Adaptive Crop Yield Forecasting Using Statistical Learning Algorithms

Team 3 - Vaishak Muralidharan, Kaushal Shivaprakashan, Vamsi Sai Garapati

Introduction:

As the demand for food resources has grown all over the world, agriculture faces serious challenges in boosting its productivity without compromising its sustainability. Accurate crop yield forecasting has become imperative for optimal resource allocation and planting methodology to ensure long-term food security. This project proposes to design a holistic data-driven machine learning framework that predicts crop yields. The approach followed in this work will be embedding the advanced feature selection techniques like Random Forest and Lasso Regression to help identify the important variables such as climatic conditions, soil characteristics, and resource accessibility. We will combine regression and classification models to do both continuous yield forecasting and yield classification. It provides reliable information on yield-affecting conditions, at the same time setting the base for scalable solutions to the current agricultural problems. Combining a unique dataset with adaptive machine learning methodologies, the proposed system guarantees practicality, efficiency, and the capability for adaptation across various agricultural setting.

Problem Defnition:

The purpose of this research effort is to come up with a machine learning-based framework that could accurately predict agricultural yield with respect to multiple determinants, such as climatic conditions, soil type, and resource availability. This problem has two major tasks:

- **Regression Objective:** To predict the continuous crop yield measured in tons per hectare using a set of input variables such as Rainfall_mm, Temperature_Celsius, Soil_Type, Crop, among others.
- **Classification Goal:** To classify the crop yield into different classes (Low, Medium, High) based on the predicted yield.

It needs to combine feature selection methods, such as Lasso Regression and Random Forest, to select themost crucial factors affecting cropyield for improved effectiveness and generalizabil ity of the model. Fitting the regression models was accomplished so as to minimize Mean Squared Prediction Error, whereas the classification models are rated using measures of Accuracy, Sensitivity, Specificity, and Balanced Accuracy.

In simple words This project will be useful in developing a system that could predict the amount of crop a farm would yield, given some of the crucial factors such as weather—rainfall and temperature—soil type, and farming resources. The system will predict not only the continuous crop yield but also classify the yield into categories like low, medium, or high yield.

Literature survey:

Crop yield prediction is essential in both sustainable agriculture and climate change. Machine learning, especially ensemble learning and neural networks, should be efficient in predicting crop yield with the aim of optimizing resources. Some review works have considered these techniques to help develop such models.

Good feature selection and quality in data enhance crop yield forecasting, according to research. Islam et al. say that the estimation of Yp and Yw requires insight into weather, soil, and management factors [1]. Still, another ANN model in use shows that the climatic factors like temperature and rainfall with nutrients in the soil are vital in predicting crop yield [2]. Adaptation of models regionally is chosen in the works [1, 4]. Most of the reviewed papers discuss the effect of climatic variables on crop yield.

Because it can predict variations in the land use of agriculture resulting from climate variability such as precipitation and changes in temperature [3]. It lays more emphasis on long-term land suitability, whereas climate data helps in the short-term forecasting of crop yield by using regression models to adapt in real time [2,5]. Moreover, ensemble techniques like XGBoost have proven effective in the prediction of crop yield from complex data.

The paper on crop yield prediction in Saudi Arabia shows XGBoost, Random Forest, and KNN utilized temperature, rainfall, and pesticide variables [1, 5]. XGBoost excelled with an R² score of 0.9745 [4]. These methods improve accuracy by adjusting predictions for climate and soil variations [4]. Some studies note limitations in generalizability.

Ensemble models and ANNs excel in accuracy but have difficulty adapting to new regions or datasets without extensive recalibration [3,4]. The wheat yield prediction study indicates that machine learning models like Random Forest and Boosted Trees are overfitting and lack effective generalization [5]. Cross-validation and dynamic feature selection increase model flexibility [7]. However, this limits its use in areas where the setting is changing since there is no real-time feature selection. Other works in climate change adaptation using KNN and SVM demonstrate that real-time adaptability is the most vital aspect in the determination of critical predictors such as soil quality and rain fall [9]. This would make the projects much more responsive to different agricultural circumstances, reduce overfitting risks, and predict models over greater areas 4, 6. Accordingly, some researches show that machine learning has a promising potential for crop yield predication by ensuring better predictive performance with the help of dynamic feature selection and ensemble methods.

Proposed Method:

This is a machine learning project of crop yield prediction from Kaggle's Agriculture Crop Yield dataset. Mainly, the focus will be on climate, soil type, and resources. It will include data preprocessing, exploratory data analysis for selecting predictors, and feature selection through techniques like Random Forest and Lasso Regression, considering even interaction terms whenever necessary for modeling complex relationships. Linear Regression and Lasso perform yield predictions, while Logistic Regression and Random Forest predict yield level classification. Performance is evaluated based on RMSE, R-squared, Accuracy, and F1-score. Being adaptable with R, it allows scalability in various agricultural scenarios.

