Crop yield Prediction Using Machine Learning Techniques

Sai Vivek Kodam
Computer Science
New Mexico State University
vivekk@nmsu.edu

Humesh Reddy Venkatapuram
Computer Science
New Mexico State University
humeshrv@nmsu.edu

Mohith Krishna Reddy D
Computer Science
New Mexico State University
dmohith@nmsu.edu

Olabisi Adebisi
Molecular Biology
New Mexico State University
gadebisi@nmsu.edu

***Abstract*—Agriculture, a cornerstone of human civilization, faces a new dawn of transformation through the application of machine learning. In this context, our research takes center stage, focusing on the utilization of machine learning algorithms to predict crop yields—a critical factor in farmers' livelihoods. By employing various machine learning techniques and comparing their outcomes, we offer a beacon of hope for farmers. The predictions generated by these algorithms empower farmers with the knowledge needed to make well-informed decisions about crop selection, enabling them to optimize yields while considering variables like temperature, rainfall, and regional conditions. Through the incredible potential of machine learning, our study presents a promising solution to enhance agricultural productivity and, in turn, the financial well-being of farmers.**

In essence, this research transcends technology; it represents a commitment to the future of agriculture. It seeks to provide a lifeline to farmers by mitigating risks associated with climate uncertainties and market fluctuations. Furthermore, it aligns with global goals of ensuring food security and sustainable agricultural practices, ultimately promising a brighter and more prosperous future for farmers and our food systems as a whole.

I. Introduction

The Agricultural Crop Recommendation System stands at the forefront of technological advancements in agriculture, leveraging the power of artificial intelligence to revolutionize the way farmers make crucial crop decisions. This innovative software is a beacon of hope for farmers facing the daunting task of crop selection, particularly in regions where financial hardships are all too common.

In many nations, farmers often find themselves grappling with significant financial losses caused by the consequences of poor crop choices. The consequences can be devastating, ranging from crop failure due to unpredictable weather patterns to devastating pest infestations. In dire circumstances, some farmers even contemplate extreme measures, such as suicide, as they struggle to cope with the overwhelming burdens of agricultural setbacks.

In these trying times, the Agricultural Crop Recommendation System emerges as a lifeline for distressed farmers, offering them tailored advice that has the potential to avert losses and alleviate their financial burdens. Through the guidance of this system, farmers are empowered to optimize their crop selection, leading to increased production, reduced costs, and enhanced overall efficiency.

This transformative approach not only has the power to reshape the agricultural landscape but also holds the promise of enhancing the lives of individual farmers. It grants them the autonomy to make decisions that are not only economically viable but also environmentally sustainable, addressing pressing global issues such as food security and climate change.

The Agricultural Crop Recommendation System, in essence, serves as an indispensable tool, extending a lifeline to struggling farmers and paving the way for their financial stability and a brighter future in agriculture. As it continues to evolve, its impact on agriculture is poised to be nothing short of revolutionary, fostering a more resilient, sustainable, and prosperous agricultural sector for generations to come.

II. Motivation

Our motivation to delve into the realm of crop prediction and recommendations is driven by a deep understanding of how these advancements can profoundly impact the lives of farmers and the sustainability of our agricultural practices. We are motivated by a sincere desire to help alleviate the challenges faced by farmers, especially in regions where their livelihoods are most susceptible to the ups and downs of both the environment and the economy.

Our research is centered on harnessing the power of data-driven insights and state-of-the-art technology to offer practical solutions to farmers. Through the development and assessment of crop prediction and recommendation systems, our goal is to provide a glimmer of hope to those whose lives

depend on agriculture. Moreover, we recognize the broader significance of this research in tackling global issues like food security and the responsible use of agricultural resources.

The urgency of our work is emphasized by the critical need to address the adverse impacts of climate change and promote more efficient and sustainable farming practices. Through our research, we aspire to contribute to a brighter and more secure future for farmers and to advance a sustainable approach to food production that benefits our world as a whole.

III. DATA

Upon examining the datasets, we've found them to be well-suited for forecasting crop production. Notably, the pesticide dataset, with its 4350 rows and 7 columns, offers insights into pesticide usage across regions and its influence on crop production. We see this as a valuable resource for understanding the precise relationship between pesticides and crop yields. Additionally, the rainfall dataset, comprising 6728 rows and 3 columns, provides essential information on regional rainfall patterns, a critical factor impacting crop yields. Analyzing the correlation between rainfall and crop production enables us to make predictions by region.

Likewise, the temperature dataset, with its 71312 rows and 3 columns, details average temperatures in various regions—an equally crucial factor in crop production. Examining the temperature-crop yield relationship allows us to forecast production for specific regions. We've also been impressed by the yield dataset, which records crop yields in various regions. This historical yield data is invaluable for accurately predicting future crop production.

To streamline our analysis, we've created the `yield_df` dataset—a cleaned and merged version of the aforementioned datasets. This customized dataset optimizes our study's specific needs. With it, we aim to analyze how factors like rainfall, temperature, pesticides, and crop yield interrelate to predict crop production accurately. Our datasets are reliable, sourced from reputable channels, and our cleaning and merging efforts save valuable time and resources.

