CROP YEILD PREDICTION

By
Ayush Taparia
Aashirwad Wardhan
Pritam Chowdhury
Aman Singh
Oishik Mukhopadhyay

B.Tech
CSE (Artificial Intelligence and Machine Learning)

CONTENT

- Aim
- Objectives
- Problem Analysis
- Solutions
- Assumptions
- Data Preprocessing
- Exploratory Data Analysis
- Results
- Performance of the Algorithms
- Observations
- Visualization
- Conclusion

WHAT IS OUR MODEL?
AND
WHY IS OUR MODEL?

OUR AIM

To design a precise and effective crop yield estimation model by utilizing weather data, soil information, and nutrient analysis.

OBJECTIVES

- The core objective of crop yield estimation is to achieve higher agricultural crop production .
- Crop Yield Prediction (CYP) including supporting decisions on what crops to grow and what to do during the growing season of the crops.
- Machine learning techniques will enhance the productivity of the field along with the reduction in the input efforts of the farmers

OUR PROBLEMS

- Data Quality and Quantity
- Feature Selection
- Model Complexity
- Temporal and Spatial Variability
- Climate and Weather Conditions
- Model Evaluation
- Scale and Accessibility
- Collaboration and Stakeholder Involvement
- Soil Quality and Nutrient Management

OUR SOLUTIONS

- Ensure that you have high-quality and sufficient data. Collect data from reliable sources.
- Conduct a thorough analysis to identify the most relevant features affecting crop yield. Use domain knowledge and statistical methods to select the best features
- Choose a model that suits your dataset size and complexity.
- Regularization techniques can help prevent overfitting, and cross-validation can be used to assess model performance.
- Consider temporal and spatial factors in your model.
- Use time-series analysis for temporal variations and geospatial data for spatial variations.
- Integrate weather data into your model. Machine learning models can leverage historical weather patterns to make predictions.
- Include soil data in your model. Analyze soil nutrient levels and incorporate this information into the prediction model.
- Factor in information about agricultural practices. Understand the farming methods used in the region and incorporate this knowledge into the model.

ASSUMPTIONS

- **Stationarity of Variables:** Assumption of stationary weather and soil conditions; sudden unpredictable changes not explicitly considered.
- **Homogeneous Crop Growth:** Assumes uniform conditions; localized variations like microclimates not explicitly incorporated.
- **Consistency in Data Sources:** Assumes consistency and accuracy in satellite imagery and historical records; discrepancies not explicitly addressed.
- **Linear Relationships**: Assumes linear relationships between input variables and crop yield; may not capture non-linear interactions.
- **Uniform Crop Management Practices:** Assumes consistent crop management practices; variations in irrigation and fertilization not explicitly considered.
- **Predictive Power of Historical Data:** Assumes reliability of historical yield data for future predictions; may not account for novel influences on yields.

