Corn yield Forecast

Hamed Goldoust
Politecnico di Torino
Turin, Italy

Parsa Taati
Politecnico di Torino
Turin, Italy

Afsoun Abbasi
Politecnico di Torino
Turin, Italy

Abstract

Accurate corn yield prediction is paramount to food security, efficient utilization of agricultural resources, and policy formation. Yield prediction is, however, confounded by the complex interrelationships between weather variables, soil conditions, and management. This study explores the application of advanced machine learning techniques in improving the accuracy of yield prediction. This present study employs a sequence of methods, including Random Forest, XGBoost, Gradient Boosting, Support Vector Machines, and regression models, to examine historical meteorological data with corn yield data. Preliminary preprocessing activities involve data cleaning, feature engineering, normalization, and dimensionality reduction. Model performance is assessed using Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). Results show that Gradient Boosting and Random Forest (with GridSearch) outperform other models, exhibiting higher predictive power. The study highlights the importance of selecting relevant features and applying ensemble learning methods to improve predictive performances. Despite constraints such as data availability and the potential for generalizing findings across various regions, the proposed method presents a good foundation for crop yield prediction with potential improvement through incorporation of remote sensing technologies and deep learning models in subsequent studies.

Project is available at: https://github.com/adsp-polito/2024-P6-CYF

1 Introduction

1.1 Background

Corn is one of the staple crops and provides a large portion of human calorie intake. Its accurate yield forecast is essential to guarantee food security, optimize resources, and make appropriate policy decisions. However, yield forecasting in agriculture is difficult because of many dynamic factors that come into play: weather variability, soil properties, and management practices, among other factors.

1.2 Problem Statement

Traditional yield prediction techniques are likely to employ linear models that are not capable of capturing non-linear and temporal relationships between environmental factors and crop development. This project aims to bridge this gap through the application of advanced machine learning models that can handle complex datasets.

1.3 Objectives

- Create a data preprocessing pipeline to combine meteorological and yield data.
- Apply various machine learning models, such as Random Forest, XGBoost, Gradient Boosting, and regression-based models.
- Compare the model performance with all features and subset of features using MAE and MAPE.
- Offer actionable suggestions to enhance the process of predicting yield.

2 Data Sources and Processing

2.1 Data description

2.1.1 *Meteorological Data:*

- Daily weather variables: mean temperature, maximum temperature, minimum temperature, precipitation, solar radiation, and wind speed.
- Timeframe: 2012–2023.
- Source: Local meteorological stations.

2.1.2 *Corn Yield Data Features:*

- Fresh ear yield
- dry ear yield
- FAO cycle
- yearly records

2.1.3 *Field Management Data:*

Includes sowing and harvest dates, essential for identifying the crop's growth stages.

2.2 Preprocessing Steps

2.2.1 *Data Cleaning:*

- Replaced commas with decimals in yield data
- Converted all numerical values to a standard format
- Dropped rows with missing target values

2.2.2 *Feature Engineering:*

- Variables like temperature, precipitation intensity, and solar radiation were aggregated monthly.
- Created new features such as:
  - Dry days: Days with zero precipitation.
  - Precipitation intensity: Total precipitation divided by the number of wet days.
  - Solar variability: Standard deviation of solar radiation.

2.2.3 *Normalization:*

Applied Min-Max scaling to weather features for uniformity.

2.2.4 *Feature Reduction:*

- Performed correlation analysis to identify low-importance features (Figure 1).

- Retained only features with a correlation >0.3 (Figure 2) with the target variables.

Figure 3 comparison of the distributions for the Targets features in the train dataset using histograms and Kernel Density Estimation (KDE).