Our approach involves employing statistical models, particularly regression analysis, to explore the connections between these factors and crop yield. Specifically, we'll delve into the impact of pesticides, rainfall, and temperature on crop yields. Through this examination of historical data, we're developing a model that can accurately forecast future crop yields.

IV. RESEARCH PROBLEMS

Agriculture remains a foundational pillar for societies worldwide, fueling economies and providing essential sustenance. As global populations grow and demands intensify, there is an ever-increasing pressure on the agricultural sector to ensure both the quantity and quality of crop yields. Yet, one of the most persistent challenges facing this sector is the unpredictability of weather patterns, compounded by the consistent variations in humidity levels. These variables significantly impact crop growth, making it challenging for farmers to accurately predict yields and decide which crops to cultivate each season.

The unpredictability of weather directly affects crop selection and potential yields. Without accurate forecasts, farmers might invest time, resources, and effort into cultivating crops that eventually underperform due to unforeseen weather conditions. This uncertainty not only risks their livelihood but also has broader implications for food security and the stability of global agricultural markets. In a world where climate change exacerbates such unpredictabilities, there's a critical need to find robust solutions that can guide farmers with greater precision.

Harnessing the power of data analytics, particularly through analyzing historical weather patterns, soil conditions, and other vital agricultural parameters, can provide invaluable insights. By developing a predictive model that incorporates these data sets, we can offer farmers a tool that aids in making informed decisions about crop cultivation. This model wouldn't just be a reflection of past patterns but an anticipatory guide for future crop yields, enabling adaptive farming techniques tailored to expected conditions.

By successfully integrating data analytics into the realm of agriculture, we aim to revolutionize the way farmers approach crop cultivation, making it more strategic and informed. This not only ensures food security by optimizing yields based on predicted conditions but also promotes the economic resilience of farming communities globally. In essence, the intersection of agriculture and technology has the potential to shape a future where food supply is both secure and sustainable, underpinning the prosperity of societies around the world.

V. LITERATURE REVIEW

Paper 1: Agriculture Crop Selection and Yield Prediction using Machine Learning Algorithms:

This paper presents a crop recommendation and yield prediction model based on weather parameters, utilizing the Random Forest algorithm and comparing it with Support Vector Machine (SVM) and Multivariate Regression. The dataset employed comprises 300 instances with six diverse features encompassing temperature, wind, rainfall, soil moisture, humidity, and precipitation. Random Forest demonstrates superior performance with 90 accuracy in crop classification, compared to SVM's 65, and also excels in yield prediction with 90 accuracy, surpassing Multivariate Regression (85) and SVM (65). The success of Random

Forest is attributed to its utilization of bagging and feature sampling techniques during tree construction, making it a valuable tool for assisting farmers in selecting appropriate crops and estimating production based on prevailing weather conditions.

The selection of these methods was based on several key considerations. The dataset featured both categorical (crop type) and continuous (yield, weather data) variables, necessitating the use of algorithms capable of handling both types. The authors prioritized the attainment of highly accurate models, leading to the choice of ensemble methods like Random Forest and kernel methods such as SVM. Given the relatively small dataset size of 300 samples, methods well-suited for working with limited data, such as SVM, were preferred. Additionally, the authors placed importance on ease of interpretation, leading to the utilization of tree-based models like Random Forest and linear models like regression, as opposed to more complex neural networks.

Paper 2: Crop Yield Prediction using Machine Learning Techniques:

This study explores machine learning algorithms, Naive Bayes and K-Nearest Neighbors (KNN), for crop yield prediction. KNN outperforms Naive Bayes with an accuracy of 87 due to its instance-based learning approach. The dataset categorizes crops into four seasons and is applied to two state datasets - Uttar Pradesh and Karnataka.

KNN's non-parametric, instance-based approach proves superior for crop classification compared to Naive Bayes' probabilistic model. This application has the potential to assist farmers in crop selection, potentially reducing losses and improving productivity. The methods used include Naive Bayes for yield classification and KNN for actual yield prediction. These methods are integrated into a Java application, with KNN demonstrating better accuracy, as confirmed by ROC curve analysis.

Paper 3: Impact of Machine Learning Techniques in Precision Agriculture

This paper explores AI and machine learning methods for crop recommendation systems, considering factors like weather, soil, temperature, and crop growth rate. Random Forest consistently outperforms other models in crop classification, highlighting the importance of robust data collection and accurate weather prediction for precision agriculture. The paper introduces a basic architecture for crop recommendation, including data collection, preprocessing, feature extraction, model training, evaluation, and crop recommendation. Experiments involving Random Forest, Logistic Regression, Naive Bayes, and Decision Table models classify crops into seasons, with Random Forest consistently proving its effectiveness.

Various methods are employed, such as ensemble models like Random Forest, Majority Voting, and similarity-based models like Pearson Correlation. Machine learning algorithms like K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Neural Networks, and Decision Trees (e.g., CHAID, Random Tree) are used for crop classification. Comparative evaluations reveal the strengths and weaknesses of different techniques for this application, enabling the implementation of models that offer personalized crop suggestions based on location, soil, weather data, and more.