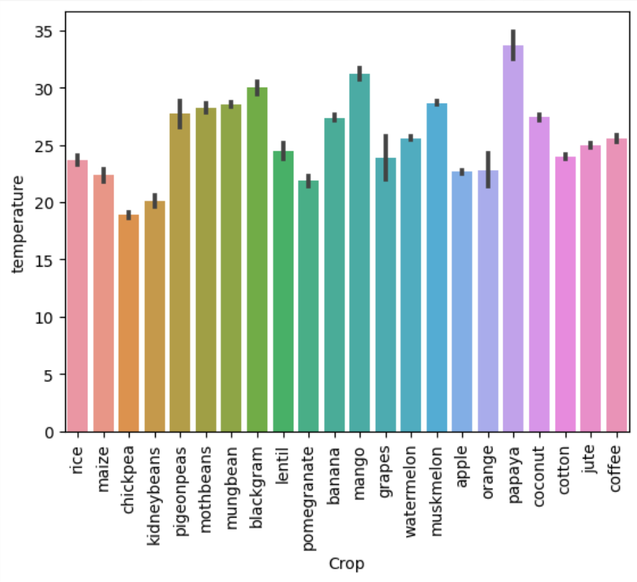
DATA PREPROCESSING

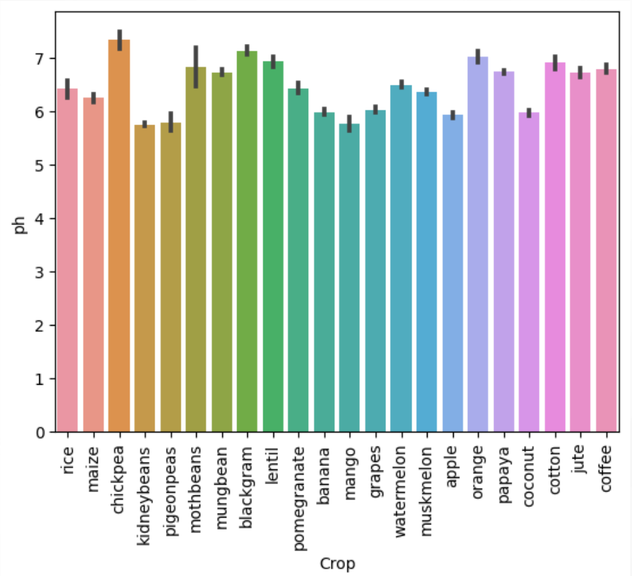
Data preprocessing is a vital process that involves preparing raw data and transforming it into a format suitable for a machine learning model. It serves as the initial step in developing an effective machine learning model.

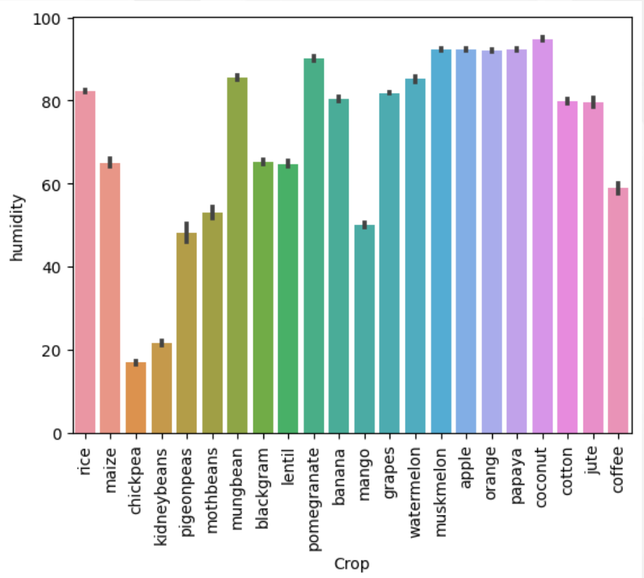
- **Data Cleaning -** From duplicates and outliers to missing numbers, it fixes them all.
- **Data Integration-** This process integrates (merges) information extracted from multiple sources to outline and create a single dataset.
- **Data Transformation-** This process entails putting the data in a format that will allow for analysis. Normalization, standardization, and discretisation are common data transformation procedures.
- **Data Reduction-** Data reduction is the process of lowering the dataset's size while maintaining crucial information.

EXPLORATORY DATA ANALYSIS

After performing cleaning and normalizing the datasets we performed Exploratory Data Analysis to get a better understanding of the agricultural scenario of India. Through plotting and analyzing the graph we came to know a lot of useful information and could understand the dataset better.







RESULTS

After our Exploratory Analysis of the dataset we arrived at various facts about the Indian agricultural scenario as mentioned in observation. We trained 4-5 Machine Learning Models on the final dataset. We have calculated the R2 score and results are as follows:

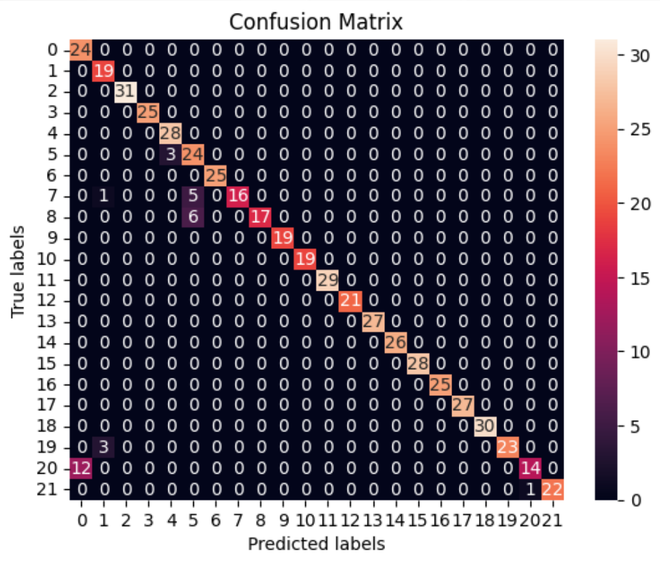
PERFORMANCE OF THE ALGORITHMS

- **Naive Bayes Classification**
- **Decision Tree Classification**
- **Random Forest Classification**
- **KNN Classifier**
- **Gradient Boosting**

ALGORITHMS

Naive Bayes Algorithm

It is a classification technique based on Bayes' Theorem with an independence assumption among predictors. It belongs to the family of generative learning algorithms, which means that it models the distribution of inputs for a given class or category.

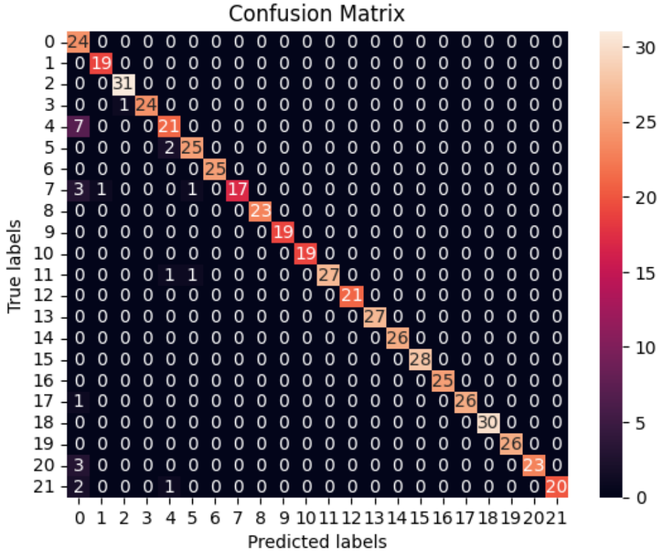


Accuracy of the Algorithm Naive Bayes Classification was 0.9709090909090909. Given below is the confusion matrix of Naive Bayes Classification algorithm.

ALGORITHMS

Decision Tree Algorithm

The Decision Tree algorithm embodies a hierarchical structure wherein nodes serve as decision points, contingent upon specific features ranging from weather conditions and soil attributes to historical yield data.

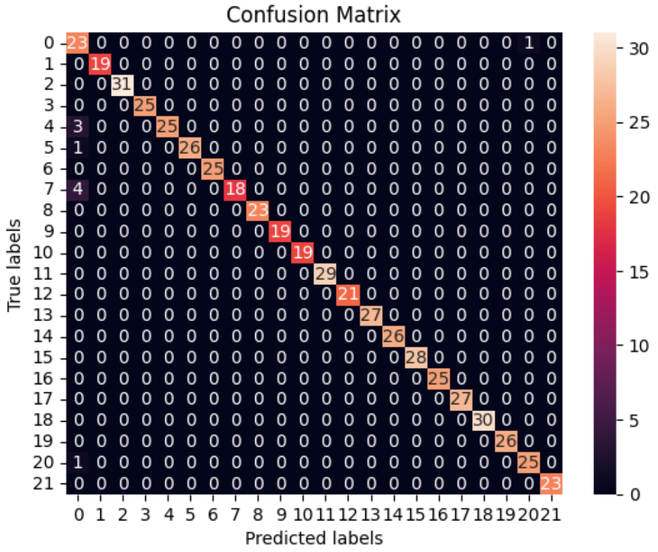


Accuracy of the Algorithm Decision Tree Classification was 0.9563636363636364. Given below is the confusion matrix of Decision Tree Classification algorithm.

ALGORITHMS

Random Forest Algorithm

At its core, the Random Forest algorithm harnesses the collective wisdom of multiple decision trees within an ensemble. Each decision tree is constructed using a random subset of the training data and features, mitigating the risk of overfitting and enhancing the model's ability to generalize to diverse agricultural conditions.

