** TEAM HACKADEMICS**

**ENVISION 2k25**

**ROuND-II**

**Problem Statement :- Crop Yield Prediction**

**Description: Use agricultural datasets (soil quality, weather conditions, crop type) to predict crop yield and recommend agricultural practices for optimization.**

**DATA ANALYSIS, PRE-PROCESSING AND VISUALIZATION**

We follow a step by step process for the same

**STEP-1:- Import all essential libraries.**

We import all the necessary libraries using the **import keyword**

We import libraries like

NumPy-This library is used to work with multidimensional data and matrices.

Pandas-This library is used to create and modify DataFrames and allow us to work with structured data.

Matplotlib-This library is used for basic visualizations.

Seaborn-This library is for visualizations like bar plot ,scatter plot,graphs ,etc.

**STEP-2:-Import .csv (data)**

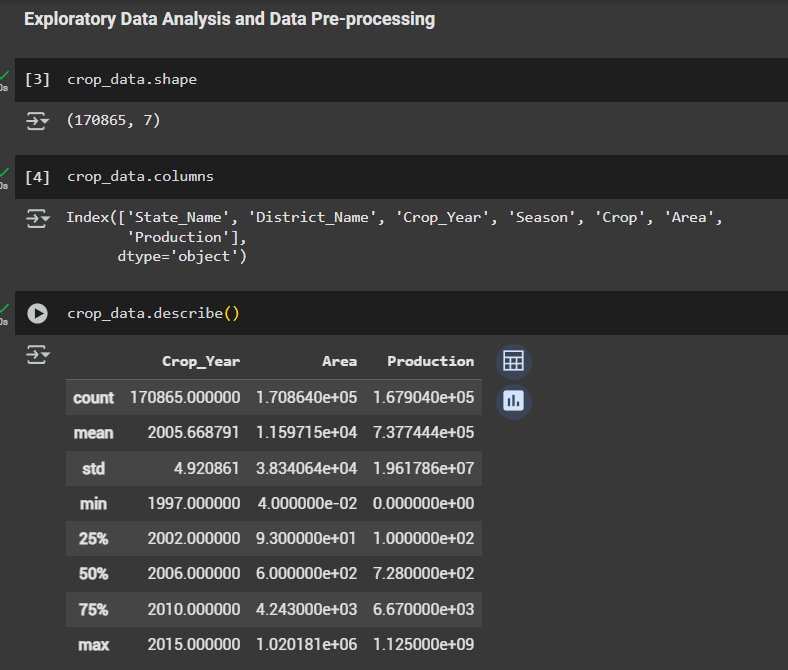
We import data on which we are going to perform the further processes using the **Pandas library and the read\_csv ()** to import data.

**STEP-3:-Perform basic EDA**

EDA Stands for exploratory data Analysis .In EDA we try to identify the underlying patterns and relation ships in the data also we try to analyse the dataset.

The various functions used for EDA are

* Shape()-This function returns the no of rows and columns in the dataset
* Describe()-This function gives the overall description of the dataset which includes count , mean ,std Deviation etc.
* Info()-This function gives information about datatypes in the dataset.
* Column()-this function returns the columns in the dataset.
* Unique()-this returns the unique elements of the array or column.



**STEP-4:-Data Preprocessing**

Data preprocessing is a very important step after performing the EDA

In data preprocessing we actually identify the missing values, outliers ,NULL values or NaN(Not a Number) values and remove the

These values are unimportant to the dataset and the model as they could lead to inefficient and incorrect models.