Paper 4: Prediction of Crop Yield Production with Novel Lasso Regression against Polynomial Regression Algorithm to Achieve Higher Accuracy

In this research paper, the primary focus is on leveraging machine learning algorithms to enhance the accuracy of crop yield predictions. Utilizing a dataset comprising 1,112 entries labeled with parameters such as "temperature," "humidity," "rainfall," and "production," the study compares the effectiveness of Lasso Regression (LR) with Polynomial Regression (PR). The results underscore Lasso Regression's superior performance, achieving an impressive accuracy rate of 94.20 in contrast to Polynomial Regression's 75.

The research accentuates the transformative impact of modern technology in Indian agriculture, spotlighting Lasso Regression's unparalleled crop yield prediction accuracy. Its superiority over other methods earmarks it as a future-forward tool in agricultural machine learning. The study integrates Long Short-Term Memory (LSTM) networks with 1D Convolutional Neural Networks (CNN). While LSTM adeptly forecasts multi-step weather variables, 1D CNN extracts key sequential data features. Combined, the LSTM-CNN hybrid outperforms standalone models, excelling in seven-day weather predictions, which can refine crop yield estimations and farming strategies.

Paper 5: Extreme Gradient Boosting (XGBoost) Regressor and Shapley Additive Explanation for Crop Yield Prediction in Agriculture

This research paper explores the potential of the XGBoost regressor in predicting crop yields by leveraging data on meteorological conditions, soil properties, and historical crop yield records. Known for its robust prediction capabilities, XGBoost is fine-tuned using cross-validation. Beyond prediction accuracy, the study places a strong emphasis on model interpretability, introducing the use of Shapley Additive Explanation (SHAP) values to elucidate the decision-making process. SHAP helps identify influential features such as temperature, precipitation, and soil characteristics, contributing significantly to crop yield predictions. The integration of XGBoost and SHAP not only enhances prediction precision but also provides interpretable insights,

offering valuable benefits to farmers and stakeholders in the agricultural sector.

In addressing the problem, Random Forest, XGBoost, and Artificial Neural Networks (ANN) models are trained on a dataset containing factors like crop type, fertilizer, pesticides, and rainfall. Comparative analysis reveals that XGBoost achieves the highest accuracy at 97.6 for crop yield prediction, surpassing Random Forest and ANN. XGBoost's regularization techniques and advanced gradient boosting methods effectively prevent overfitting and enhance generalization. The XGBoost model stands out by providing accurate crop yield forecasts, empowering farmers to maximize productivity and profitability in their agricultural endeavors.

Paper 6: Comparative Analysis of Machine Learning Techniques for Disease Prediction in Crops

The study highlights the importance of agriculture in India and the challenges posed by crop diseases. By leveraging machine learning, specifically utilizing the WEKA data mining tool, the authors explored various algorithms on a soybean disease dataset from the UCI machine learning library to determine their efficacy in predicting crop diseases. The research tested methods such as Naive Bayes, Logistic Regression, and Random Forest, among others, to classify crops as healthy or infected, using metrics like accuracy, sensitivity, and specificity.

The results showcased that the Simple Logistic algorithm was the most accurate, achieving a 98.13 accuracy rate, closely followed by the SMO and Random Forest methods. Through a comparative analysis, it became evident that machine learning could provide a precise and efficient solution for predicting agricultural diseases. The insights gained from this study suggest that these machine learning techniques can be employed for automated disease diagnosis, aiding farmers in taking proactive measures.

Paper 7: Crop Recommendation in Precision Agriculture using Supervised Learning Algorithms

This paper focuses on the primary challenges farmers face in selecting the optimal crop considering factors like soil type, climate, and geography. To address this, the paper introduces a crop recommendation system built using an ensemble model, incorporating methods like K-Nearest Neighbors, Logistic Regression, and Support Vector Machine. This system, trained on a Kaggle dataset, suggests crops tailored to specific site parameters, achieving a prediction accuracy of 92. As a result, farmers can optimize yield and profitability by planting the most suitable crops.

Emphasizing the significance of precision agriculture, the study guides farmers in crop selection, especially with shifting climate patterns. The research suggests expanding the dataset and integrating more attributes, including climatic data, to refine the recommendation system. The Random Forest classifier, with its highest accuracy, was pivotal in this endeavor. It not only provided the most accurate crop suggestions based on land characteristics but also underscored the potential of data-driven approaches in enhancing agricultural productivity.

Paper 8: Crop Recommendation System using Random Forest Algorithm in Machine Learning

The paper presents a Crop Recommendation System built using the Random Forest Algorithm, designed to suggest the ideal crop for specific sites considering parameters such as nitrogen, phosphorus, potassium, and humidity. Although various algorithms like KNN, Decision Tree, Naïve Bayes, SVM, and ANN were evaluated, Random Forest stood out for its exceptional accuracy, reaching a remarkable 99 using data sourced from Kaggle. This data encompassed various factors like temperature, humidity, rainfall, and nutrient levels for diverse crops.

Highlighting the importance of precise crop selection tailored to land quality, the study asserts that such accurate predictions can significantly boost agricultural output. The choice of Random Forest stems from its ensemble nature, combining multiple decision trees, its proficiency in managing noisy and high-dimensional data, and its robustness against overfitting. With this model, farmers receive tailored crop recommendations based on diverse factors, enhancing productivity. The research further envisions hosting this system on cloud platforms, thereby broadening its accessibility to farmers across the nation.