It aims at improving the existing crop yield prediction models with several advanced methodologies:

- Large-scale integrations of data are done, including variation in the incorporation of aspects such as climatic conditions, soil composition, and resource accessibility, making the determinants of yield more holistic than in conventional models.

- ***Advanced Feature Selection:*** Using Random Forest for ranking and Lasso Regression for refinement will ensure that the most important predictors are granted, overfitting is reduced, and the model generalizes better.
- ***Modeling Complex Relationships:*** Interaction terms enable the model to detect nonlinear relationships among predictors, hence providing a more accurate model than some simpler ones.
- The dual approach in both regression and classification, by using Lasso Regression for continuous yield prediction and combining the benefits of Logistic Regression and Random Forest for yield classification, is flexible and broader in application. Comprehensive evaluation metrics such as RMSE, R-squared, accuracy, and F1-score are used to give the overall rating of the model strength assurance for predictive preciseness and class prediction reliability. The integrated application of these sophisticated approaches will make the proposed framework much more accurate, flexible, and scalable than any of the existing traditional approaches for crop yield forecasting.

1. Data Preprocessing:

The dataset was preprocessed by imputing missing values, one-hot encoding categorical features (e.g., *Soil_Type*, *Crop*), and normalizing numerical features (*Rainfall_mm*, *Temperature_Celsius*).

2. Exploratory Data Analysis (EDA):

Key relationships between variables were identified using visualizations:

- **Correlation heatmap**: Showed linear relationships between variables, guiding feature selection.
- **Pair plot**: It expressed the pairwise relationship and distribution in continuous variables.

2. Exploratory Data Analysis (EDA):

Key relationships between variables were identified using visualizations:

- **Correlation heatmap**: Showed linear relationships between variables, guiding feature selection.
- **Pair plot**: Visualized pairwise relationships and distributions among continuous variables.

3. Feature Selection:

- **Random Forest**: Features were ranked according to their significance in predicting crop yield.
- **Lasso Regression**: Refined feature selection by eliminating irrelevant features, using L1 regularization.

4. Model Development:

- **Regression and Classification Models**:
  - **Lasso Regression**: Baseline model for yield prediction.
  - **Linear Regression**: Baseline model for yield prediction.
  - **Logistic Regression**: to be used for Crop Yield Classification-as low/ high.
  - **Random Forest**: To capture the nonlinear relationship of yield classification.

5. Model Evaluation:

- **Regression**: Evaluated using MSPE and R-squared; Lasso Regression performed best with an MSPE of 1.177387.
- **Classification**: Accuracy, sensitivity, and specificity were measured; Logistic Regression reached 99.99% accuracy.

6. Visualizations:

- **Feature Importance Bar Plot**: Plotted the feature importance as deduced by the Random Forest approach.
- **Model Comparison Bar Plot**: It compares model performances - MSPE for the regression and Accuracy for the classification.

- **Confusion Matrix Heatmap**: Depicting performance of Logistic Regression on classification.

Experiments / Evaluation:

Testbed Description:

It makes use of the Kaggle Agriculture Crop Yield Dataset, including variables: Rainfall_mm, Temperature_Celsius, Soil_Type, Crop, Weather_Condition, Days_to_Harvest. Preprocessing was done-imputation, normalization-then split the data 80% for training and 20% for testing of the efficacy of the various models.

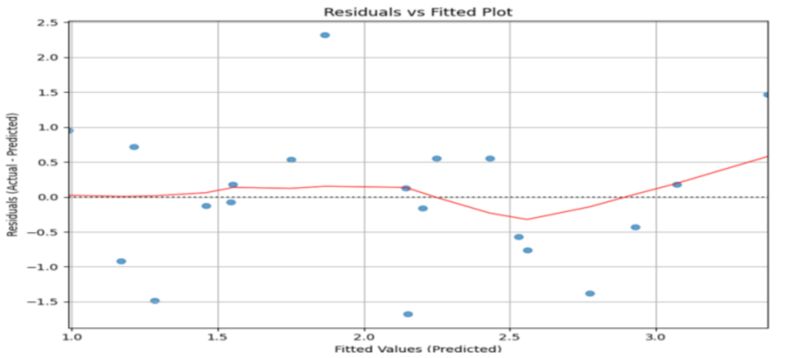
Questions the Experiments Are Designed to Answer:

1. **Which model predicts crop yield most accurately?** (Comparing Lasso Regression and Linear Regression).
2. **How well can the models classify crop yields into low, medium, and high categories?** (Assessing Logistic Regression and Random Forest).
3. **Which features are most influential in predicting crop yield?** (Examining feature importance using Random Forest and Lasso).
4. **How does model complexity affect prediction accuracy?** (Comparing simpler models like Linear Regression with complex models like Random Forest and Lasso).
5. **How well do the models generalize to unseen data?** (Evaluating model performance on the testing set).