We perform various steps like

1. Data Cleaning
2. Data transformation
3. Feature Engineering
4. Data integration

**1.DATA CLEANING**

In this process we

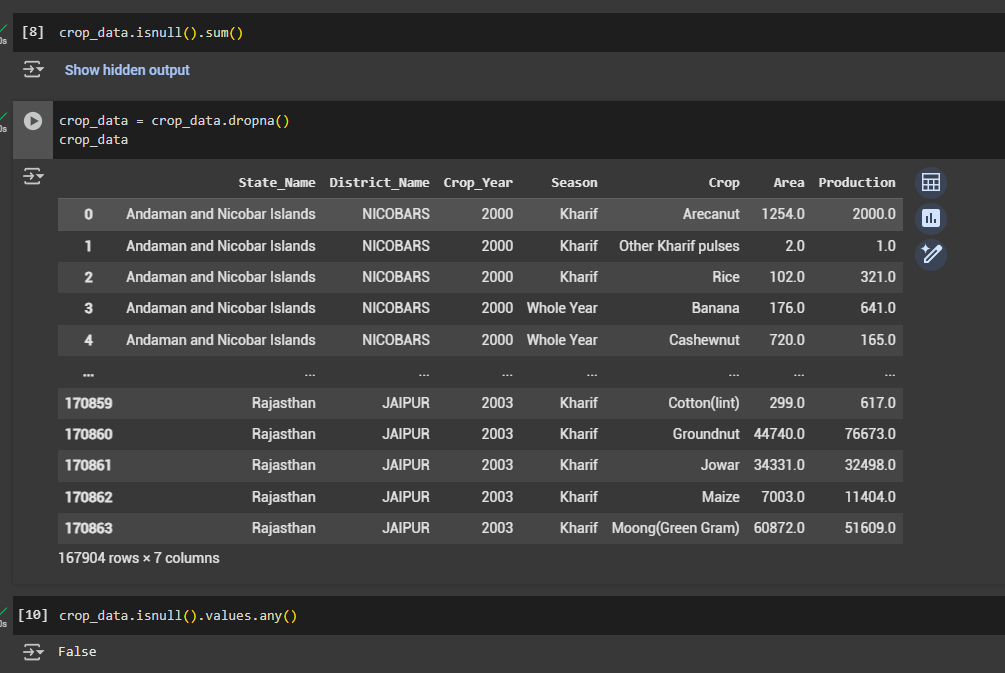
**Handle missing data, remove duplicates, handle outliers**

We use functions like

Isnull()-This identifies missing values.

Isnull.sum()-This gives the count of null values.

Duplicated()-this shows the duplicated values.



**2.Data transformation**

In data transformation we perform scaling ,encoding etc.

**3.Feature Engineering**

In this step we perform

Feature selection

Feature Extraction

Feature Reduction

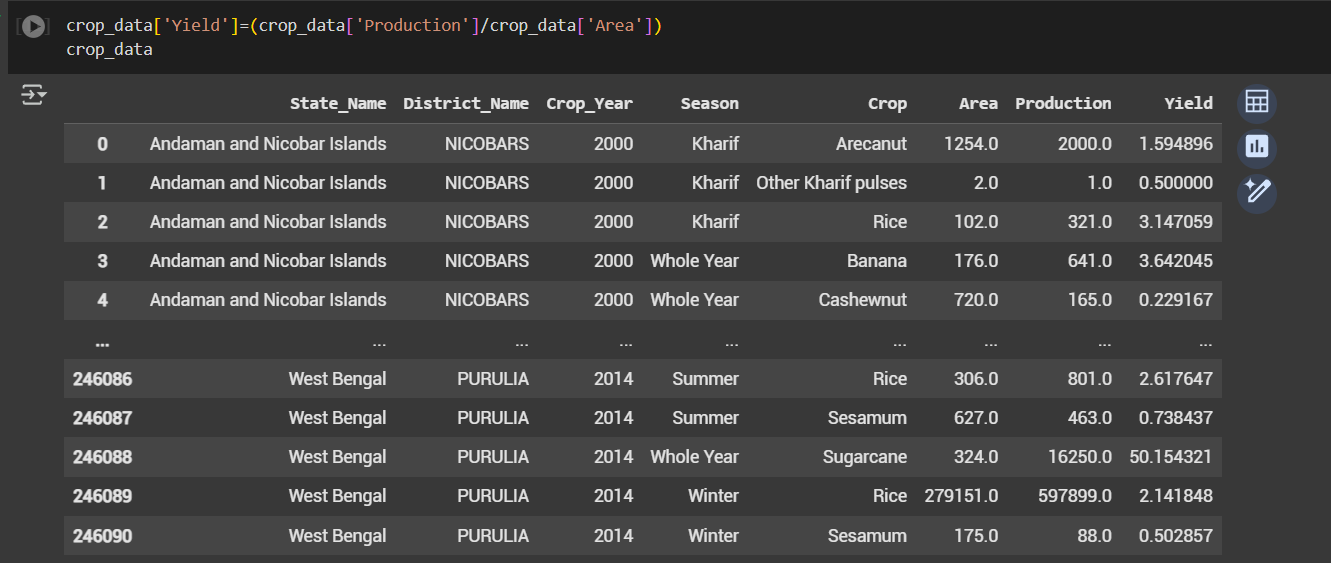
For which we use methods like PCA and LDA.

**4.Data Integration**

In this step we add or remove columns from our dataset depending on our need and necessity.

On performing EDA on our dataset we have seen that we don’t have any feature called Yield In our dataset **.Therefore we add new column for yield in our dataset.**

**WE USE THE FORMULA  
yield=production/area**

****

**STEP-5:-Data Visualization**

In this step we use the matplotlib and seaborn library to add visualizations to our dataset.