Paper 9: Machine Learning based Smart Crop Recommender and Yield Predictor

This study introduces a comprehensive machine learning system for crop recommendations and yield predictions. Deploying models like decision tree, light GBM, naive bayes, random forest, and Xgboost, the system achieved an impressive 99 accuracy in suggesting ideal crops using soil and environmental factors. Additionally, it provides insights into potential profits from trade policies and suggests appropriate fertilizers and pesticides. With data sourced from Kaggle, the system also aids in yield predictions, which play a crucial role in making informed export-import decisions, ultimately augmenting farmer profits.

The research employed a variety of machine learning models, training them on datasets encompassing soil attributes, climatic conditions, and crop types. Among them, Naive Bayes, Random Forest, and XGBoost stood out, offering 99 accuracy in crop recommendations. These models also proved instrumental in predicting yields, further assisting

in estimating potential profits. In a nutshell, this study delivers an all-encompassing approach to crop recommendations, with promising implications for bolstering agricultural profitability. Future endeavors may integrate disease detection and its impact on yields and revenue.

Paper 10: Smart Crop Recommender System- A Machine Learning Approach

The "Smart Crop Recommender System" introduced in this paper harnesses machine learning to recommend a selection of 22 crops, employing a three-tiered framework consisting of data preprocessing, classification, and performance evaluation. The study delves deep into feature analysis, using correlation plots and density distribution, and employs ensembling techniques for classification. Notably, the Naïve Bayes classifier stands out with an impressive accuracy of 99.54, outshining the 98.52 achieved by the majority voting ensembler. The research underscores the value of machine learning in precision agriculture, aiming to address challenges farmers face from fluctuating economic conditions by offering crop recommendations based on myriad factors, from soil characteristics to temperature.

A multi-class classification approach is central to the study, with a keen focus on ensemble techniques and majority voting. Algorithms like Decision Tree, Random Forest, Naive Bayes, and SVM were trained on a rich dataset encompassing soil properties, climatic variables, and crop types. Among these, Naive Bayes emerged as the top performer, laying the foundation for the crop recommendation system. This system, built on a data-driven approach, seeks to enhance agricultural productivity by guiding farmers towards the most optimal crop choices. The research paves the way for a future where farmers make more informed, profitable decisions, backed by sophisticated machine learning tools.

VI. Solutions and action plans

Data Cleaning, Preprocessing: The data preparation script prepares a dataset for further analysis by performing several operations, such as handling missing values, encoding categorical variables, and aggregating data. Its primary purpose is to prepare the data for further analysis.

A. Loading Data

The script starts by checking if a CSV file named "yield_df.csv" exists. If it does, it loads a data frame called `yield_df` from this file. If not, it loads four separate CSV files: "pesticides.csv," "rainfall.csv," "temp.csv," and "yield.csv."

B. Data Cleaning and Preprocessing for Each Data Source

For each of the four data sources (`pesticides_df`, `rainfall_df`, `temp_df`, and `yield_df`), several data cleaning and preprocessing steps are applied:

1) Unnecessary columns are dropped to retain only relevant information.
2) Column names are standardized for consistency.
3) Year values are converted to integers for uniform data types.

C. Data Merging

After cleaning and preprocessing individual data sources, the script merges them into a single data frame named `yield_df`. This is achieved by performing outer joins based on the "Country" and "Year" columns, combining data from different sources into one cohesive dataset.

D. Data Cleaning and Filtering

Further cleaning is applied to the `yield_df` DataFrame:

1) Rows with missing values in the "Yield" column are removed.
2) Rows with "Value" less than or equal to 0 are filtered out.
3) Rows with "Rainfall" and "Temperature" marked as "No Data" are excluded.
4) The string ".." in the "Rainfall" column is replaced with NaN.
5) Missing values in the "Rainfall" column are imputed with zeros.
6) Data types for numeric columns are adjusted to integers or floats as needed.

E. One-Hot Encoding

Categorical variables, including "Country," "Area Code," and "Item," are one-hot encoded using the OneHotEncoder from scikit-learn. This transformation converts categorical variables into binary vectors, making them suitable for machine learning algorithms.

F. Data Type Inspection

The script prints the data types of the columns in `yield_df` to help users understand the data types after preprocessing. This step aids in verifying that the data is in the correct format.

G. Data Aggregation

Finally, the script aggregates the `yield_df` DataFrame by grouping it based on the "Item" and "Year" columns. It calculates the mean of each group's "Yield" column, creating a new data frame called `grouped_df`. This aggregated data can be useful for analyzing trends in agricultural yield over time for different items.

```
Country          object
Year              int64
Value           float64
Rainfall          int64
Temperature     float64
Area Code         int64
Item             object
Yield           float64
dtype: object
```

Data Visualization and Analysis: Data visualization is a critical tool in agricultural research and analysis. It enables researchers and analysts to gain insights, identify patterns, and make informed decisions based on agricultural data. In this document, we explore a Python script that leverages the power of the Seaborn and Matplotlib libraries to visualize and analyze agricultural data.