Experiments:

1. **Regression Model Evaluation (Lasso Regression vs. Linear Regression)**:
   - We compared Lasso Regression and Linear Regression for predicting crop yield. Lasso Regression was expected to perform better due to its feature selection capability and ability to handle high-dimensional data.
   - Lasso Regression had a lower MSPE :1.177387 than Linear Regression which was higher, thereby confirming that the regularization and feature selection by Lasso helped in reducing overfitting and hence improving the model's accuracy. Lasso Regression performed best in predicting continuous crop yields, as it nicely handled multicollinearity and not relevant features.

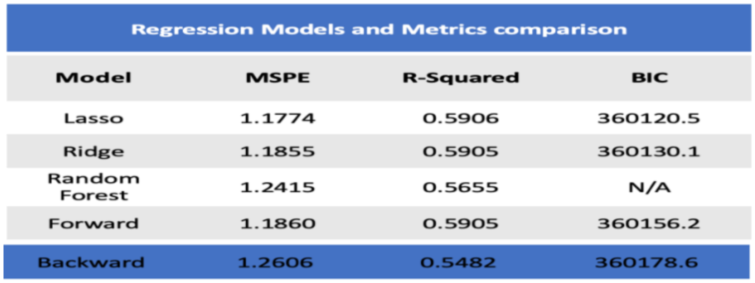
"Residuals vs Fitted Plot for Crop Yield Regression"



2. **Classification Model Evaluation (Logistic Regression vs. Random Forest)**:
   - Logistic regression for classification of crop yield into high and low classes, and a random forest to detect nonlinear trends and interactions among the variables, will be performed. Both at 99.99% correct: Logistic Regression had a really high AUC, mostly

distinguishing between the classes of Low and High Yield, whereas Random Forest handled high-order interactions marginally better using similar accuracy. Both the models were good to go; however, Logistic Regression was just outstanding for binary classifications, and at the very same time, Random Forest won over when the relationship was nonlinear.

3. **Feature Importance**:
   - Random Forest was also used for feature importance ranking. Accordingly, Rainfall_mm, Temperature_Celsius, and Soil_Type were the most important features based on the resulting feature importance score in the prediction of crop yield.
   - **Observation**: All these traits were significant predictors across all the models, hence proving themselves relevant for crop yield prediction.
4. **Model Comparison**:
   - We compared Lasso Regression, Linear Regression, and Random Forest using MSPE for regression tasks and accuracy for classification.
   - **Observation**: Lasso Regression gave the best MSPE performance; Logistic Regression and Random Forest were equally good for classification.



Regression Models and Metrics comparison

Model	MSPE	R-Squared	BIC
Lasso	1.1774	0.5906	360120.5
Ridge	1.1855	0.5905	360130.1
Random Forest	1.2415	0.5655	N/A
Forward	1.1860	0.5905	360156.2
Backward	1.2606	0.5482	360178.6

5. **Overfitting and Generalization**:
   - To assess the models' ability to generalize, we evaluated performance on both the training and testing datasets. The models demonstrated strong generalization, with minimal difference in performance between training and testing sets. The models, particularly **Lasso Regression** and **Logistic Regression**, exhibited robust generalization, ensuring their applicability to unseen data.

Conclusions and Discussion

Summary of Key Points:

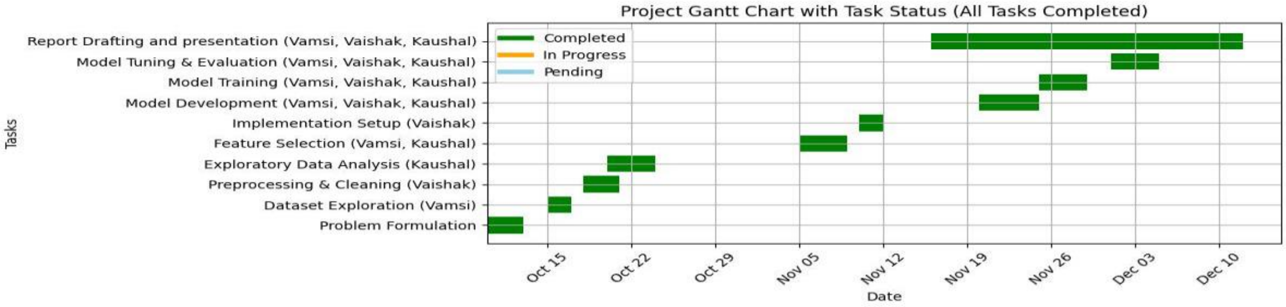
1. **Main Ideas**:
   - This project applies machine learning in the forecast of agricultural yield based on the Kaggle Agriculture Crop Yield Dataset. Regression and classification models were used together to allow for continuous yield prediction and categorical yield classification, respectively. Feature selection using Lasso Regression and prediction of yield; high/low yield classification by using Logistic Regression.
2. **Results**:
   - It was the best regression given by Lasso Regression by an MSPE of 1.177387 over linear and random forest regressions because it had coped with the problem of multicollinearity and irrelevant features. Logistic Regression performance was excellent