In our dataset we have implemented plots like

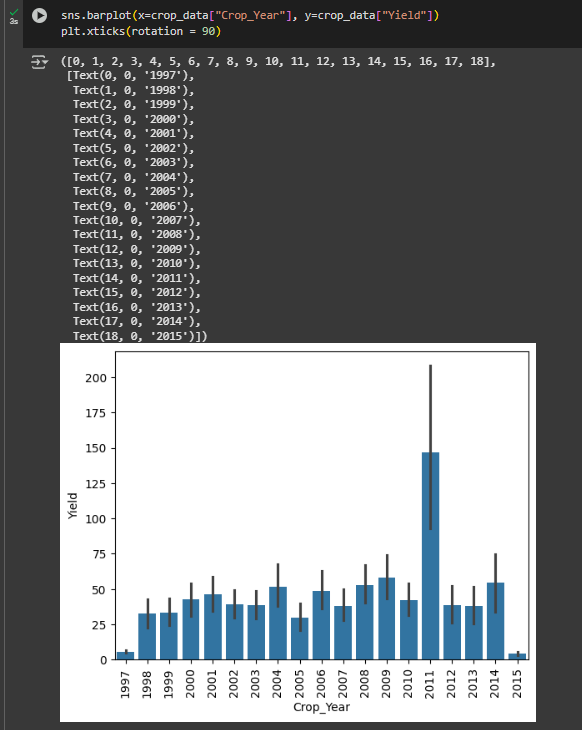
* Barplot-It is a rectangular bar that is used to represent data categories.
* Lineplot-used to display information as a series of data points connected by a straight line
* Jointplot- It is used to explore relationships between bivariate data as well as their distributions

By using visualizations we can gather various insights on our dataset

We can obtain pattern, comparisons, relationships, trends, .etc from our dataset.

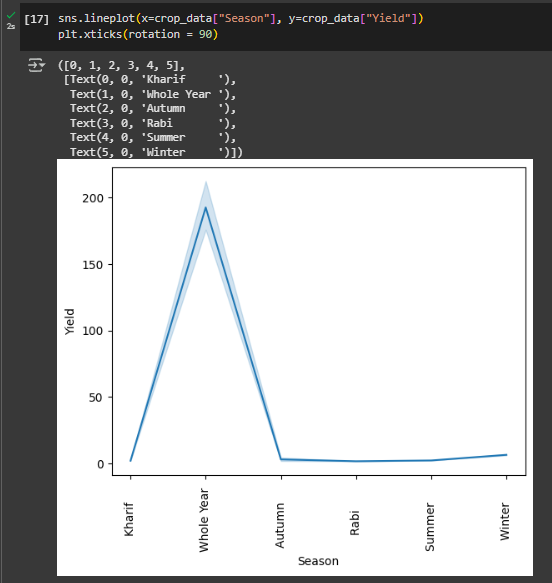
**We have implemented various visualizations to our dataset**

**Like**

****

**This bar plot shows year wise yield of the crops.**

We can conclude from this that the yield was maximum in the year 2011



This line plot shows that the crops that produce the most yield are the crops that are cultivated for the whole year.