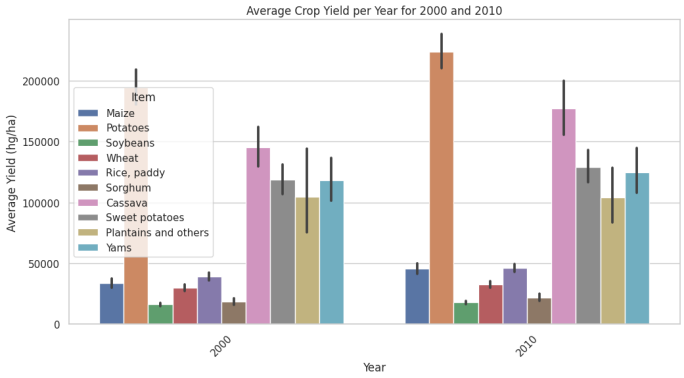
It is designed to work with a preprocessed agricultural dataset named `yield_df`. This dataset contains information about crop yields, temperature, and other factors over multiple years and for various agricultural items.

H. Filtering Data for Specific Years:

The script begins by selecting specific years, 2000 and 2010, from the `yield_df` dataset. This step involves the use of the ‘isin‘ function to filter rows where the ”Year” column matches either 2000 or 2010. The resulting dataset is stored in a new data frame called `filtered_df`.

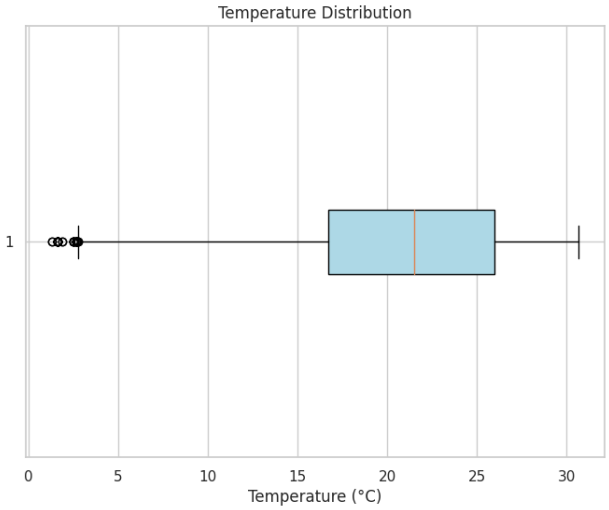
Setting Seaborn Plot Style: Seaborn is a data visualization library that enhances Matplotlib’s capabilities. The script sets the Seaborn plot style to a ”white grid.” This style provides a clean and visually appealing background with gridlines, enhancing the clarity of the plots. The core of the code focuses on visualizing average crop yields for the years 2000 and 2010, broken down by different crop items. This visualization is achieved through Seaborn’s ‘barplot‘ function.

The script initializes a Matplotlib figure with a specified size (12x6 inches). key visualization is created using Seaborn’s ‘barplot‘ function. It specifies the x-axis as ”Year”. The y-axis is ”Yield”. The Hue (color grouping) is ”Item”. This creates a grouped bar plot that allows for a visual comparison of yield variations across different crop items for the years 2000 and 2010. The script customizes the plot with a title, x-axis label, y-axis label, and rotation of x-axis labels for better readability. The ‘show‘ function is used to display the finalized bar plot. This visualization provides valuable insights into how crop yields have evolved over the selected years and how different crop items compare.



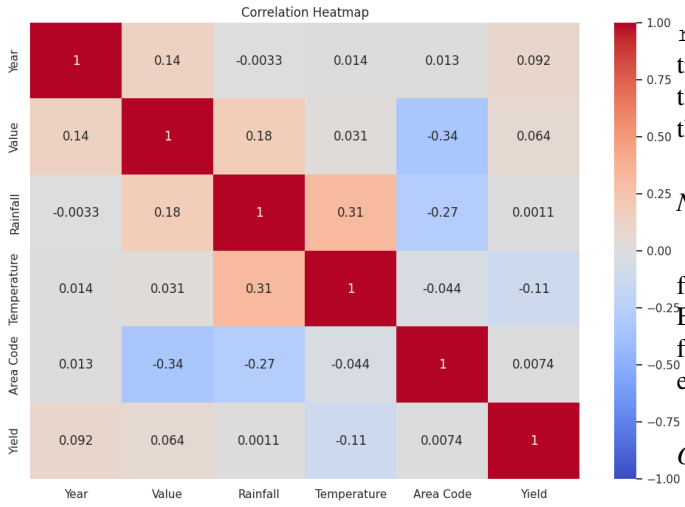
I. Visualizing Temperature Distribution with Matplotlib:

In addition to crop yield visualization, the script also includes a visualization of temperature distribution using Matplotlib’s box plot. A new Matplotlib figure is initialized with a specified size (8x6 inches). The script generates a box plot using the ‘boxplot‘ function, focusing on the ”Temperature” column of the `yield_df` dataset. The box plot provides insights into the distribution of temperature data, including measures of central tendency and spread. Various styling options are applied, including setting the plot title, and x-axis label, enabling gridlines, and choosing a light blue color for the box plot elements. The ‘show‘ function is used to display the finalized box plot. This visualization aids in understanding the distribution of temperature data in the agricultural dataset, which is crucial for assessing the impact of temperature on crop yields.