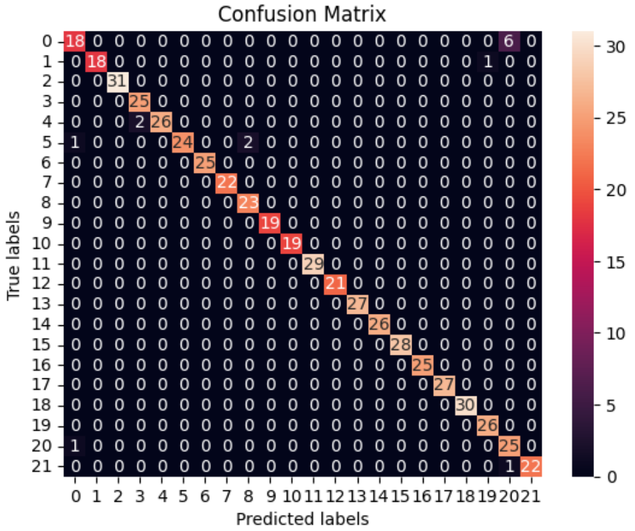


Accuracy of the Algorithm Random Forest Classification was 0.98. Given below is the confusion matrix of Random Forest Classification algorithm.

ALGORITHMS

K-Nearest Neighbors Algorithm

KNN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. In order to determine which data points are closest to a given query point, the distance between the query point and the other data points will need to be calculated.

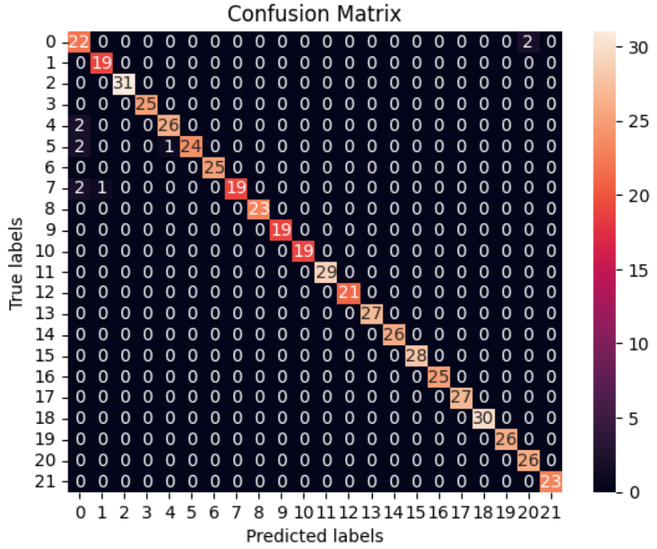


Accuracy of the Algorithm KNN Classifier was 0.9745454545454545. Given below is the confusion matrix of KNN Classifier algorithm.

ALGORITHMS

Gradient Boosting Algorithm

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models. Its key innovation lies in the strategic incorporation of regularization techniques and parallel processing, enhancing predictive accuracy and computational efficiency.



Accuracy of the Algorithm Gradient Boosting was 0.9818181818181818. Given below is the confusion matrix of Gradient Boosting algorithm.

OBSERVATIONS

- The Machine Learning Algorithms Gradient Boost and Random forest Classification had the highest accuracy of about 84% and 84%-76% respectively.
- Ridge Regression and Gradient Boosting Regression performed fairly well with accuracy of about 54% and 66% respectively.
- The study demonstrated the efficacy of CYPA by training and verifying five models using optimal hyper-parameter settings for each machine learning technique.
- Decision Tree Classifier, Random Forest Classifier, and KNN Classifier, the score is equal to 0.95, 0.98, and 0.97, respectively.
- Accuracy of the Algorithm Naive Bayes Classification was 0.97.Accuracy of the Algorithm Gradient Boosting was 0.98.