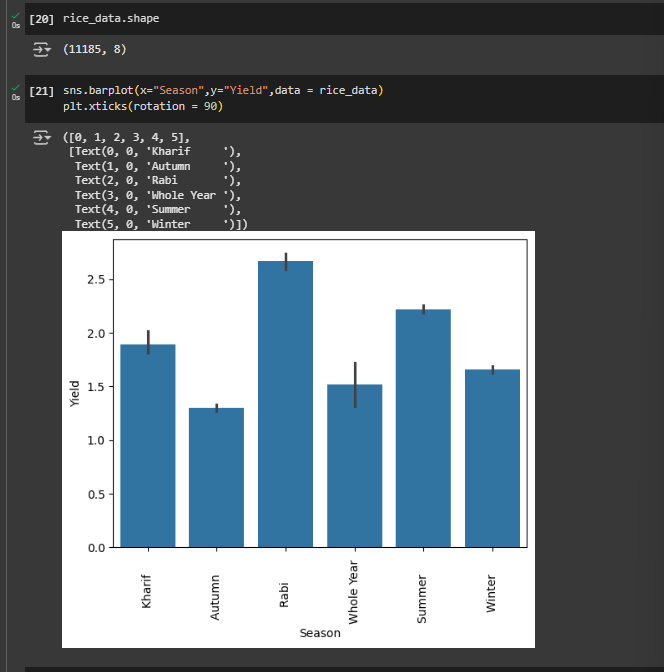
**CONCLUSIONS :**

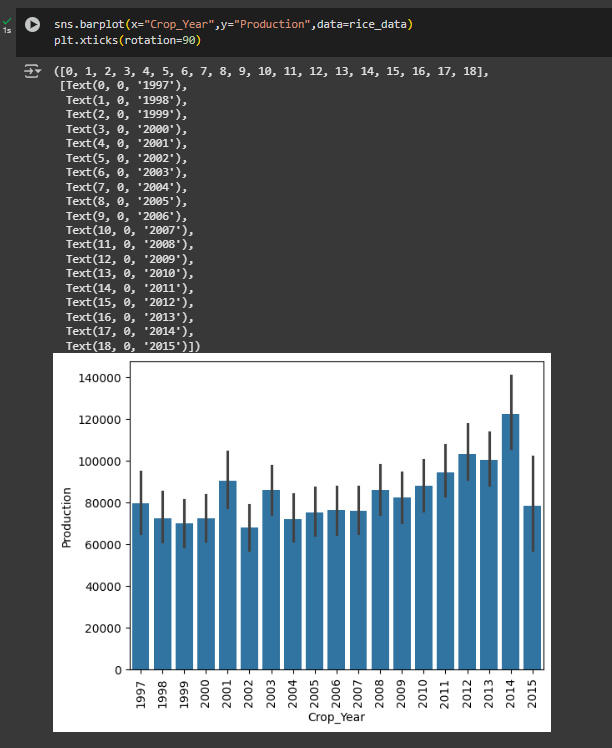
**Conclusions Obtained from the dataset:**

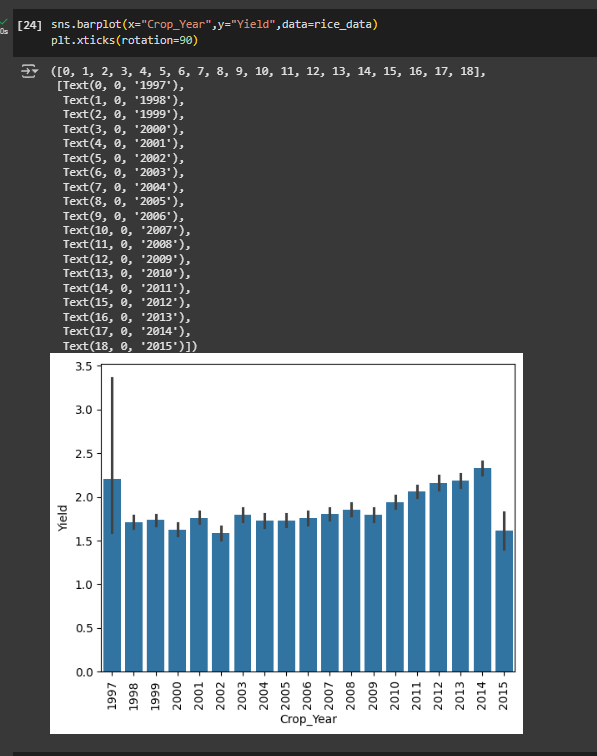
1. **Kerela is the largest producer of crop in India.**
2. **Production per unit area (Yield ) of Puducherry is maximum.**
3. **In the year 2011, crop yield was maximum.**

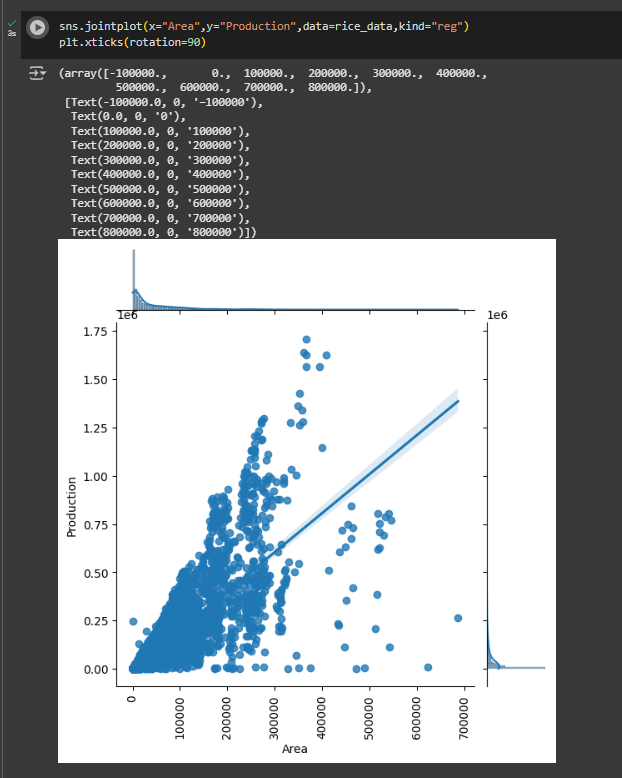
**After this we performed analysis on each cultivated crops.**