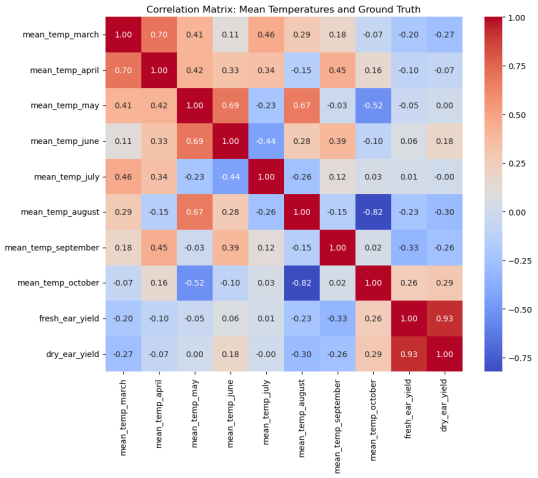

Figure 1

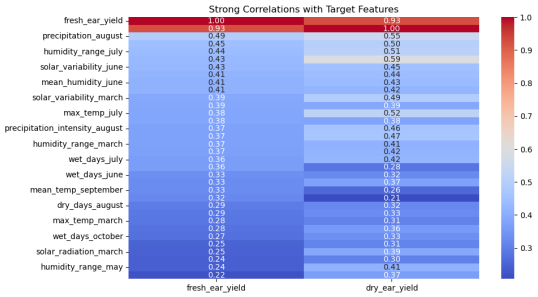

Figure 2

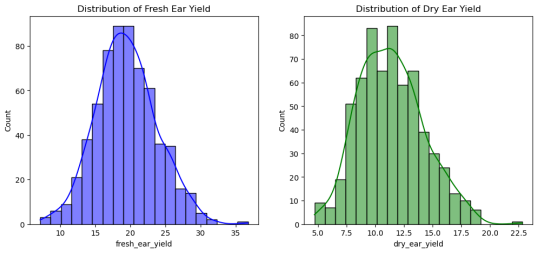

Figure 3

3 Methodology

To develop an effective predictive model for maize crop yield, we explored a range of machine learning techniques, including ensemble methods, kernel-based approaches, gradient boosting models, and regression-based models. These models were selected to capture both linear and non-linear relationships between meteorological variables and crop yield.

RandomForest:
Random Forest is an ensemble learning method that constructs multiple decision trees during training and averages their outputs to enhance predictive performance. This approach helps mitigate overfitting while maintaining robustness against noisy data. Key hyperparameters, such as the number of estimators (n_estimators), maximum tree depth (max_depth), and minimum samples per leaf (min_samples_leaf), were optimized through grid search to improve accuracy and generalizability. The use of bootstrap sampling and feature randomness enables Random Forest to perform well even with high-dimensional datasets [2].

XGBoost:
XGBoost (Extreme Gradient Boosting) is a high-performance implementation of gradient boosting, designed for scalability and computational efficiency. It builds decision trees sequentially, correcting the errors of previous trees while minimizing loss. The model's performance was optimized by tuning hyperparameters such as the learning rate (eta), maximum depth of trees (max_depth), and subsample ratio, which controls the fraction of observations used for each boosting iteration. XGBoost is particularly effective in handling missing data and capturing complex non-linear patterns in yield prediction [4].

Support Vector Machines (SVM):
SupportVectorMachines (SVMs) were employed to model non-linear relationships between meteorological variables and yield. The Radial Basis Function (RBF) kernel was used to map input features to a higher-dimensional space where a linear separation was possible. Additionally, a linear kernel was tested to compare performance on linearly separable data. The regularization parameter (C) and the kernel coefficient (gamma) were fine-tuned to balance model complexity and avoid overfitting. SVMs have been widely applied in agricultural predictions due to their effectiveness in handling high-dimensional feature spaces [5].

Gradient Boosting:
GradientBoosting is another ensemble learning technique that builds trees sequentially, where each tree corrects the residual errors of its predecessors. Unlike XGBoost, which uses optimized tree structures and parallel computations, the GradientBoostingRegressor from scikit-learn was employed in this study. Key hyperparameters, including learning rate, maximum depth, and minimum samples per split, were adjusted to reduce overfitting. Gradient Boosting has demonstrated strong predictive capabilities in agricultural applications, especially when combined with feature engineering techniques [6].