J. Visualizing Correlations Heatmap with Seaborn:

Lastly, the script includes a correlation heatmap generated using Seaborn. This heatmap provides insights into the relationships between different variables in the dataset. The script uses Seaborn's 'heatmap' function to create a correlation heatmap. This heatmap visualizes pairwise correlations between numeric columns in the `yield_df` dataset. The heatmap is customized with options such as displaying correlation values within each cell, choosing a color map ("coolwarm") to represent positive and negative correlations, and setting the color scale's minimum and maximum values. The 'show' function displays the correlation heatmap. This visualization is essential for identifying and understanding the relationships between different agricultural variables, helping researchers make data-driven decisions.



In summary, the script showcases the power of data visualization in agricultural research and analysis. It allows researchers and analysts to explore and understand agricultural data by visualizing crop yields, temperature distributions, and correlations between variables. These visualizations are invaluable for making informed decisions, identifying trends, and gaining insights into agricultural datasets. By leveraging libraries like Seaborn and Matplotlib, researchers can unlock the full potential of their data and contribute to more effective agricultural practices and policies.

Data Splitting, Cross-Validation, and Model Evaluation: This section focuses on data splitting, cross-validation, and the evaluation of regression models for predicting crop yields. These steps are essential in machine learning and data analysis for assessing model performance.

K. Data Splitting:

The script begins by splitting the data into training and testing sets using the `train_test_split` function from scikit-learn. The independent variables are stored in `X`, while the dependent variable (Yield) is stored in `y`. The split ratio is set at 70 for training and 30 for testing, with a specified random seed for reproducibility. This separation is crucial for evaluating how well the models generalize to unseen data.

L. Regression Models:

The script defines several regression models for analysis:

1) Lasso Regression
2) Random Forest Regressor
3) K-Nearest Neighbors (KNN) Regressor
4) Gradient Boosting Regressor

Each model is instantiated with specific hyperparameters.

M. Model Evaluation Function:

The script includes a function called `runtime_model_prediction_` to fit the model to training data, make predictions on test data, and measure the time taken for training and prediction. This function returns the model, predicted values, and runtime.

N. Model Evaluation and Results:

The `get_analysis` function evaluates each model's performance using metrics such as R-squared and Mean Squared Error (MSE) on the testing data. It also calculates the runtime for training and prediction. The results are stored in a list for each model and displayed in the console.

O. Data Visualization with Pandas:

The code prepares a Pandas DataFrame called `df_test` to visualize the model predictions alongside actual yield values. This enables a visual comparison of how well each model performs in predicting crop yields.

P. Model Instances and Evaluation:

Four models, namely Lasso, Random Forest, KNN, and Gradient Boosting, are instantiated, each with a unique name. These models are included in the list `models_to_evaluate` for systematic evaluation.

Q. Analysis and Display:

The code evaluates each model using the `get_analysis` function, which returns metrics such as R-squared and MSE, as well as runtime for training and prediction. The results are stored in a Pandas DataFrame (`results_df`) and displayed in the console.

	Model	R-squared	Mean Squared Error	Runtime (seconds)
0	Lasso	0.750733	1.803388e+09	6.333452
1	Random Forest	0.987083	9.345274e+07	24.444924
2	KNN	0.358257	4.642863e+09	3.155028
3	Gradient Boosting	0.871889	9.268556e+08	7.628975

R. Cross-Validation:

Next, the code sets up a K-Fold Cross-Validation (KFold) scheme with five splits using `KFold` from scikit-learn. K-Fold Cross-Validation is employed to assess model performance while mitigating overfitting or underfitting risks. It shuffles the data and divides it into training and testing sets across multiple iterations.

Within the cross-validation loop, the script further divides the data into training and testing subsets for each fold. This ensures that each data point is used for testing once and for training multiple times (in this case, four times, as there are five folds). The data for each fold is appended to respective lists (`X_train_cross_list`, `X_test_cross_list`, `y_train_cross_list`, `y_test_cross_list`) for later use.

	Model	R-squared	Mean Squared Error	Runtime (seconds)
0	Lasso	0.755422	1.712823e+09	8.673018
1	Random Forest	0.988086	8.343266e+07	39.163556
2	KNN	0.363788	4.455505e+09	2.232435
3	Gradient Boosting	0.867835	9.255779e+08	19.080139

In summary, this script presents a comprehensive pipeline for assessing the performance of various regression models in predicting crop yields. It emphasizes data splitting, cross-validation, model instantiation, and evaluation. The code is essential for selecting the most suitable model for crop yield prediction, providing valuable insights for agricultural research and decision-making. It also adheres to good practices in machine learning, such as cross-validation and evaluating multiple models to ensure robust and reliable predictions.

Visualizing Model Predictions for Crop Yield: This section focuses on visualizing and comparing actual crop yield values with predictions generated by multiple regression models. The script utilizes the Matplotlib library to create a scatter plot that visually represents the performance of these models.

S. List of Models

The script begins by defining a list named `models_to_evaluate`, which contains various machine learning models, such as `lasso_model`, `rf_model`, `knn_model`, and `gb_model`. It is assumed that these models have been previously trained.

T. Dictionary for Predictions

Next, an empty dictionary named `all_predictions` is initialized. This dictionary will be used to store the predictions made by each model.

U. Model Evaluation Loop

The code then enters a for loop that iterates through the `models_to_evaluate` list. For each model in the list, the following steps are performed:

1) Prediction: It utilizes the `model.predict(subset_X_test)` method to make predictions on a subset of the test data referred to as `subset_X_test`. This subset is assumed to be a NumPy array or pandas DataFrame containing the testing features.