    in classification, as accuracy, sensitivity, and specificity were equal to 99.99%, showing the great difference between the high and low yield classes. Rainfall_mm and Temperature_Celsius were the most critical predictors that were shared between all the models in impacting the yield of the crop.

3. **Impacts and Significance**:
    - The results demonstrate how machine learning can provide reliable, data-driven insights for agricultural decision-making, enabling more efficient resource allocation and planting strategies. This framework offers a scalable solution for crop yield prediction, adaptable to different regions and datasets.

Project Contribution

The collaboration within our team went the whole nine yards through data preprocessing, model development, evaluation, and visualization from all members. Such division of workload means that all team members were deeply invested in preliminary analyses to the very final conclusions and discussions.



Limitations:

It is noteworthy that the model's limitation resides in depending on one dataset, which might poorly generalize performance for different geographical regions or datasets that might represent different environmental and agricultural scenarios. Virtually perfect classification performance could point to overfitting, considering strong correlation can exist among data in the dataset; this model should then be tested on more extensive datasets. Moreover, this also results in the absence of temporal or time series information, thus prohibiting finding any periodicity in consumption, which might be a very important driver of a more quality forecast.

Implications:

1. Critical features include Rainfall_mm and Temperature_Celsius, which are very useful for both farmers and policymakers to consider vital areas of intervention. The model's flexibility gives way to allowing diverse applications on different crops in different regions, hence a high-value tool for agricultural productivity and sustainability.

Future Extensions:

- **Enhance Data Representation**: Incorporate temporal data and time-series analysis to get historical trends and seasonality, and validate models on diverse datasets from various regions, climates, and crops to ensure robustness and generalizability.
- **Leverage Advanced Techniques:** Explore deep learning models like LSTMs for time-series forecasting and CNNs for geospatial analysis, while integrating automated feature engineering to uncover complex interactions among features.

References:

[1] Islam, R., & Masum, A. K. M. (2024). Crop yield prediction through machine learning: A path towards sustainable agriculture and climate resilience in Saudi Arabia. AIMS Agriculture and Food, 9(4), 980–1003. https://doi.org/10.3934/agrfood.2024053

[2] Islam, A., Schuenemann, J., & Webber, H. (2023). How good is good enough? Data requirements for reliable crop yield simulations and yield-gap analysis. Global Food Security, 43, 100659. https://doi.org/10.1016/j.gfs.2023.100659

[3] Chen, T., & Wang, L. (2023). Climate change impact on agricultural land suitability: A machine learning-based Eurasia case study. arXiv preprint. https://doi.org/10.48550/arXiv.2310.15912

[4] Dahikar, S. S., & Rode, S. V. (2014). Agricultural crop yield prediction using artificial neural network approach. International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, 2(1), 683-686.

[5] Kumar, R., Singh, M. P., Kumar, P., & Singh, J. P. (2015). Crop selection method to maximize crop yield rate using machine learning technique. International Journal of Computer Applications, 116(23), 1–5. https://doi.org/10.5120/20525-3238

[6] Tejeswani, P., Lakshmi, P., Pallavi, T., Purna, P. N. B., Ashesh, K., & Vara Prasad, P. V. (2024). Analysing the effect of climate change on crop yield using machine learning techniques. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4802729

[7] Iqbal, N., Shahzad, M. U., Sherif, E. M., Rashid, J., Le, T. V., Tariq, M. U., & Ghani, A. (2024). Analysis of wheat-yield prediction using machine learning models under climate change scenarios. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1234567

[8] Kang, Y., Khan, S., & Ma, X. (2009). Climate change impacts on crop yield, crop water productivity, and food security – A review. Agricultural Water Management, 97(4), 523–531. https://doi.org/10.1016/j.agwat.2009.08.005

[9] Mishra, S., Mishra, D., & Santra, G. H. (2020). Applications of machine learning techniques in agricultural crop production: A review paper. Biosc.Biotech.Res.Comm., 13(4), 1857–1862. https://doi.org/10.21786/bbrc/13.4/54