Regression Models:
To establish baseline performance, we implemented multiple regression-based models:

Ridge Regression: A linear regression model with L2 regularization to penalize large coefficients and mitigate multicollinearity.

LassoRegression: Similar to Ridge but employs L1 regularization, which can lead to feature selection by driving some coefficients to zero.

ElasticNetRegression: A hybrid of Ridge and Lasso, balancing

between L1 and L2 regularization.
SGDRegressor: A stochastic gradient descent-based linear regression model that is efficient for large-scale datasets.
These regression models served as benchmarks, providing insights into the effectiveness of more complex non-linear methods in capturing yield variability [7].

3.1 Experimental Step

Training and Test Splits:

- Training: 2012–2020 (year)
- Test: 2021–2023 (year)

Evaluation Metrics:
To assess the performance of the predictive models used in this study, we employed two key evaluation metrics: Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are used to evaluate the accuracy of the models in estimating maize crop yield. These metrics are assessed separately for fresh yield and dry yield to provide a clearer understanding of the model's performance in predicting each type.
MAE measures the average absolute difference between the predicted and actual values, giving an indication of how much, on average, the predictions deviate from the true yield. This metric is useful for understanding the overall error magnitude in the same unit as the yield.
MAPE, on the other hand, expresses the error as a percentage of the actual values. This allows for an easier comparison of prediction accuracy across different scales, highlighting the model's relative performance.
For fresh yield, these metrics help evaluate how accurately the model predicts the weight of freshly harvested maize. For dry yield, they assess the model's ability to estimate yield after the drying process. By analyzing these errors separately, we can determine whether the model performs better in predicting fresh or dry yield and identify areas for improvement.
Mean Absolute Error(MAE)
The Mean Absolute Error (MAE) measures the average absolute differences between predicted and actual values. It provides an intuitive understanding of model accuracy in the same unit as the target variable. A lower MAE indicates a model with better predictive accuracy. The MAE is computed as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

Where:

- y_i represents the actual value of the target variable (fresh or dry ear yield) for the ith sample.
- $\hat{y}_i$ is the predicted value.
- n denotes the total number of observations.

MAE is particularly useful for interpreting prediction errors in real-world scenarios. Since it does not square the differences, it does not disproportionately penalize large errors, making it robust against outliers [3].
Mean Absolute Percentage Error(MAPE):
The Mean Absolute Percentage Error (MAPE) is a scale-independent metric that measures the average percentage error between actual and predicted values. It is widely used in forecasting and agricultural yield prediction because it expresses errors as a percentage, making it easier to compare across datasets with different scales. MAPE is calculated using the formula:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{2}$$

where the terms y_i, $\hat{y}_i$, and n are defined as in the MAE formula. MAPE provides a clear understanding of the relative error in predictions, making it highly interpretable. However, one limitation of MAPE is its sensitivity to very small actual values (y_i), which can inflate the percentage error disproportionately [1].

3.2 Hyperparameter Tuning

Random Forest with Grid Search is an optimization technique used to improve model performance by systematically searching through a predefined set of hyperparameters. It evaluates different combinations of hyperparameters, such as the number of trees (n_estimators), maximum tree depth *(max_depth)*, minimum samples for splitting *(min_samples_split)*, and feature selection criteria *(max_features)*. By using cross-validation, GridSearch identifies the best-performing parameters, enhancing model accuracy and robustness while preventing overfitting.

4 Result and analysis

4.1 Model Performance

The evaluation results from Table 1: Evaluation Metrics for Machine Learning Models Using All Features indicate that Random Forest (GridSearch) and Gradient Boosting performed best, achieving the lowest MAE and MAPE values for both fresh and dry ear yield predictions. XGBoost also showed competitive performance, particularly after hyperparameter tuning. In contrast, SGDRegressor had the highest error values, especially for fresh ear yield, making it unsuitable for this task.
Table 2: Evaluation Metrics for Machine Learning Models Using Reduced Features shows that feature reduction had minimal impact on high-performing models like GradientBoosting, RandomForest, and XGBoost, demonstrating their robustness. However, SGDRegressor and Ridge Regression suffered significant performance drops, indicating their reliance on a larger feature set. SVM showed slight improvements with reduced features, suggesting some noise reduction benefits.
Overall, based on Table 1 and Table 2, Gradient Boosting is the most reliable model for maize yield prediction, followed closely by RandomForest(GridSearch). Feature selection should focus on maintaining only the most relevant features to optimize model efficiency without compromising accuracy. SGDRegressor and Ridge Regression should be avoided due to their poor generalization performance.

