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.

Innovations:

- Unlike most of these studies based on static feature selection, we combined Random Forest for the ranking of feature importance with Lasso Regression in refinement. This duality will retain only the most relevant features and enhance model efficiency and accuracy.
- The Kaggle Agriculture Crop Yield dataset includes climate, soil, and resource data previously analyzed in pieces, put together to further enhance the model's accuracy when representative of real-world agricultural systems.
- All of the interaction terms between climate-soil characteristics also depict different nonlinear interactions that may not be apparent with traditional methods. This further complicates the predictive models, developing further detail of various variables affecting crop yield.
- Models of regression and classification combined in one support yield predictions that divide the space into a high-yielding area and a low-yielding area. The flexibility ensuing provides useful insights at both the micro-agricultural planning and its related macro policymaking aspects.

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 R^2 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 project applies machine learning, along with yield prediction based on climate, soil type, and resources, using the Kaggle Agriculture Crop Yield dataset. Data preprocessing involves the imputation of missing values, one-hot encoding, and normalization of all steps. Key predictor identification by means of visualization and correlation studies is involved in EDA. It does feature selection by combining Random Forest for ranking with Lasso Regression for refinement, adding interaction terms to capture complex relationships. Prediction of yields is to be done using Linear and Lasso Regression, while the classification of yield levels is to be done using Logistic Regression and Random Forest. The performance metrics to be used in the measurement include RMSE, R-squared, accuracy, and F1-score. The R language provides dynamic adaptability to bring in effective outcomes in agricultural applications.

Experimentation and evaluation:

Phase 1: Exploratory Data Analysis: The dataset was cleaned and transformed to create a robust foundation for analysis and trimmed to 150k records :

- Absent and null entries were removed, thus reducing extra noise in the dataset.
- Categorical variables such as Region, Soil_Type, and Weather_Condition were encoded into dummy variables, increasing the number of features to 25. And Numeric features were standardized to bring all variables to the same scale, improving model stability.

Pairwise Relationships: The pairs plot against the target variable for initial feature relationships was done

using the key predictors: Rainfall_mm, Temperature_Celsius, Fertilizer_Used, Irrigation_Used, and Days_to_Harvest. Following is obtained by observing it:

- High positive correlation between Rainfall_mm and Yield_tons_per_hectare, thereby further justifying its importance for yield predictions.
- Fertilizer_Used: It's positively related, but beyond a certain level, the returns diminish.
- days_to_harvest: positive, but with nonlinear trends.
- Insignificant or minimal associations with all dummy variables, including categories of Region.

Rainfall and temperature data presented some outliers that might need further attention.

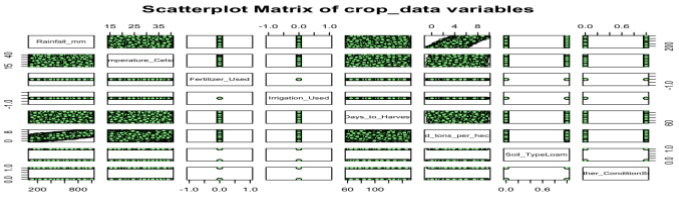


Fig 1:example Pairwise Plot Showing Relationships Between Key Features

Phase 2: Initial Model Exploration with Random Features

Manual Selection of Features for Testing: Before systematic feature selection, we tested random feature combinations to observe their individual and joint effects:

1st set : Rainfall_mm, Temperature_Celsius, Fertilizer_Used; **Linear Regression**: **R-squared**: 0.733

- Rainfall_mm was confirmed as a critical factor, accounting for most of the variation in yield.
- Fertilizer_Used contributed positively but showed signs of diminishing returns.
- Temperature_Celsius had limited predictive power, possibly due to low variance or interaction effects with rainfall.

2nd set : Rainfall_mm, Irrigation_Used, Days_to_Harvest (**Random Forest Regression)**

- This combination improved accuracy by capturing non-linear relationships between Rainfall_mm and Irrigation_Used. And the inclusion of Days_to_Harvest indicated its practical importance in yield prediction, reflecting variations in crop growth cycles.