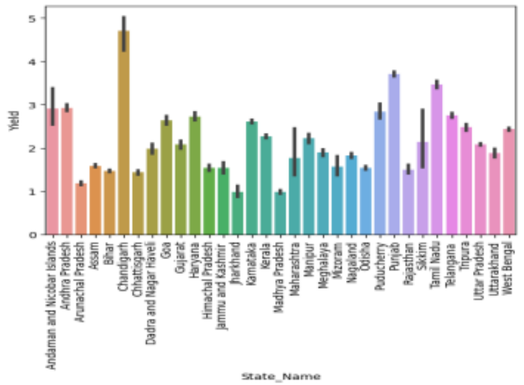
VISUALIZATION

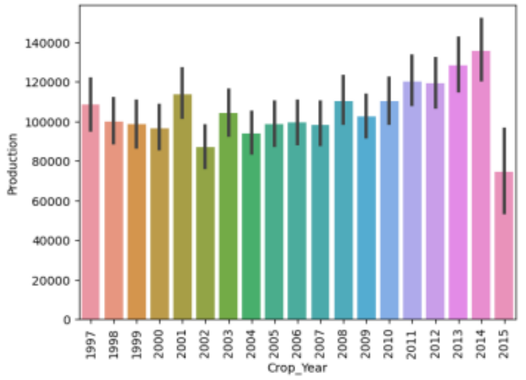
Analyzing each type of Crop

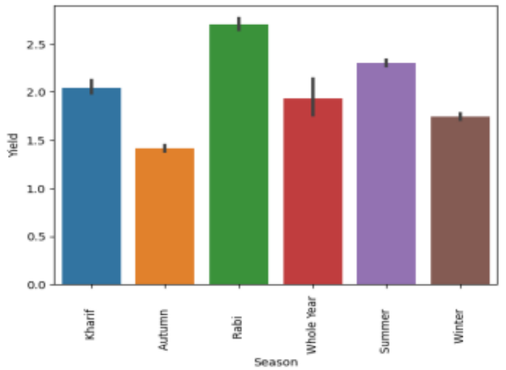
- **Rice**

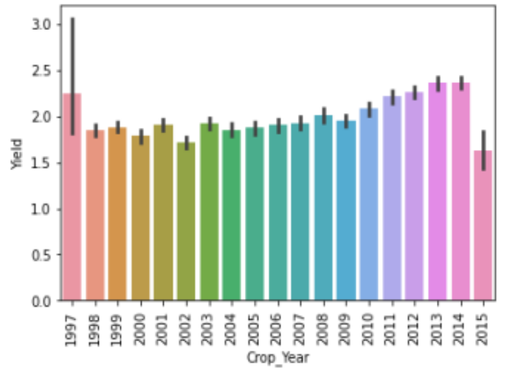
Observations Obtained:

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.







