**AGRICULTRE DATASET**

**INTRODUCTION**

we have observed the emerging concept of smart farming that makes agriculture more efficient and effective with the help of high - precision algorithms. The mechanism that drives it is Machine Learning — the scientific field that gives machines the ability to learn without being strictly programmed. It has emerged together with big data technologies and high - performance computing to create new opportunities to unravel, quantify , and understand data intensive processes in agricultural operational environments.

Machine learning is everywhere throughout the whole growing and harvesting cycle. It begins with a seed being planted in the soil — from the soil preparation , seeds breeding and water feed measurement — and it ends when neural networks pick up the harvest determining the ripeness with the help of computer vision.

**PROBLEM DEFINATION FOR DATASET**

**The Toxic Pesticides**

Though , many of us don't appreciate much , but a farmer's job is real test of working ability and determination. Once the seeds are sure own , he works days and nights to make sure that he cultivates a good harvest at the end of season. A good harvest is ensured by several factors such as availability of water , soil fertility , protecting crops from rodents , timely use of pesticides & other useful chemicals and nature. While a lot of these factors are difficult to control for , the amount and frequency of pesticides is something the farmer can control.

Pesticides are also special because while they protect the crop with the right dosage. But if you add more than required they may spoil the entire harvest. A high level of pesticide can deem the crop dead / unsuitable for consumption among many outcomes. This data is based on crops harvested by various farmers at the end of harvest season. To simplify the problem , you can assume that all other factors like variations in farming techniques have been controlled for. You need to determine the outcome of harvest season, that whether the crop healthy, damaged by pesticides or damaged by other reason.

**Data Description**

Variable

ID: UniqueID

Estimated\_Insects\_Count: Estimated insects count per square meter

Crop\_Type: Category of Crop(0,1)

Soil\_Type: Category of Soil (0,1)

Pesticide\_Use\_Category: Type of pesticides uses (1- Never, 2-Previously Used, 3-Currently Using)

Number\_Doses\_Week: Number of doses per week

Number\_Weeks\_Used: Number of weeks used

Number\_Weeks\_Quit: Number of weeks quit

Season: Season Category (1,2,3)

Crop\_Damage: Crop Damage Category (0=alive, 1=Damage due to other causes, 2=Damage due to Pesticides).

**DATA ANALYSIS**

I tried to understand the dataset and analysis that feature and the description for solving the problem of dataset I analyzed the data and did the necessary transformation , normalization , on it. And also some visualization process to understand the dataset and achived the target and also use the Logisticregression, Gridsearchcv,Randomforestclassifier and check them with the train and test data function I used tools Pandas, Matplotlib, Sklearn . So by the help of these algorithms we understand the dataset and achive the target of the dataset. Basicly our approached to evaluate the dataset and find out the accuracy .

**EDA**

I started with df.head() , df.describe() and df.info() that are uses for define indicate dataset , types , null values and column , shape Then I do the describe process in dataset and understand the mean, mini , max values of the dataset and the check their differences. After that check the dataset train pesticide with their category and also with find out the null value of dataset number week pesticide and similarly that process with their define value count. Then after that we used to also with indicate the histogram, after these process fill the na(nan) value of the train and test category which define number of week used and pesticide used category Data columns(index) are shown the I describe the dataset which the given pesticide use category, for define the total doses uses by multiplying the no. doses week and no. doses used. we will also find out no. doses monthly by total doses used divide by 30, for find out no. doses daily uses by divide total doses divide by 360, number of days quite find out by number week quit multiply 7 then describe the dataset. That eda process well understand by seeing the formula and working with dataset(main file).

**Data Preprocessing and transformations.**

We understand In the section we built a stages to understand the preprocessing requirement for our data. It is now time to form a preprocessing pipeline design based on our learning from the section. We will define our preprocessing pipeline in three stages: We will use a Transformer to do the required transformations.

1. Drop the columns that are not required for model training used the data for training and testing purpose.

2.The training label are done for training and testing data then check the shape and also concatation process for train and test dataset.

3. That process are necessary to find out the accuracy and their working with dataset.

That are the three important steps of preprocessing and these steps are also included with the ‘run’ the dataset we understand to see working in dataset.

**MODEL BUILDING**

Now that we are done with the basic pre-processing steps , we can go ahead and build simple machine learning models over this data. We will try models here – StandardScaler, LogisticRegressor and Gridsearchcv Random Forest Regressor, DecisionTreeClassifier to predict the result. In that dataset that are also check the values for count and nan with appropriate conditions and also fill the nan values , describe check shape then the data preprocessing step for train the dataset and concate the value of train, test . In spliting the dataset process final train and test the data shape and standard scaler process also done fit and transform the train , test dataset. creating cross validation and training model on train dataset with the use of logistic regressor train and predict the value then also find out the accuracy score use gridsearchcv and decisiontreeclassifier for predict for predicting final model that parameters are best suited for accuracy.

**CONCLUDING REMARK**

In these dataset we will do the process train , test , fit and predict the model and all are the process we will understand the target dataset and also find the their accuracy and which logistic classifier are best suited for the dataset and also understand the target. And find out the accuracy for harvest season about the use of pesticide and reaction on crops. so I am trying to achieve our target because I tried to cover all the dataset to given output and reduce minimum data that’s reason trying to something new from my side and trying to achieve dataset.