



Abstract

This research paper delves into the application of machine learning techniques for the accurate prediction of crop yields, addressing the challenges posed by climate variability and the increasing demand for food production. The study employs a diverse set of machine learning algorithms, including regression models and ensemble methods, to analyze a comprehensive dataset comprising meteorological variables, soil characteristics, and historical crop yield data.

Machine Learning is extremely important for predicting crop yield as it helps us to understand the best crops which can grow under the conditions provided. On the basis of this we also built a recommendation system which will be used to suggest the crops to grow under specific conditions for optimal use of the conditions around us.

A lot of factors influence the yield of crops including humidity, minerals present in soil etc. Using Machine Learning will help to enhance the economy as we will be able to choose the optimal crops for the specific conditions.

The methodology involves preprocessing and feature engineering to extract meaningful patterns from the input data, followed by a systematic evaluation of various machine learning models.

Our main aim which was to help predict the yield of crops under given circumstances was achieved by using different algorithms. Our research has helped us conclude that out of all the algorithms we used, the random forest algorithm stood out.

Introduction

Agriculture, as the cornerstone of human civilization, has played a pivotal role in shaping societies and economies across the globe. The demand for food, feed, and fiber is reaching unprecedented levels with a rising population. This surge in demand, coupled with the challenges posed by climate change, resource scarcity, and the need for sustainable practices, has compelled the agricultural sector to explore innovative solutions. One such solution that has gained substantial traction in recent years is the integration of machine learning (ML) techniques into agricultural processes.

Machine learning, a subset of artificial intelligence, offers a paradigm shift in addressing the multifaceted challenges faced by modern agriculture. By leveraging algorithms that can learn patterns from data, ML empowers farmers, researchers, and policymakers with tools to make informed decisions. The ability to process vast datasets, identify trends, and predict outcomes positions ML as a transformative force in optimizing agricultural practices.

The models must be trained with datasets that have some prior knowledge to the objective we want to achieve.



Novelty

Our project plans to seamlessly integrate financial data, real-time information, and on-site soil condition testing to provide a comprehensive and dynamic solution for farmers. Unlike traditional models, our approach goes beyond yield prediction alone. By incorporating financial data related to crops for the given year, we not only forecast yield but also calculate estimated revenue, offering farmers a holistic view of their potential earnings. Additionally, our system harnesses the power of real-time data through various APIs, eliminating the need for farmers to manually input information. Leveraging on-site soil condition testing further refines predictions, ensuring that recommendations are tailored to the unique characteristics of each field. What sets our project apart is our commitment to inclusivity; recognizing the challenges faced by farmers in remote areas without internet access, we go the extra mile by providing physical reports. This tangible output empowers every farmer, regardless of their technological resources, to make informed decisions and optimize their agricultural practices. In this way, our project pioneers a new era of precision agriculture, where technology becomes a practical and accessible tool for farmers across diverse landscapes.

System Analysis

Python 3.8.5(Jupyter Notebook): Python is the coding language used as the platform for machine learning analysis. Jupyter Notebooks illustrates the analysis process and gives out the needed result.

Visual Studio Code: With built-in tools for web development, Visual Studio Code (VS Code) is well-suited for crafting frontend code for associated websites. As the ideal IDE for Python-based machine learning, it seamlessly supports Jupyter Notebooks and key ML libraries. Its user-friendly interface, integrated debugging, and version control make it a versatile and efficient choice for collaborative and reproducible ML workflows.

Python Flask Framework (Version 2.0.1): Flask is a micro framework in python. Flask is based on WSGI(Web Server Gateway Interface) toolkit and Jinja2 template engine. In this paper flask is used as the back-end framework for building the application. It is the collection of modules and libraries that helps the developer to write applications without writing the low-level codes such as protocols, thread management, etc.

Methodology

A) Crop Recommendation System:

1.1. Collecting the Raw Data

The practice of cumulating and scrutinizing data from different sources is known as data collection. Data collection is a way to keep track of past occurrences so that one can utilize data analysis to detect repetitive patterns. The 'Crop Recommendation' dataset is collected from the Kaggle website. The dataset takes into account 22 different crops as class labels and 7 features-

(i) Nitrogen content ratio

(N) (ii) Phosphorus content ratio (P) (iii) Potassium content ratio (K) in the soil, (iv)

Temperature expressed in degree Celsius (v) Percentage of Relative Humidity (vi) ph value and

(vii) Rainfall measured in millimeters.



Shri Vile Parle Kelavani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

(Autonomous College Affiliated to the University of Mumbai)

NAAC Accredited with "A" Grade (CGPA : 3.18)



	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice
...
2195	107	34	32	26.774637	66.413269	6.780064	177.774507	coffee
2196	99	15	27	27.417112	56.636362	6.086922	127.924610	coffee

Dataset Sample

1.1. Data Preprocessing

The process of modifying raw data into a form that analysts and data scientists can use in machine learning algorithms to find insights or forecast outcomes is called Data preprocessing. In this project, the data processing method is to find missing values. Getting every data point for every record in a dataset is tough. Empty cells, values like null or a specific character, such as a question mark, might all indicate that data is missing. The dataset used in the project didn't have any missing values.