2) Storage: The predictions generated by the model are stored in the `all_predictions` dictionary, with the model's name serving as the key. Therefore, for each model, the corresponding predictions are stored in the dictionary.

V. Creating a DataFrame for Comparison

After obtaining predictions from all the models, the code creates a pandas DataFrame named `all_predictions_df`. Initially, it includes a column labeled 'Actual Yield,' which contains the actual target values (y) obtained from the test dataset (`subset_y_test`).

W. Adding Predicted Values to DataFrame

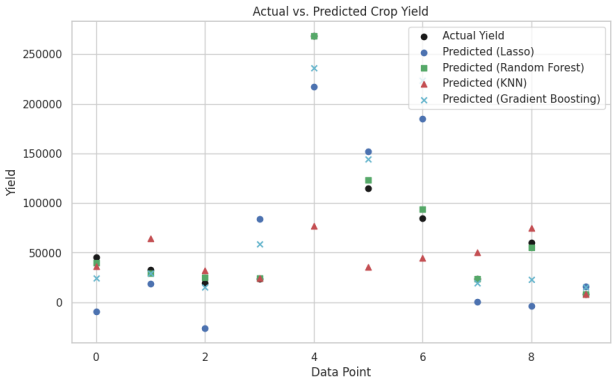
For each model, a new column is added to the DataFrame, with the model's name incorporated into the column name. These columns contain the predicted values for the target variable. Consequently, the resulting data frame contains both actual and predicted values for all the models, facilitating a straightforward comparison of their performance.

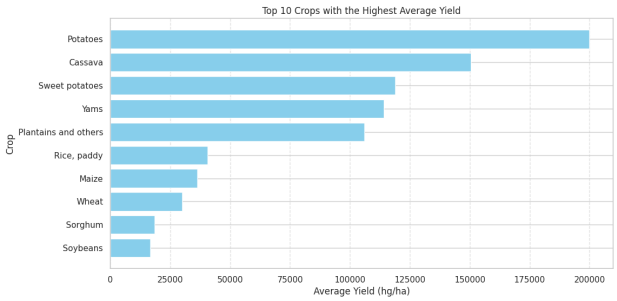
	Actual Yield	Predicted Yield (Lasso)	Predicted Yield (Random Forest)	\
18798	45266.0	-9617.634650	39827.05	
26337	32884.0	18703.957110	29123.53	
7125	19295.0	-26275.075334	24985.85	
16581	24000.0	84162.534318	24125.52	
20367	268098.0	217342.141585	268209.70	
10193	114676.0	151813.372846	123333.66	
6952	84888.0	184877.632975	93582.10	
3498	23724.0	277.737298	23727.09	
2041	60000.0	-3987.614543	55508.90	
10664	8136.0	15924.694737	8136.00	

	Predicted Yield (KNN)	Predicted Yield (Gradient Boosting)
18798	36464.2	24334.115365
26337	64616.2	28943.771867
7125	32091.4	15332.433175
16581	24427.8	59035.715158
20367	77011.2	236095.935831
10193	35469.8	144605.680056
6952	45064.8	223256.299318
3498	50297.4	19698.638705
2041	74578.2	22684.089651
10664	8136.0	15354.649672

X. Creating a Scatter Plot:

After creating the DataFrame, the code uses matplotlib to create a scatter plot for visualizing the actual vs. predicted values. It defines markers and colors for each model. Initializes figure and axis using plt.figure and plt.gca(). Plots actual values as black dots with the 'Actual Yield' label. Loops through models, plotting predicted values with associated colors and markers. Adds title, axis labels, and legend After, it displays the scatter plot using plt.show(). This plot allows you to visually compare how well each model's predictions align with the actual values, making it easier to assess the model's performance.





Now it performs an analysis of crop yield data and generates a horizontal bar plot to visualize the top crops with the highest average yield over the years.

Y. Calculate Average Yield by Crop

The script calculates the average yield for each crop using the `groupby` method. It groups the data by the "Item" (crop) column and computes the mean (average) yield for each group.

Z. Sort Crops by Average Yield

Crops are sorted in descending order based on their average yield, resulting in the creation of a DataFrame named `best_crops`.

. Print the Top N Crops:

The script prints the top N crops with the highest average yield. The value of `top_n_crops` can be customized according to specific requirements.

Top 10 crops with the highest average yield:

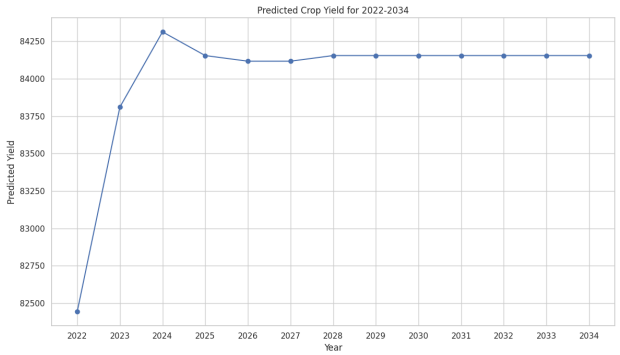
	Item	Yield
3	Potatoes	199801.549579
0	Cassava	150433.403517
7	Sweet potatoes	118999.490664
9	Yams	114140.345927
2	Plantains and others	106041.320144
4	Rice, paddy	40730.434770
1	Maize	36300.137764
8	Wheat	30116.267825
5	Sorghum	18635.777229
6	Soybeans	16731.092771