4.2 Visualizations

Figure 4: Mean Absolute Error (MAE) Comparison for Fresh Ear Yield

The plot indicates that RandomForest (GridSearch) and Gradient Boosting have the least MAE values, suggesting good predictive power. SGDRegressor has the highest MAE, suggesting poor generalization. Feature reduction has little impact on top-performing models but a considerable impact on poorer models such as SGDRegressor and Ridge Regression.

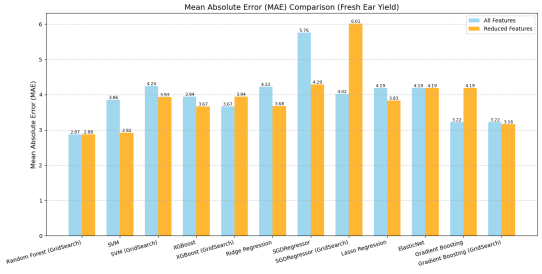

Figure 4

Figure 5: Mean Absolute Percentage Error (MAPE) Comparison for Fresh Ear Yield

MAPE values also reinforce the observation that RandomForest (GridSearch) and GradientBoosting are the best predictors with minimal relative error. Models like SVM and SGDRegressor possess higher MAPE values, which reflect worse predictive ability. Feature reduction slightly increases the percentage error in weaker models.

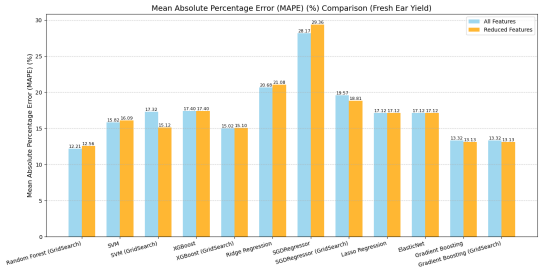

Figure 5

Figure 6: Mean Absolute Error (MAE) Comparison for Dry Ear Yield

Like fresh ear yield, RandomForest (GridSearch) and GradientBoosting have the lowest MAE values for dry ear yield, showing their dominance. Feature reduction impacts neither of these models but greatly reduces SGDRegressor and Ridge Regression performance, suggesting their dependence on a larger set of features.

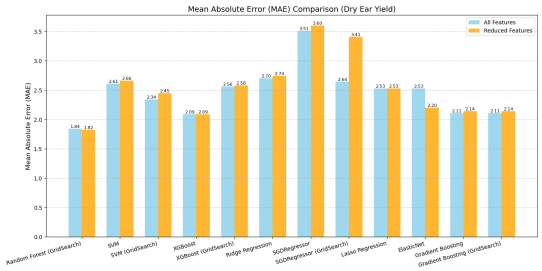

Figure 6

Figure 7: Mean Absolute Percentage Error (MAPE) Comparison for Dry Ear Yield

This line indicates that RandomForest (GridSearch) and Gradient Boosting continue to lead with the lowest MAPE values, which provide good relative predictions. On the contrary, SGDRegressor and Ridge Regression have the highest errors, reaffirming their poorer performance. Feature reduction incrementally raises MAPE in a majority of the models but impacts weaker models more. In general, RandomForest (GridSearch) and GradientBoosting are the most stable models with good predictive strength and robustness to feature reduction. SGDRegressor and Ridge Regression struggle with both MAE and MAPE, particularly when fewer features are included.