1.2. Train and Test Split

It is a process of splitting the dataset into a training dataset and testing dataset using `train_test_split()` method of scikit learn module. 2200 data in the dataset has been divided as 80% of a dataset into training dataset-1760 and 20% of a dataset into testing dataset-440 data.

1.3. Fitting the model



Modifying the model's parameters to increase accuracy is referred to as fitting. To construct a machine learning model, an algorithm is performed on data for which the target variable is known. The model's accuracy is determined by comparing the model's outputs to the target variable's actual, observed values. Model fitting is the ability of a machine learning model to generalize data comparable to that with which it was trained. When given unknown inputs, a good model fit refers to a model that properly approximates the output.

1.4. Checking the score over a training dataset

Scoring, often known as prediction, is the act of creating values from new input data using a trained machine learning model. Using `model.score()` method calculating the score of each model over a training dataset shows how well the model has learned.

1.5. Predicting the model

When forecasting the likelihood of a specific result, "prediction" refers to the outcome of an algorithm after it has been trained on a previous dataset and applied to new data. Predicting the model using `predict()` method using test feature dataset. It has given the output as an array of predicted values.

Models	Model Accuracy
 DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING (Autonomous College Affiliated to the University of Mumbai) NAAC Accredited with "A" Grade (CGPA : 3.18)	
Logistic Regression	96.36
Naive Bayes	99.54
Support Vector Machine	96.81
K-Nearest Neighbors	95.90
Decision Tree	98.63
Random Forest	99.31
Bagging	98.86
AdaBoost	14.09
Gradient Boosting	98.18
Extra Trees	92.04

Machine learning classifiers used for accuracy comparison and prediction were Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbours, Decision Tree, Random Forest, Bagging, AdaBoost, Gradient Boosting and Extra Trees. These ten classifiers were trained on the dataset and the model accuracy was calculated. Of the ten classifiers used, Naïve Bayes resulted in highest accuracy and was used as the midway to predict the crop that can be grown on a particular location at the respective time.

Website Application

A website has been developed to query the results of machine learning analysis. The pages were written using HTML, CSS, Bootstrap and JavaScript language and later the ML model was deployed on the website using Python Flask framework.

The website has a simple, easy-to-use interface requiring only few taps to retrieve desired results. Just only giving the district location, soil and climatic conditions of the field the website gives the name of right crop to be grown there. Which is further used to calculate yield.

By accessing the user entered details, website queries the machine learning analysis. The retrieved weather and soil data gets acquired by machine learning classifier to predict the crop and calculate the yield. The output is then fetched by the server to portray the result in application.

The main activities in the application were account creation, explore_system and results_fetch. The account_creation helps the user to actively interact with application interface. The user fill's the fields in the explore_system to move onto the results activity. The retrieved data passed to machine learning model and crop is recommended which can be further used to predict the yield.



Crop Recommendation System

Nitrogen

Enter Nitrogen

Phosphorus

Enter Phosphorus

Potassium

Enter Potassium

Temperature

Enter Temperature in °C

Humidity

Enter Humidity in %

pH

Enter pH value

Rainfall

Enter Rainfall in mm

Get Recommendation

Yield Forecast System

Quantifying Weather Impacts to Predict Crop Yield

Enter your details

Crop Name

Choose...

District Name

1234 Main St

Country

Choose...

State

Choose...

Pincode

Temperature in Celcius

Wind speed

Humidity

Season Name

Choose...

Soil Type

Choose...

Area(where crop is grown)

Content in soil

Nitrogen:

Nitrogen

Phosphorous:

Phosphorous

Potassium:

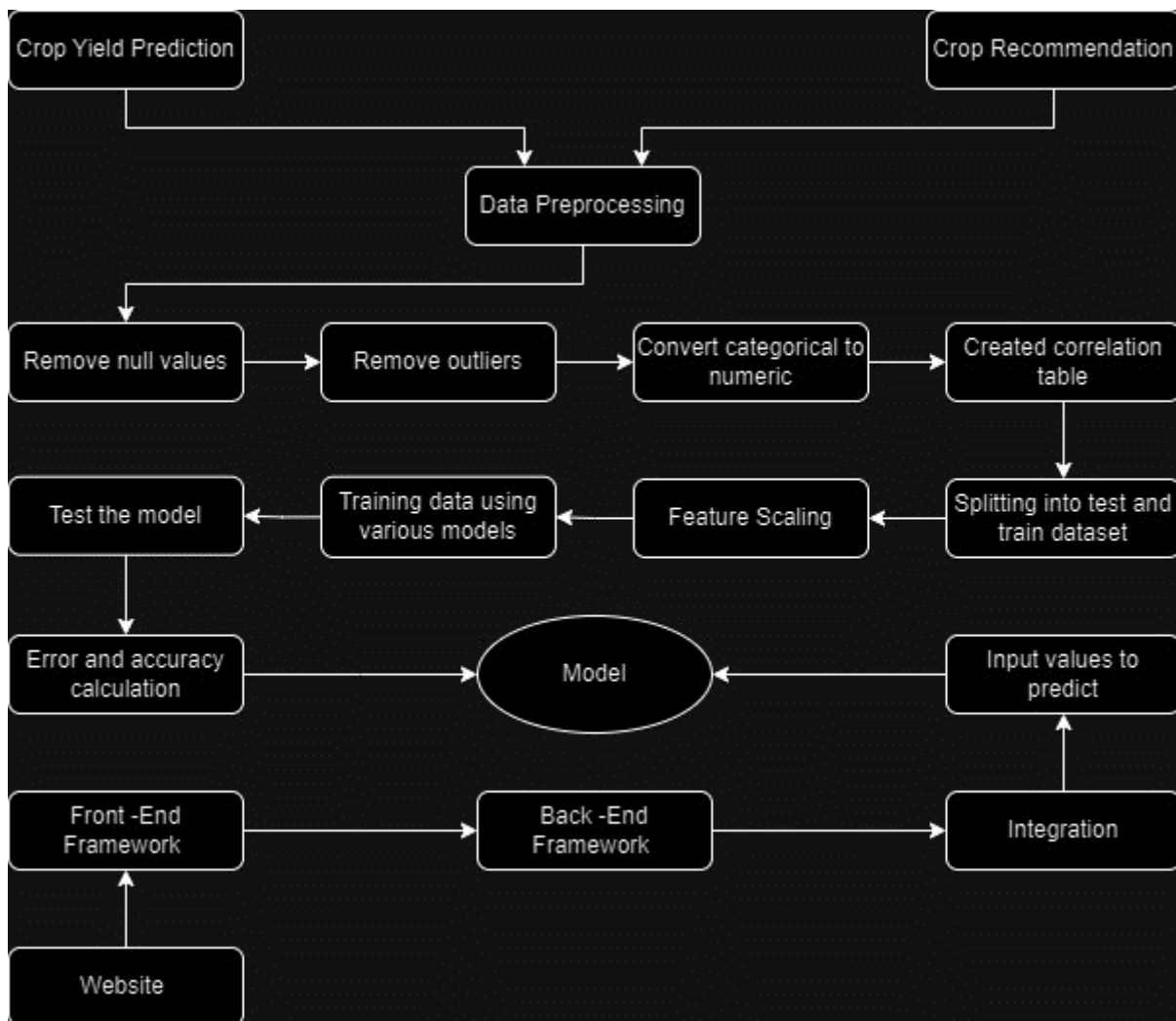
Potassium

Predicted Yield is:

Result

Predict Yield

System Architecture



PREDICTION MODEL METHODOLOGY

B) Yield Prediction

PREPROCESSING:

The original dataset contained data about soil, weather conditions and production of various crops over many regions. We decided to include areas within Maharashtra. First the non desirable entries were removed and only entries within Maharashtra were taken. The features present in the dataset were state, district, crop, area, temperature, wind speed , pressure, humidity, soil type, Potassium, Sodium and Phosphorus levels in soil , and finally net production.

