering and association comes under unsupervised learning.

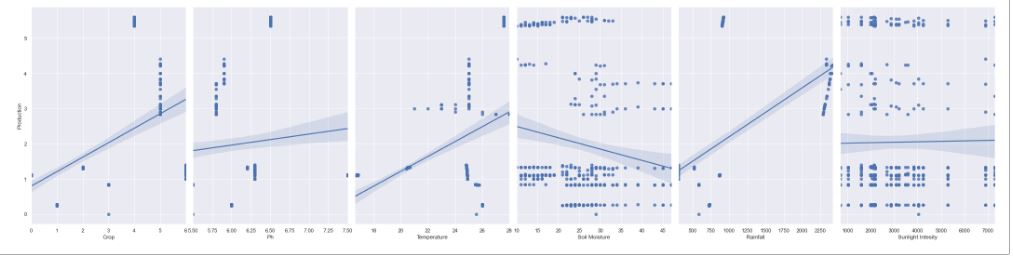
A few algorithms can come under multiple types. Considering the problem statement and the desired output of the project, the most suitable type of algorithm would come under regression.

**REGRESSION:**

Regression is when the output variable is a real value, such as dollars or weight. Example: Linear Regression, Logistic Regression, etc.

Before choosing an algorithm and working with it further, many algorithms were explored and the error rates and accuracy were checked for each.

This model is used to obtain relationships between the production of crops and soil parameters. This algorithm will help us in predicting the amount of production for a particular crop. We are using linear regression to identify the relationship of production with each of the parameters.



*Representation of relationship between production of crops and different parameters.*

1. **Linear Regression**

It is the simplest form of regression. The dependent variable is continuous in nature in a technique. The relationship between the dependent variable and independent variables is assumed linear in nature.

When you have only one independent variable and one dependent variable, it is called simple linear regression.  
When you have more than one independent variable and one dependent variable, it is called multiple linear regression.

**The equation of multiple linear regression listed below -**

|  |
| --- |
| <https://2.bp.blogspot.com/-xbqTM5K3bIU/WkzhtHMPEmI/AAAAAAAACFs/RULnlMKw_0U14oRWOUcuETJNt9TBYiJEgCLcBGAs/s1600/b.jpg> |

Here 'y' is the dependent variable to be estimated, and X are the independent variables and ε is the error term. βi’s are the regression coefficients.

**Estimating the parameters**

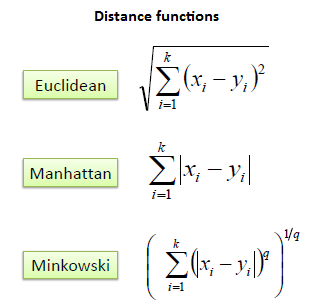
To estimate the regression coefficients βi’s we use principle of least squares, which is to minimize the sum of squares due to the error terms i.e.   
  
On solving the above equation mathematically we obtain the regression coefficients as:



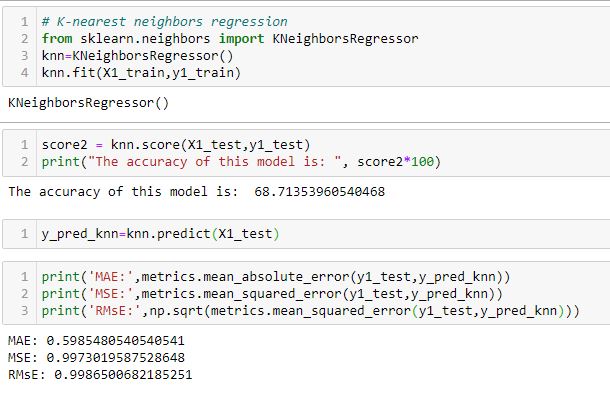


1. **K-Nearest Neighbors Regression(KNN)**

This simple algorithm stores all available cases and predict the numerical target based on a similarity measure. KNN regression uses the same distance functions as KNN classification.



The above three distance measures are only valid for continuous variables. In case of categorical variables, you must use the Hamming distance, which is a measure of number of instances in which corresponding symbols are different in two strings of equal length.



1. **Decision Tree Regression**

It observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

