**IMPLEMENTATION**

**MODULES:**

* Data Collection
* Dataset
* Data Preparation
* Model Selection
* Analyze and Prediction
* Accuracy on test set
* Saving the Trained Model

**MODULES DESCSRIPTION:**

**Data Collection:**

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

We attached data set in the document , The file name is crop\_modified.csv

**Dataset:**

The dataset consists of 2232 individual data. There are 9 columns in the dataset, which are described below.

Id: unique id

State\_Name: India is a union of States and Union Territories for the purposes of administration, India is divided into 29 States

District\_Name: District in india

Crop\_Year: which year crop harvested.

Season:

\*Winter, occurring from December to February. ...

\*Summer or pre-monsoon season, lasting from March to May. ...

\*Monsoon or rainy season, lasting from June to September. ...

\* Post-monsoon or autumn season, lasting from October to November.

Crop: crop name

Area: how much area? they harvested.

Production: production amount

cat\_crop: cat\_crop name

**Data Preparation:**

we will transform the data. By getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain.

Next we drop or remove all columns except for the columns that we want to retain.

Finally we drop or remove the rows that have missing values from the data set.

**Model Selection:**

While creating a machine learning model, we need two dataset, one for training and other for testing. But now we have only one. So lets split this in two with a ratio of 80:20. We will also divide the dataframe into feature column and label column.

Here we imported train\_test\_split function of sklearn. Then use it to split the dataset. Also, test\_size = 0.2, it makes the split with 80% as train dataset and 20% as test dataset.

The random\_state parameter seeds random number generator that helps to split the dataset.

The function returns four datasets. Labelled them as train\_x, train\_y, test\_x, test\_y*.* If we see shape of this datasets we can see the split of dataset.

We will use Random Forest Classifier, which fits multiple decision tree to the data. Finally I train the model by passing train\_x, train\_y to the fit method.

Once the model is trained, we need to Test the model. For that we will pass test\_x to the predict method.

\***Lasso Regression**:

Lasso regression is a type of **linear regression**that uses [shrinkage](https://www.statisticshowto.com/shrinkage-estimator/). Shrinkage is where data values are shrunk towards a central point, like the [mean](https://www.statisticshowto.com/mean/). The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of [muticollinearity](https://www.statisticshowto.com/multicollinearity/) or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

The acronym “LASSO”stands  Least Absolute Shrinkageand Selection Operator.

**\*Kernel Ridge:**

Kernel ridge regression is a non-parametric form of ridge regression. The aim is to learn a function in the space induced by the respective kernel kk by minimizing a squared loss with a squared norm regularization term.

The solution can be written in closed form as:

α=(K+τI)−1yα=(K+τI)−1y

where KK is the kernel matrix and αα is the vector of weights in the space induced by the kernel. The learned function can then be evaluated as f(x)=∑Ni=1αik(x,xi)f(x)=∑i=1Nαik(x,xi).

We used stacked averaging models is one of the most powerful methods that is used in machine learning for regression problems. The stacked averaging models comes in the category of the supervised regressor algorithm.

**Analyze and Prediction:**

In the actual dataset, we chose only 6 features :

**State\_Name**: India is a union of States and Union Territories for the purposes of administration, India is divided into 29 States

**District\_Name**: District in india

**Crop\_Year**: which year crop harvested.

**Season**:

\*Winter, occurring from December to February. ...

\*Summer or pre-monsoon season, lasting from March to May. ...

\*Monsoon or rainy season, lasting from June to September. ...

\* Post-monsoon or autumn season, lasting from October to November.

**Crop**: crop name

**Area**: how much area? they harvested.

**Accuracy on test set:**

We got a accuracy of 80.1% on test set.