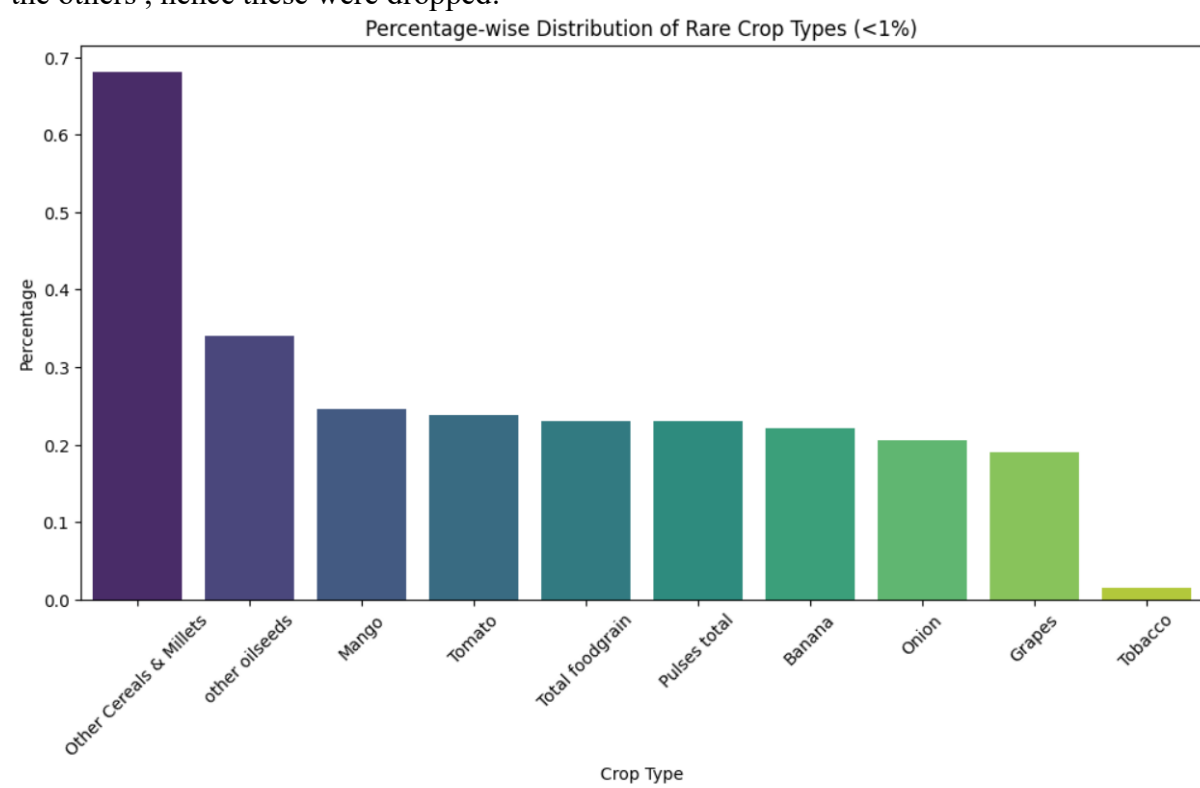
We wanted to estimate how much yield a piece of land would produce based on the weather, soil conditions, area and crop type. We made a new variable called yield that is production upon area



, so that the potential yield could be estimated irrespective of area.

After the changes were made to the dataset, it was imported into a Jupyter notebook to begin further analysis. Missing values needed to be handled and only around 200 data entries contained missing values out of a total of 12300. Hence it would not make a significant difference on the model so we simply dropped these values.

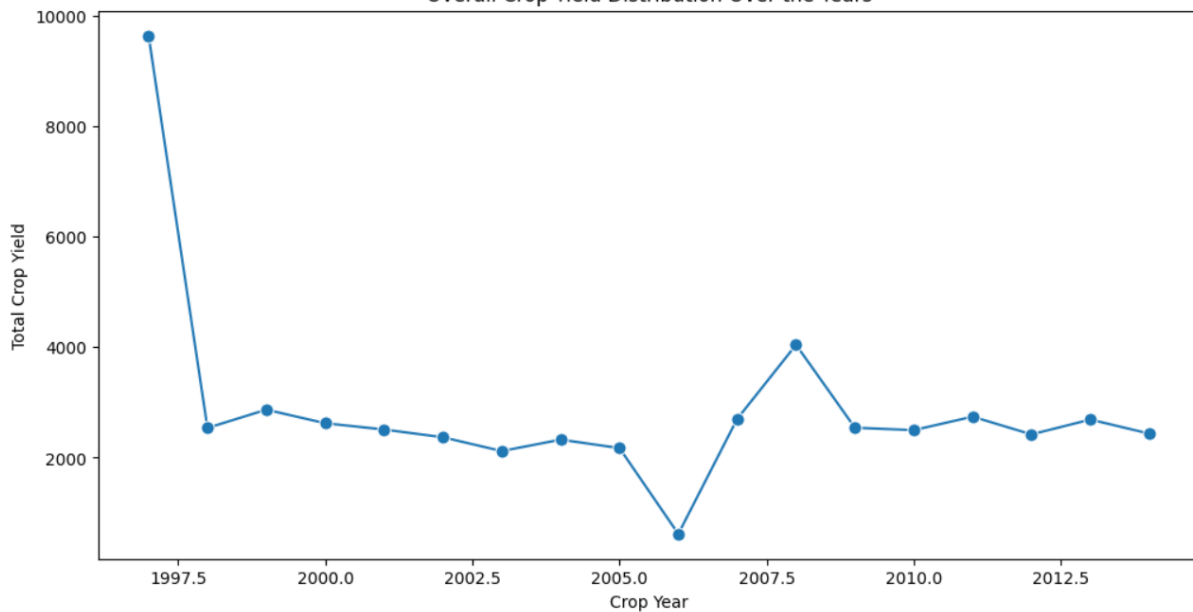
Basic data analysis was performed to get an idea of how the data is structured and its various features, correlation between features were seen and also between the features and the target. On analyzing the data, we noticed that there were a few features that had classes with very few entries as compared to the other classes. Since we want a robust model that generalizes well, these must be handled, crop types that accounted for less than 1 percent of the total number of entries were removed from the data set, barring cereals since it had almost twice the number of entries as other members of this group . After this process, 8 classes of crops were eliminated , and then we moved onto the other categorical features to check for value counts. Soil type had relatively even distribution of data entries, as did season names hence they were left as is. Districts had two classes, Mumbai and Palghar with substantially lower number entries than all the others , hence these were dropped.



