**METHODOLOGY**

KNN (K-Nearest Neighbors) classification is a supervised machine learning algorithm used for both regression and classification tasks. It is a non-parametric algorithm that assigns a class label to an input based on the majority class of its k nearest neighbors in the feature space. In KNN classification, the algorithm builds a model by memorizing the entire training dataset. When a new input is provided, it calculates the distances between the input and all the training samples. The "k" in KNN represents the number of nearest neighbors to consider. The algorithm selects the k nearest neighbors based on the calculated distances. For classification, KNN uses majority voting to assign a class label to the input. The class label assigned is based on the class labels of the k nearest neighbors. The most common class label among the neighbors is chosen as the predicted class for the input. However, it has some considerations, such as the choice of the distance metric, the value of K, and the potential impact of imbalanced datasets. Additionally, the algorithm's computational complexity increases as the size of the training dataset grows, making it less suitable for large-scale datasets.

1. **Data collection**

The data set took from Harvard Dataverse. The Harvard Dataverse Repository is a free data repository open to all researchers from any discipline within and outside the Harvard community, where you can share, document, cite, access, and explore research data. This dataset was made by augmenting optimum soil and environmental characteristics for crop growth. This data set was created in 2019-11-02.

1. **About data set**

in this data set Users can build a predictive model using the data in this collection to suggest the ideal crops for a farm depending on a variety of factors. This dataset was created by combining rainfall, climate, and fertilizer data from India. Based on environmental information, farmers may be uncertain about which crops to plant to ensure good yields. By spotting patterns in previous data, they will be able to form a more accurate view of what they need to develop to achieve their goal. This data set can help us understand how rainfall, ph level, humidity, and nutrition like N, P, and K affect plant growth, and which crops are better to grow in different conditions like high rainfall, high N ratio. There are 8 columns and 1697rows in the data set. In the data set the first 7 columns show the different parameters to recommend the crop and the 8th column shows the crops. The first 7 main variables are,

• N- Ratio of nitrogen content in soil

• P- Ratio of phosphorus content in soil

• K- Ratio of Potassium content in soil

• Temperature- Temperature in degree Celsius

• Humidity – humidity in %

• ph- ph value of the soil

• Rainfall- rainfall in mm

In First 7 columns there are independent variables and last column (Label) there are the dependent variables. It depends on the independent variables.