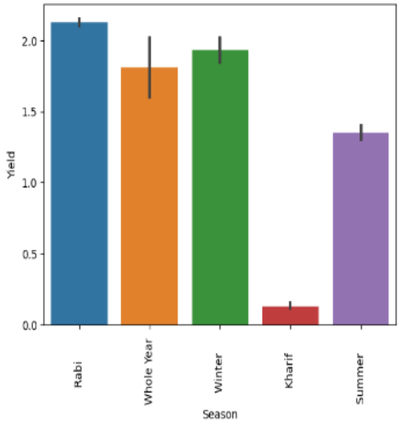


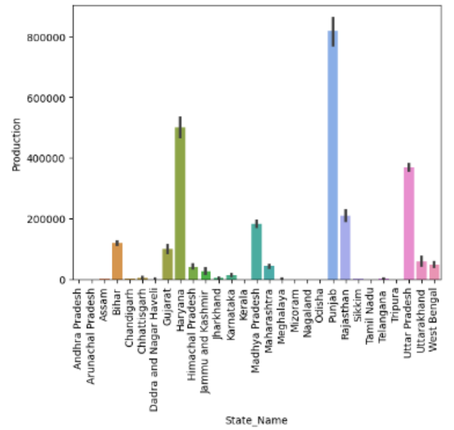
VISUALIZATION

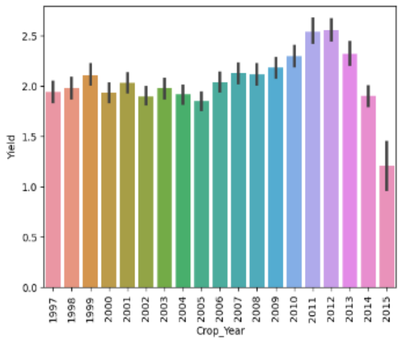
- **Wheat**

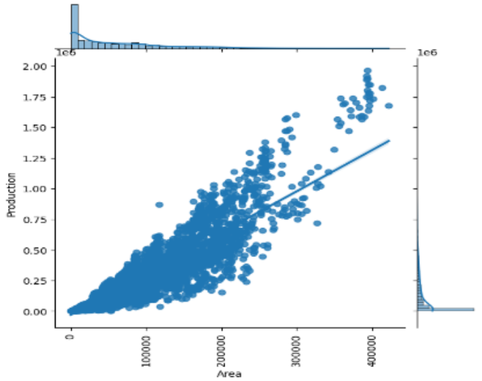
Observations Obtained:

1. Wheat yield is maximum in Rabi season.
2. Wheat yield is maximum in Punjab.
3. Wheat yield is decreasing in the year 2012 to 2015







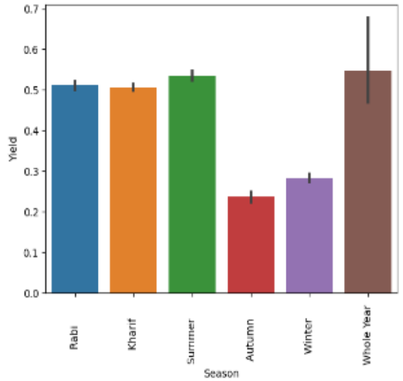


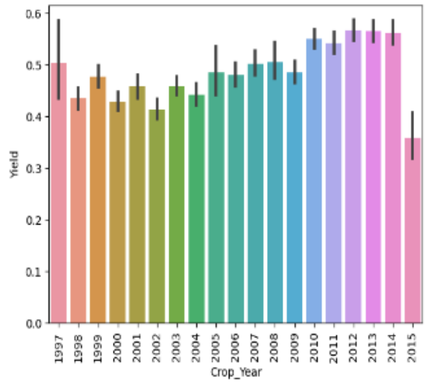
VISUALIZATION

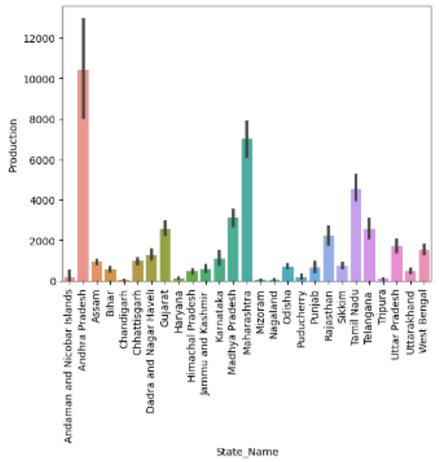
- **Coconut**

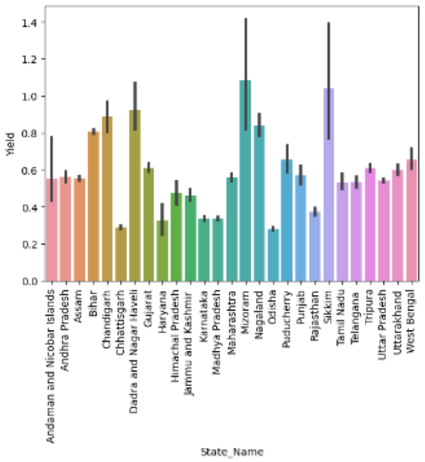
Observations Obtained:

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







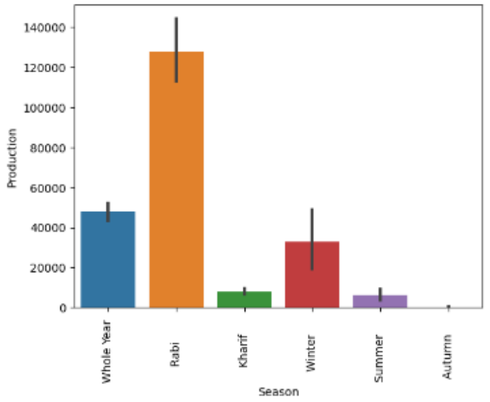


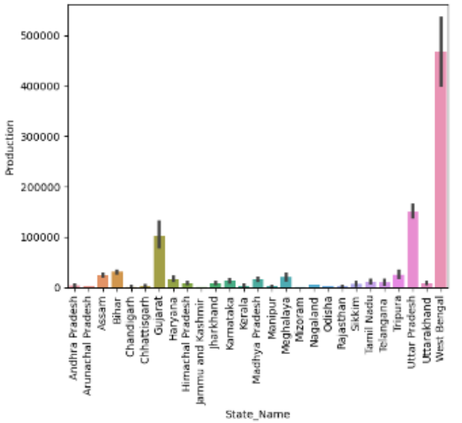
VISUALIZATION

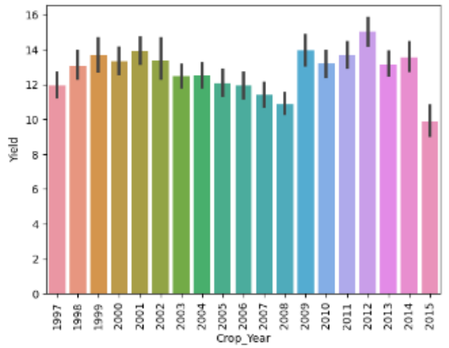
- **Potato**

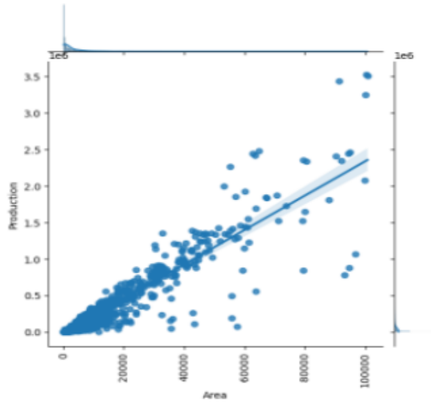
Observations Obtained:

1. Potatoes are a Rabi crop.
2. West Bengal is the largest producer of potatoes.
3. Potato yield is decreasing in the year 2001 to 2008.