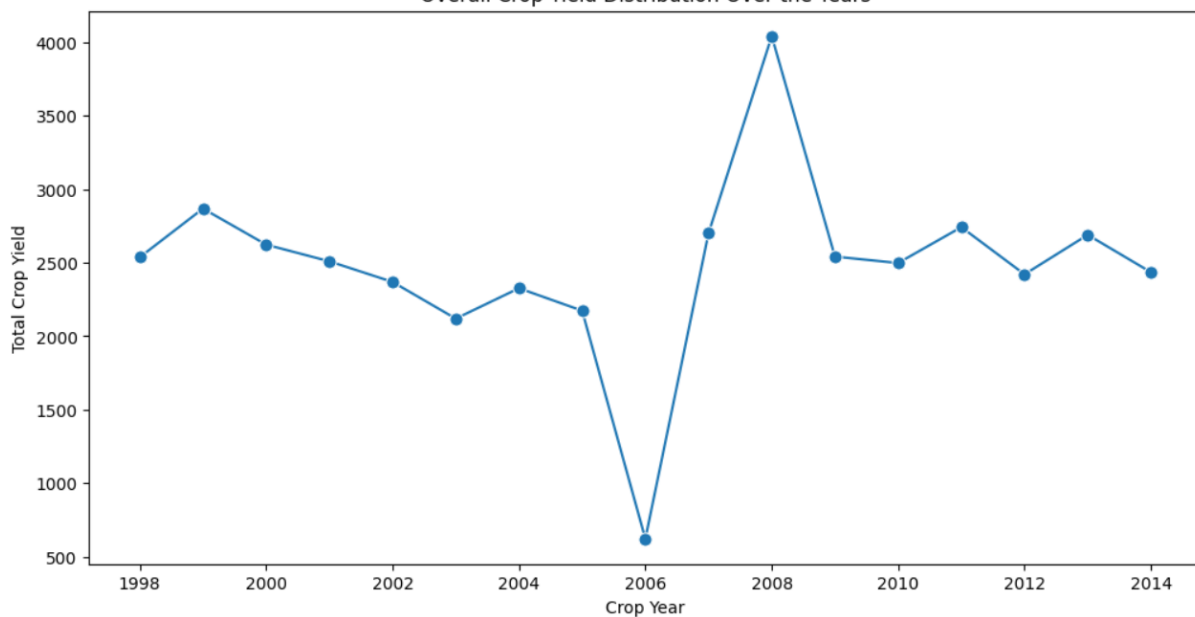
The distribution of yield over the various years were analyzed. On examining this, we found that the entries in 1997 were much larger than the years ahead , and since we have data from over 16 years , it was acceptable to drop all entries from this year. Average yield per year was taken and analyzed on a line plot to view the distribution, in order to prevent a discrepancy in number of entries per year from skewing the distribution. There was a significant deviation in yields around the years of 2006-2008, we speculate that this could have been the result of droughts and floods that occurred in these time periods, since the deviation was not too large , it was left as it was.