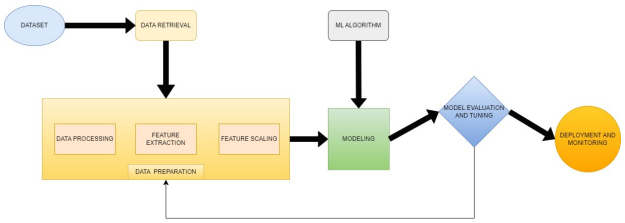
. Creating a Bar Plot:

A horizontal bar plot is generated using the Matplotlib library. This plot displays the average yield on the x-axis and crop names on the y-axis. It includes a title, x-axis label, y-axis label, and other plot customizations to enhance clarity. The y-axis is inverted to present the highest yield at the top of the plot, and grid lines are added to the x-axis for reference.

This script undertakes a process of predicting crop yields based on certain influential factors. Initially, the data representing these factors, termed as `X_relevant`, is scaled using the 'StandardScaler' to have a mean of 0 and standard deviation of 1. A Random Forest regressor is then trained on this scaled data. Concurrently, the code groups another dataset, `merged_df`,, by year and calculates the mean for each variable. Using the ARIMA (AutoRegressive Integrated Moving Average) model, the code forecasts values for specific columns, such as pesticide values and average temperatures, for the years 2022 to 2034. These forecasted values are then scaled using the previously defined scaler. Using the trained Random Forest model, crop yields for the forecasted years are predicted based on these scaled influential factors. Finally, the predictions are organized into a DataFrame for a clear visualization and are printed out.

	Year	Predicted Yield
0	2022	82443.3512
1	2023	83810.0000
2	2024	84312.7800
3	2025	84154.5978
4	2026	84117.0000
5	2027	84117.0000
6	2028	84154.5978
7	2029	84154.5978
8	2030	84154.5978
9	2031	84154.5978
10	2032	84154.5978
11	2033	84154.5978
12	2034	84154.5978





REFERENCES

[1] M. Aruna Devi, D. Suresh, D. Jeyakumar, D. Swamydoss, and M. Lilly Florence, *"Agriculture Crop Selection and Yield Prediction using Machine Learning Algorithms,"* in *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Coimbatore, India, 2022, pp. 510-517, doi: 10.1109/ICAIS53314.2022.9742846.

[2] R. Medar, V. S. Rajpurohit, and S. Shweta, *"Crop Yield Prediction using Machine Learning Techniques,"* in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, Bombay, India, 2019, pp. 1-5, doi: 10.1109/I2CT45611.2019.9033611.

[3] R. Katarya, A. Raturi, A. Mehndiratta, and A. Thapper, *"Impact of Machine Learning Techniques in Precision Agriculture,"* in *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, Jaipur, India, 2020, pp. 1-6, doi: 10.1109/ICETCE48199.2020.9091741.

[4] V. A and S. M. S, *"Prediction of Crop Yield Production with Novel Lasso Regression against Polynomial Regression Algorithm to Achieve Higher Accuracy,"* in *2022 International Conference on Edge Computing and Applications (ICECAA)*, Tamilnadu, India, 2022, pp. 1478-1482, doi: 10.1109/ICECAA55415.2022.9936193.

[5] D. A. -L. Mariadass, E. G. Moung, M. M. Sufian, and A. Farzamnia, *"Extreme Gradient Boosting (XGBoost) Regressor and Shapley Additive Explanation for Crop Yield Prediction in Agriculture,"* in *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, Mashhad, Iran, Islamic Republic of, 2022, pp. 219-224, doi: 10.1109/ICCKE57176.2022.9960069.

[6] V. Choudhary and A. Thakur, *"Comparative Analysis of Machine Learning Techniques for Disease Prediction in Crops,"* in *2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)*, Indore, India, 2022, pp. 190-195, doi: 10.1109/CSNT54456.2022.9787661.

[7] R. Kavitha, M. Kavitha, and R. Srinivasan, *"Crop Recommendation in Precision Agriculture using Supervised Learning Algorithms,"* in *2022 3rd International Conference for Emerging Technology (INCET)*, Belgaum, India, 2022, pp. 1-4, doi: 10.1109/INCET54531.2022.9824155.

[8] S. R. Sani, S. V. Sekhar Ummadi, S. Thota, N. Muthineni, V. S. Srinivas Swargam, and T. S. Ravella, *"Crop Recommendation System using Random Forest Algorithm in Machine Learning,"* in *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, Salem, India, 2023, pp. 501-505, doi: 10.1109/ICAAIC56838.2023.10141384.

[9] S. Chhikara and N. Kundu, *"Machine Learning based Smart Crop Recommender and Yield Predictor,"* in *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, Greater Noida, India, 2022, pp. 474-478, doi: 10.1109/ICCCIS56430.2022.10037678.

[10] R. K. Ray, S. K. Das, and S. Chakravarty, *"Smart Crop Recommender System-A Machine Learning Approach,"* *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 2022, pp. 494-499, doi: 10.1109/Confluence52989.2022.9734173.