**For ex: Rice**

****

****

****

****

**AFTER ANALYSING THE DATA USING VISUALS WE CAN CONCLUDE**

**RICE**

1. **Rice yield is maximum in Rabi season.**
2. **Rice yield is maximum in Chandigarh.**
3. **Rice yield has been growing a little from the year 2009 to 2014.**

**SIMILARLY WE CAN ANALYSE THE REMAINING CROPS**

**WHEAT**

1. **Wheat yield is maximum in Rabi season.**
2. **Wheat Production is maximum in Punjab.**
3. **Wheat yield has declined after 2013.**

**COCUNUT**

1. **Andhra Pradesh is the largest producing coconut states.**
2. **Production per unit area is higher in Mizoram and Sikkim.**
3. **Coconut yield is decreasing in the year 2012 to 2015.**

**POTATO**

1. **Potato is a Rabi crop.**
2. **West Bengal is the largest producer of potatoes.**

**ONION**

1. **Onion is a Rabi crop.**
2. **Gujrat and Maharashtra are the major onion-producing states.**

**MAIZE**

1. **Maize is produced in the autumn season**
2. **Telangana is the major maize-producing states.**
3. **There was a sudden decline in maize production from the year 2000.**

**P.T.O**

**SOLUTION PLAN**

After thorough EDA, Data Analysis and Visualization we can come to conclusion that for a prediction model to be created we can use the following algorithms.

* Linear Regression
* Random forest
* Decision Tree

We select the algorithm with **the highest accuracy** by using the Evaluation matrices

Before applying the algorithms we split the data into training and test data

Then we train the model and apply the following evaluation metrics.

* Accuracy
* R2 Score
* Adjusted R^2
* Standard Deviation

After this step we will perform **CROSS VALIDATION**  to find the **Best Parameter for Hyper Parameter Tuning**

IN the final step we PICKLE the model and save it for further deployment.

**FOR RECOMMENDATION MODEL**

The initial steps to create a recommendation model is similar to the creation of a prediction model.

The only difference is that the same dataset used for prediction cant be used for recommendation .We need new dataset with more features for the same.

**STEP-1:- Import all essential libraries.**

* NumPy
* Pandas
* Matplotlib
* Seaborn

**STEP-2:-Import .csv (data)**

**STEP-3:-Perform basic EDA**

The various functions used for EDA are

* Shape()-This function returns the no of rows and columns in the dataset
* Describe()-This function gives the overall description of the dataset which includes count , mean ,std Deviation etc.
* Info()-This function gives information about datatypes in the dataset.
* Column()-this function returns the columns in the dataset.
* Unique()-this returns the unique elements of the array or column.

**STEP-4:-Data Preprocessing**

We perform various steps like

1. Data Cleaning
2. Data transformation
3. Feature Engineering
4. Data integration

**1.DATA CLEANING**

In this process we

**Handle missing data, remove duplicates, handle outliers**

We use functions like

Isnull()-This identifies missing values.

Isnull.sum()-This gives the count of null values.

Duplicated()-this shows the duplicated values.

**2.Data transformation**

In data transformation we perform scaling ,encoding etc.

**3.Feature Engineering**

In this step we perform

Feature selection

Feature Extraction

Feature Reduction

For which we use methods like PCA and LDA.

**4.Data Integration**

In this step we add or remove columns from our dataset depending on our need and necessity.

In this case we drop a column

**STEP-5:-Data Visualization**

Visualization using

* Hisplot
* distplot

By using we can gather various insights on our dataset

We can obtain pattern, comparisons, relationships, trends, .etc from our dataset.

After thorough EDA, Data Analysis and Visualization we can come to conclusion that for a prediction model to be created we can use the following algorithms.

* Logistic Regression
* SVC(Support Vector Classifier)
* Decision Tree
* Random Forest

We select the algorithm with **the highest accuracy** by using the Evaluation matrices

Before applying the algorithms we split the data into training and test data

Then we train the model and apply the following evaluation metrics.

* Accuracy
* R2 Score

After this step we will perform **CROSS VALIDATION**  to find the **Best Parameter for Hyper Parameter Tuning**

IN the final step we PICKLE the model and save it for further deployment