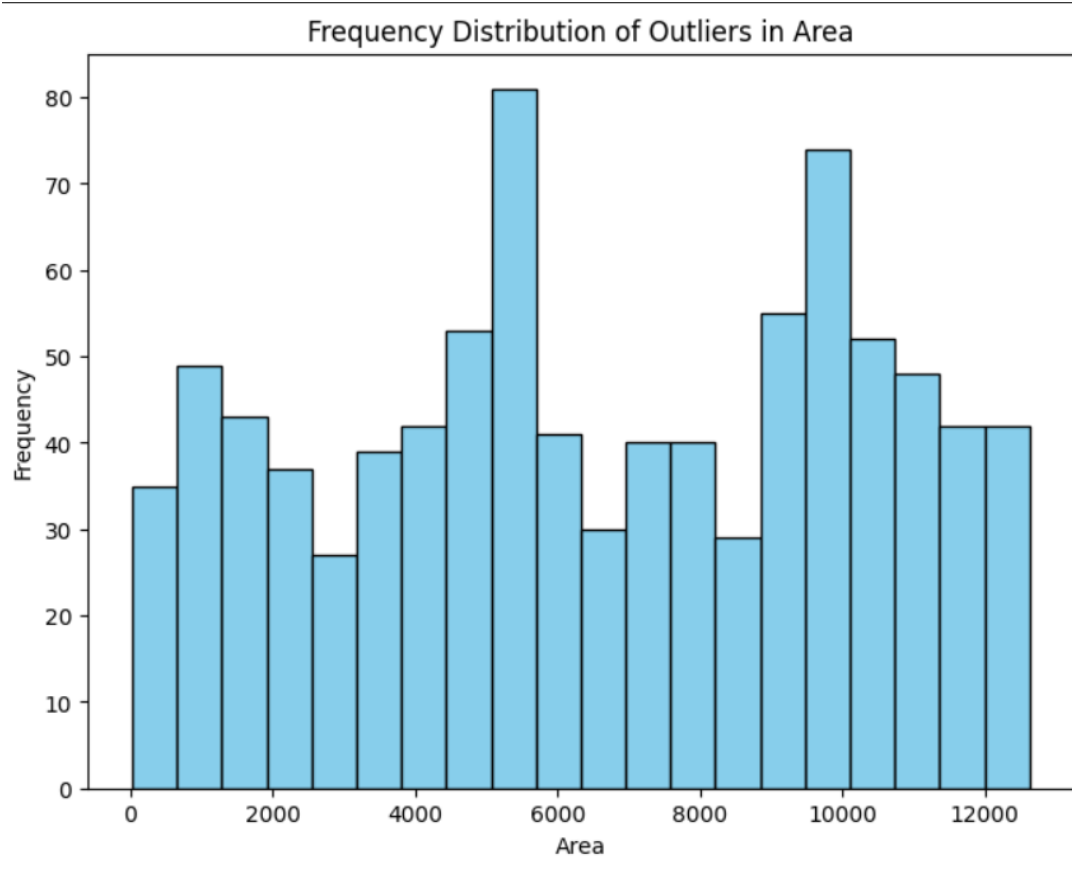
Overall Crop Yield Distribution Over the Years



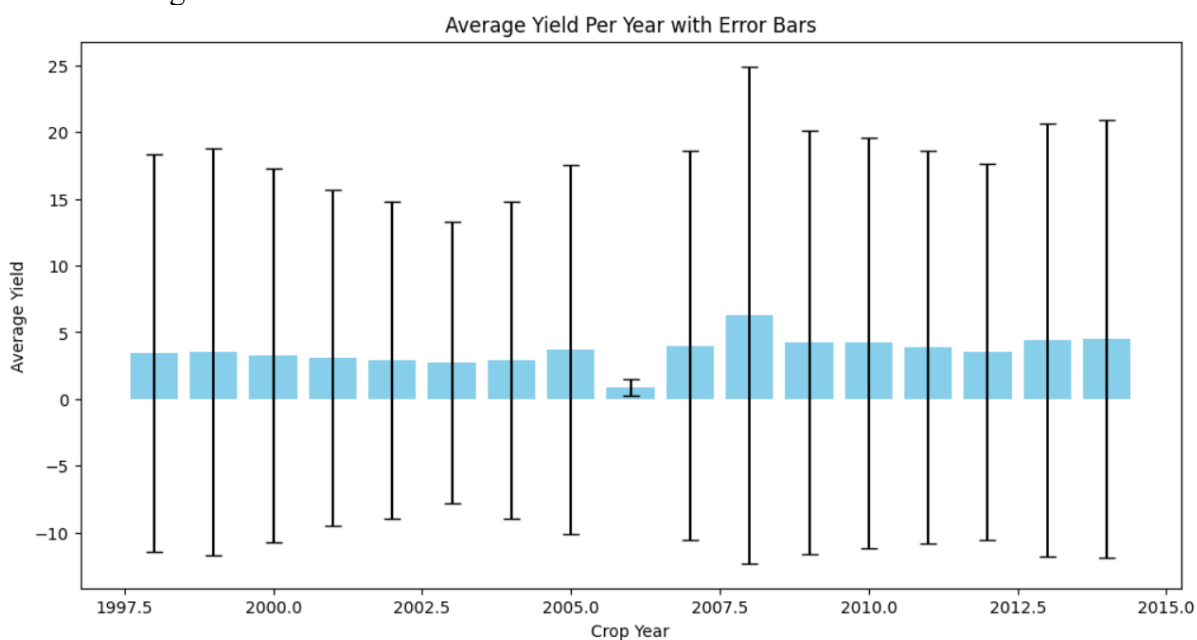
Overall Crop Yield Distribution Over the Years



Box plots were plotted to see how the data of various features were distributed, to see outliers in the data set. On examining area, there was a very large discrepancy between various entries; some entries were of farmland that had very small areas that were multiple orders of magnitude smaller than other entries in the data. This made the yield value of these entries much higher than others and skewed the data.



Finding the right threshold for area outliers could not be determined at this time , so we decided to wait it out till we tested various models on our data to predict yield , trying various thresholds of outlier removal on the yield variable. This would suffice as area and yield were highly correlated . We decided to test out various thresholds and compare the performance metrics on unseen testing data to determine what would be the ideal thresholds to filter out outlier data.