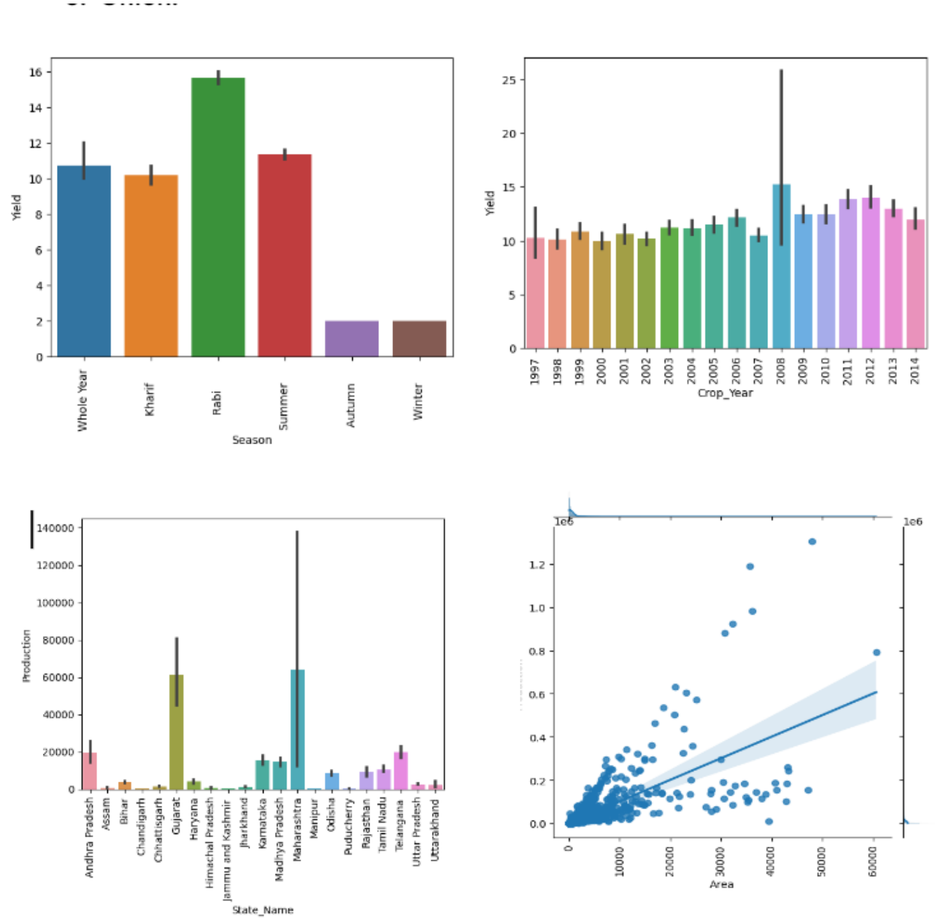


VISUALIZATION

- **Onion**

Observations Obtained:

1. Onion is a Rabi crop.
2. Gujarat and Maharashtra are the major onion-producing states.
3. Onion yield is decreasing in the year 2012 to 2014.

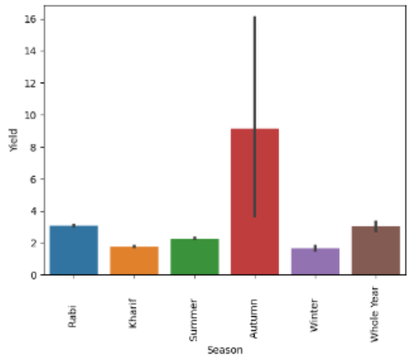


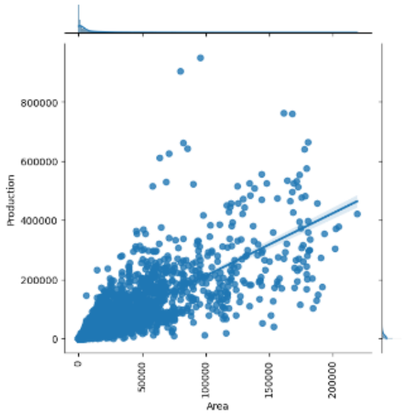
VISUALIZATION

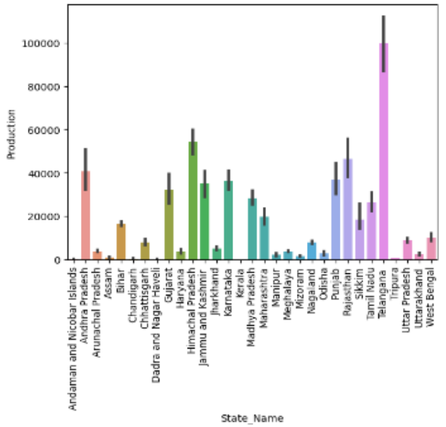
- **Maize**

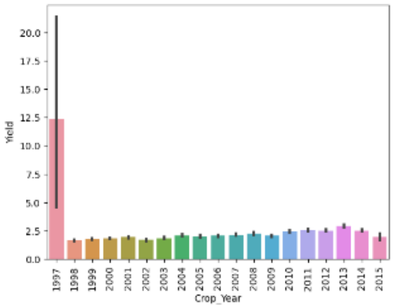
Observations Obtained:

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









CONCLUSION

The proposed Crop Yield Prediction Algorithm (CYPA) incorporates multiple data sources such as climate, weather, agricultural yield, and chemical data to anticipate annual crop yields by policymakers and farmers in their country. The study demonstrated the efficacy of CYPA by training and verifying five models using optimal hyper-parameter settings for each machine learning technique. The results indicate that CYPA can achieve high accuracy in predicting crop yields, as demonstrated by the scores of Decision Tree Classifier, Random Forest Classifier, and KNN Classifier. Additionally, the paper introduces a new algorithm based on active learning that can enhance CYPA's performance by reducing the number of labeled data needed for training. Incorporating active learning into CYPA can improve the efficiency and accuracy of crop yield prediction, thereby enhancing decision-making at international, regional, and local levels. Overall, this study highlights the potential of IoT and machine learning techniques in addressing the critical challenge of predicting crop yields, thereby facilitating informed decision-making for policymakers and farmers alike. When utilizing Decision Tree Classifier, Random Forest Classifier, and KNN Classifier, the score is equal to 0.95, 0.98, and 0.97, respectively.

THANK YOU!