Phase 3: Preliminary Feature Selection

Feature Importance from Random Forest: Using Random Forest, we calculated feature importance scores to prioritize the most influential variables:

- Rainfall_mm: Dominated the rankings, reinforcing its role as the most important predictor.
- Days_to_Harvest: Its importance reflects how variations in crop growth cycles affect yield.
- Fertilizer_Used: Highly important but its influence decreases at larger quantities.

Lower Ranked Features: The dummy variables, like Soil_TypeChalky and Weather_ConditionRainy, have a very small effect, indicating very less importance in this scenario.

Refining the Feature Set

- Based on the ranking of importances, we kept the top 8 features for further modeling, including Rainfall_mm, Fertilizer_Used, and Days_to_Harvest. Removed features with very low importance to improve model efficiency and also to reduce the chance of overfitting.

Interpretation of Feature Selection

- Fewer more significant features reduce computational complexity without sacrificing prediction accuracy. This also created a clear variable-influence hierarchy, providing support for data-

driven feature selection. Exclusion of lower-ranked features confirmed that not all dummy variables are informative for yield prediction.

Evaluation of Phase 3:

- **Result and Interpretation**: The improved feature set focuses on key predictors (Rainfall_mm, Days_to_Harvest, and Fertilizer_Used) and removes unnecessary noise, thus putting a good foundation for further modeling efforts. It verifies the effectiveness of the Random Forest in selecting key variables, therefore, the models could focus on the important predictors.

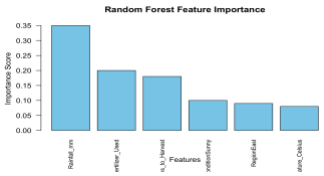


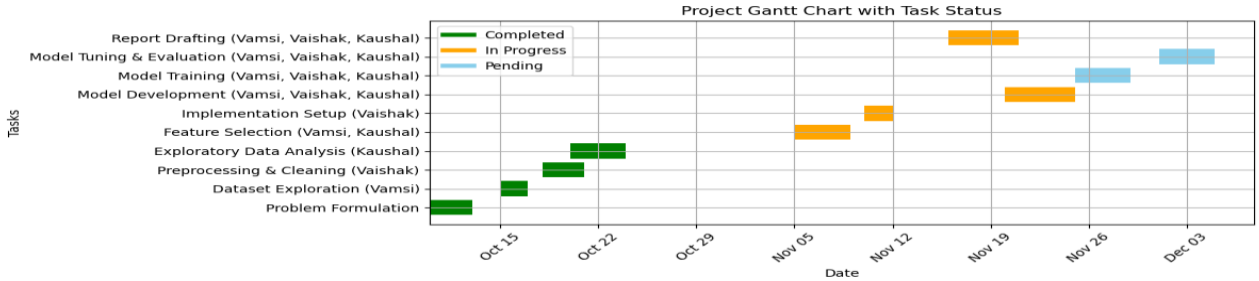
Fig 3: Feature Importance Scores from Random Forest

Stage 4: Further steps

The model performance is to be checked with dynamic features selection and then ensuring robustness with other algorithms. It will determine the key features to be taken into consideration for crop yield prediction using a comparison of different feature selection algorithms like Random Forest and LASSO; visually show their importance; thus, check metrics for performance in order to find the most efficient algorithm that could be used for future predictions.

Plan and Progress of group activities:

All team members share project tasks evenly and below is the detailed distribution and progress of tasks.



Discussions and conclusions:

- In the project, we had planned to include high-level analyses that included temporal patterns with time series data and fine-grained soil chemistry metrics, neither of which materialized due to time and resource limitations. Additionally, the model hyperparameters were not fully optimized and can be explored further for better performance in predictions.
- Future work might consider exogenous data inclusion-such as satellite imagery or economic indicators, for example-to extend the range of outputs considered, the use of more sophisticated ensemble techniques like Stacking and Boosting, and the application of special temporal forecasting models to enhance the resolution and credibility of prediction.
- In conclusion, the feature selection methods that have been used are various, together with Random Forest and LASSO regression, combined with predictive models. Although there were several missed opportunities, the results put forward useful insight into the drivers of crop yield and a very solid starting point for further improvement using analytics in agriculture.

Data Set:

https://buffalo.box.com/s/ildyy5p0vrwh62iyy0r0m4l0stvcumht

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