The models tested were Linear Regression , Decision Tree , Random Forest Regressor and XGBRegressor, from various papers that we read random forest was consistently giving better results as opposed to other models so we anticipated a similar result in our testing , and it did



Shri Vile Parle Kelavani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

(Autonomous College Affiliated to the University of Mumbai)

NAAC Accredited with "A" Grade (CGPA : 3.18)



turn out that way , before removing outliers we were getting an exceptional R2 score but mse and mae were quite poor , and after removing outliers based on the standard formula of 1.5 times inter quartile range , this resulted in much better mse and mae scores , but R2 dipped quite a bit

Trying cross validation as well as hyper parameter tuning did not yield significant changes to the performance metrics , after the best model was chosen as random forest regressor now we were left with trying out various thresholds of outlier removal to give us the best middle ground value of R2 vs mse and mae , removing a higher number of outliers seemed to improve the mse and mae but reduce R2, we were aiming for an mse and mae of less than 0 and an R2 of over 85, so that our model would be both accurate on the testing data as well as fitting well enough to explain the variation in the dataset to a satisfiable degree,

After multiple trials with various thresholds , we noticed that leaving a higher amount of large yield values compared to low yield values , as opposed to taking an even amount from both sides off , resulted in the best tradeoff between R2 and mse,mae. Hence we went ahead with this , only 0.5 times Iqr less than Q1 as the bottom threshold and 40 times Iqr more than Q3 as the ideal threshold values .

```
Minimum Yield: 0.02  
Maximum Yield: 29.0
```

```
MSE = 0.1191800151439523  
MAE = 0.22622842816962002  
R2 Score = 0.8602677409526314
```

This resulted in an R2 value of 0.86 while the mse and mae values were 0.11 and 0.22 respectively . For now this is what we went with , but plan to look deeper into the data set to find out why this was the case , and also find and explore other methods of raising accuracy , and learning more about the domain to truly understand what the outliers mean , and the best way to handle them while keeping agricultural science in mind.

MAJOR FINDINGS AND OUTCOMES:

First for the recommendation system , after testing out multiple models , we concluded that Naïve Bayes best fit our dataset the best and gave us the highest accuracy. The second model for yield prediction , Random Forest Regressor showed the best results .A major factor that determined the final accuracy , was how outliers in the data set were handled and filtered out .Other changes in the model or preprocessing did not yield as significant as changing the threshold for outlier removal . This factor heavily influenced accuracy with multiple models that were tried . In the papers we read many other models were tested , but we narrowed it down to the few that we saw consistently gave the best results. We read papers that used neural nets to predict crop yield , but for our dataset that was not as extensive as some that were used in the research papers , neural net would be unnecessary and simpler machine learning models provided a relatively close performance metrics .



Shri Vile Parle Kelavani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

(Autonomous College Affiliated to the University of Mumbai)

NAAC Accredited with "A" Grade (CGPA : 3.18)



SOCIAL RELEVANCE AND APPLICATION

The social relevance of predicting crop yield and building recommendation systems using machine learning is profound, influencing various aspects of society and communities. Accurate crop yield predictions play a crucial role in addressing food security challenges. By anticipating potential crop shortages or surpluses, governments and humanitarian organizations can strategize to alleviate hunger through timely interventions, such as food aid distribution. For many communities, especially in rural areas, agriculture is a primary source of livelihood. Reliable crop yield predictions empower farmers with insights to make informed decisions about crop selection, resource allocation, and market engagement, contributing to economic stability in these regions. Climate change poses significant challenges to agriculture, leading to unpredictable weather patterns. Accurate crop yield predictions assist communities in adapting to these changes by enabling the development of resilient agricultural practices that can withstand the impact of climate variability. Access to accurate information about crop yield predictions can promote social equity in agriculture. Farmers across different socio-economic backgrounds can benefit from such insights, leveling the playing field and reducing disparities in agricultural productivity.

Smallholder farmers, who form a significant portion of the global farming community, can benefit greatly from recommendation systems. Personalized advice helps them overcome resource constraints and improve productivity, thereby enhancing their economic well-being. As farmers adopt more efficient and sustainable practices guided by recommendation systems, the overall quality of life in rural communities can improve. This includes economic well-being, access to education and healthcare, and a healthier environment.

Conclusion and Future Scope

In conclusion, the integration of machine learning techniques for crop yield prediction and recommendation systems represents a transformative leap toward sustainable and efficient agriculture. The multifaceted applications outlined in this research underscore the profound impact these technologies can have on global food security, economic stability, and environmental stewardship.

The predictive accuracy of machine learning models in forecasting crop yields empowers stakeholders to make informed decisions in the face of dynamic environmental and market conditions. By optimizing resource allocation, mitigating risks, and aiding in market planning, these technologies contribute significantly to the resilience and adaptability of agricultural systems.

