# CROP YIELD ESTIMATION USING MACHINE LEARNING ALGORITHM

## Problem statement:

Crop yield prediction is an important agricultural problem. The Agricultural yield primarily depends on weather conditions (rain, temperature, etc), pesticides. Accurate information about the history of crop yield is important for making decisions related to agricultural risk management and future predictions. Predicting the crop yield in advance of its harvest would help the policy makers and farmers for taking appropriate measures for marketing and storage.

## Proposed Work:

In this project, the proposed system determines the suitable crop to be cultivated and its appropriate yield estimation. The system gets the input such as soil minerals, moisture, temperature (max and min), humidity, rainfall etc. for predicting the crop estimation This system analyses the crop yield estimation based on available data in the dataset. This project focuses on predicting the yield of the crop based on the existing data by using Naïve bayes classifier, binary Support Vector machine classifier and Decision tree algorithm.

From experimental results, it has been predicted that Decision tree algorithm found to be most accurate technique for crop yield prediction.

A. Dataset Collection: In this phase, we collect data from various sources and prepare datasets. And the provided dataset is in the use of analytics. There are online abstracts source such as kaggle.

B. Acquisition of training dataset:
The accuracy of a machine learning algorithm may depend on the number of parameters used and to the extent of correctness of the dataset.Our dataset contains the soil minerals, moisture, temperature (max and min), humidity,rainfall etc.Thus, by using an appropriate machine learning algorithm we can train the dataset to predict the most suitable crop that can be grown under the given input parameters.

C. Data pre-processing:
Data pre-processing is a technique that is used to convert the raw data into a clean data set. Data preprocessing is the second step and it contains two steps. Original dataset can contain lots of missing values so initially all these should be removed. Missing values are denoted by a dot in the dataset and their presence can deteriorate the value of entire data and it can reduce the performance.
So, to solve this problem we replace these values with large negative values which will be treated as outliers by the model. Generating the class labels is the second step.

D. Models/algorithms selection for comparison:

Before deciding on an algorithm to use, first we need to evaluate, compare and choose the best one that fits this specific dataset.

Usually, when working on a machine learning problem with a given dataset, we try different models and techniques to solve an optimization problem and fit the most suitable model, that will neither overfit nor underfit the model.

E. Machine Learning Algorithm:

Machine learning is an important decision support tool for crop yield prediction.

This algorithm is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training, data. Different machine learning algorithms are being used in order to make comparisons. The different algorithms used are as follows:

## Dataset used:

| | Unnamed: 0 | Moisture | rainfall | Average Humidity | Mean Temp | max Temp | Min temp | alkaline | sandy | chalky | clay | millet yield |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 12.801685 | 0.012360 | 57 | 62 | 71 | 52 | 0 | 1 | 0 | 0 | 2.0 |
| **1** | 1 | 12.851654 | 0.004172 | 57 | 58 | 73 | 43 | 0 | 1 | 0 | 0 | 0.0 |
| **2** | 2 | 12.776773 | 0.000000 | 56 | 58 | 69 | 46 | 0 | 0 | 1 | 0 | 4.0 |
| **3** | 3 | 12.942001 | 0.031747 | 62 | 56 | 70 | 43 | 0 | 1 | 0 | 0 | 0.0 |
| **4** | 4 | 12.984652 | 0.000000 | 65 | 56 | 70 | 42 | 0 | 0 | 0 | 1 | 1.0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **3972** | 3972 | 12.959730 | 0.000000 | 30 | 72 | 85 | 58 | 1 | 0 | 0 | 0 | 4.0 |
| **3973** | 3973 | 12.985416 | 0.000000 | 30 | 70 | 87 | 52 | 0 | 1 | 0 | 0 | 4.0 |
| **3974** | 3974 | 12.947405 | 0.000300 | 33 | 70 | 86 | 53 | 0 | 1 | 0 | 0 | 0.0 |
| **3975** | 3975 | 12.771689 | 0.000000 | 35 | 69 | 85 | 53 | 0 | 1 | 0 | 0 | 4.0 |
| **3976** | 3976 | 12.845779 | 0.010131 | 36 | 70 | 86 | 53 | 1 | 0 | 0 | 0 | 0.0 |

3977 rows × 12 columns

## Workflow:

- Initially load the dataset which is in csv file format using pandas library with all features.
- Clean the data if it has missing values, null values by removing that value in the dataset.
- Train the dataset in the training dataset.
- Apply various algorithms on the dataset.
- Predict the results obtained by different algorithms on the training dataset.
- Finally, the algorithms give their respective accuracy of crop yield by considering the features in the dataset.

## Reason for choosing Decision trees:

Decision trees tend to be the method of choice for predictive modeling because they are relatively easy to understand and are also very effective. the basic goal of a decision tree is to split a population of data into smaller segments.

There are two stages to prediction. The first stage is training the model, where the tree is built, tested, and optimized by using an existing collection of data. In the second stage, it uses the model to predict an unknown outcome.

## Conclusion

In this proposed system, we have applied Decision tree Algorithm and predicted the crop yield in which it is getting 99.4% of accuracy by taking statistical raw datasets which contains various fields like Average Humidity, Mean Temp, Min Temp, Moisture, Alkaline, Chalky, sandy, Max Temp, Millet Yield, Rainfall and, Sandy. As python is easy which has run time results and Jupyter is useful for the entire implementation is done by using Anaconda Software which consists of both Python IDLE and Jupyter Notebook.