Simultaneously, recommendation systems pave the way for precision agriculture, guiding farmers toward sustainable practices, technology adoption, and continuous improvement. The socio-economic implications are profound, fostering social equity, community-based decision-making, and environmental stewardship.

As we navigate the complexities of the 21st century, the collaboration between technology and agriculture emerges as a linchpin in addressing global challenges. The findings presented in this research paper underscore the importance of embracing machine learning for crop yield prediction and recommendation systems as integral components of a resilient and



Shri Vile Parle Kelavani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

(Autonomous College Affiliated to the University of Mumbai)

NAAC Accredited with "A" Grade (CGPA : 3.18)



adaptive agricultural landscape.

In the future we plan to link our recommendation system along with yield to give the best recommendation along with an estimate of yield .Also we would try to analyse financial data for crops to find out the estimated cost of the crop recommended to predict revenue of the farmers total production .We also need to look deeper into the dataset for yield prediction and figure out the best way to handle the area outliers , and hence improve the accuracy of the model.

We also want to find ways to extract real time data and feed the model automatically , as opposed to currently where each data field needs to be entered manually . Since we want this to have ease of use to farmers , we cant expect them to have to figure out attributes like wind speed , soil chemical composition etc. Finding ways to get data about the weather conditions of a farm by just inputting the location it is in would make it much more useful to the users. This could be achieved by requesting data from APIs that can give information about weather and soil conditions.

Lastly we could like to conduct research and try to implement satellite imaging data into our product , this would provide valuable insight into crop yield prediction in the following ways : vegetation health monitoring, land cover and use, climate and weather monitoring, cloud cover and yield prediction calibration

REFERENCES

- 1)Kavita Jhahariaa , Pratistha Mathura* , Sanchit Jaina , Sukriti Nijhawana. “Crop Yield Prediction using Machine Learning and Deep Learning Techniques” , In 2022 First International Conference on Machine Learning and Data Engineering .
Published by Elsevier B.V (2023), 10.1016/j.procs.2023.01.023
2. Thomas van Klompenburg a, Ayalew Kassahun a, Cagatay Catal b . “Crop yield prediction using machine learning: A systematic literature review” , Computers and Electronics in Agriculture, Volume 177, October 2020, 105709 ,
<https://doi.org/10.1016/j.compag.2020.105709>
3. Suresh Kumar Sharma, DP Sharma “CROP YIELD PREDICTIONS AND RECOMMENDATIONS USING RANDOM FOREST REGRESSION IN 3A



AGROCLIMATIC ZONE, RAJASTHAN”, April 2023 Shu Ju Cai Ji Yu Chu Li/Journal of Data Acquisition and Processing 38(2):1635-1651 DOI:10.5281/zenodo.776786

https://www.researchgate.net/publication/370029230_CROP_YIELD_PREDICTIONS_AND_RECOMMENDATIONS_USING_RANDOM_FOREST_REGRESSION_IN_3A_AGROCLIMATIC_ZONE_RAJASTHAN

4. Saeed Khaki, Lizhi Wang , Crop Yield Prediction Using Deep Neural Networks , Front. Plant Sci., 22 May 2019, Sec. Computational Genomics, Volume 10 -2019,
<https://doi.org/10.3389/fpls.2019.00621>
<https://www.frontiersin.org/articles/10.3389/fpls.2019.00621/full>
5. Javad Ansariaf , Lizhi Wang, Sotirios V. Archontoulis , “An interaction regression model for crop yield prediction”, *Scientific Reports* 11, Article number: 17754 (2021), Published: 07 September 2021
<https://www.nature.com/articles/s41598-021-97221-7>
6. K P K Devan, Swetha B, Uma Sruthi P, Varshini S, “Crop Yield Prediction and Fertilizer Recommendation System Using Hybrid Machine Learning Algorithms”, 12th IEEE International Conference on Communication Systems and Network Technologies 2023, 978-1-6654-6261-7/23/\$31.00 ©2023 IEEE
DOI: 10.1109/csnt.2023.33
7. Anakha Venugopal, Aparna S, Jinsu Mani, Rima Mathew, Prof. Vinu Williams, “ Crop Yield Prediction using Machine Learning Algorithms”, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, NCREIS - 2021 Conference Proceedings, Special Issue- 2021, Volume 9 Issue 13, Published by, www.ijert.org
8. Martin Kuradusenge , Eric Hitimana , Damien Hanyurwimfura , Placide Rukundo , Kambombo Mtonga ,Angelique Mukasine , Claudette Uwitonze , Jackson Ngabonziza , Angelique Uwamahoro , “Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize”, *Agriculture* 2023, 13, 225.
<https://doi.org/10.3390/agriculture13010225> <https://www.mdpi.com/journal/agriculture>
9. Aruvansh Nigam, Saksham Garg, Archit Agarwal, Parul Agarwal, “Crop Yield Prediction Using Machine Learning Algorithms” , 2019 Fifth International Conference on Image Information Processing (ICIIP)