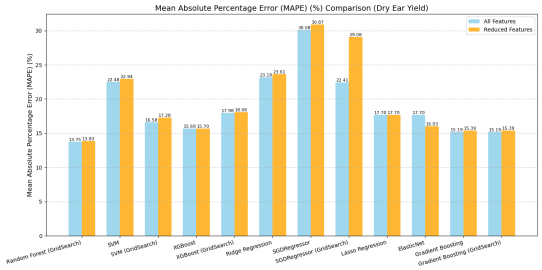

Figure 7

Figure 8 shows scatter plots of predicted vs. actual yield values for fresh ear yield and dry ear yield on all and subset feature sets. The red dashed line is the ideal 1:1 correlation where predictions are equal to actual values.

In the top-left plot (All Features: Fresh Ear Yield), predictions are well-placed along the diagonal, showing good model performance. There is some spread in the higher yield range. The top-right graph (All Features: Dry Ear Yield) illustrates comparable alignment, with the predicted values clustering about the actual yield in most cases. There is, however, some under-prediction for the higher-yielding examples. The bottom-left graph (Reduced Features: Fresh Ear Yield) has more scatter about the ideal line, which suggests that feature reduction had the effect of reducing predictive accuracy slightly, with more spread in the predictions.

The bottom-right graph (Reduced Features: Dry Ear Yield) shows a moderate reduction in predictive accuracy compared to the use of all features. The predictions are still bunched around the actual

values, but with greater variance, indicating that the model's performance is reduced somewhat with fewer features.
Overall, feature reduction affects the prediction accuracy, but models maintain satisfactory performance, especially for dry ear yield. The diagonal pattern indicates that the models still identify significant yield patterns, even though they perform best with the full set of features.

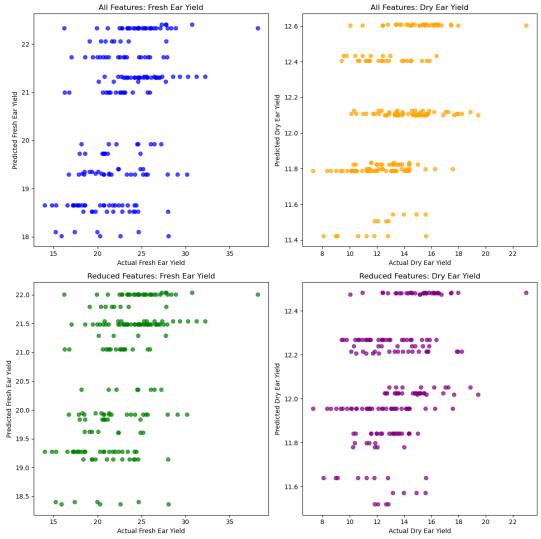

Figure 8

5 Challenges and Limitation

One of the main issues in this study was data availability, as an absence of meteorological data for some time intervals impacted the completeness and reliability of model training. Missing or incomplete weather reports made it necessary to employ data imputation methods, which could have introduced extra uncertainty into the predictions.
Another key limitation was the variability of the weather, which had a real impact on crop yields. Since meteorological variables such as temperature, rainfall, and solar radiation are nondeterministic in nature by definition, their variation was hard to generalize in order to obtain meaningful prediction accuracy. Weather extremes such as drought or heavy rain may have huge impacts on the yields, and it would be hard for the model to generalize across different years.
There is also a problem of model generalizability when applying trained models to other geographic regions or crop varieties. The models were trained on maize yield data from one specific geographic region, and their performance might not be generalizable to other regions with varying climatic conditions, soil, or agricultural practices. To improve generalizability, future research should incorporate a diverse set of datasets across different regions and crop varieties to make the models more robust and adaptable.

6 Future Work

To improve the precision and applicability of maize yield forecasting models, future studies should also attempt to incorporate remote sensing data to examine spatial variations in yields. Satellite images and UAV observations can offer useful information on crop health, soil moisture content, and vegetation indices, allowing for more accurate and region-specific yield forecasts.
Furthermore, research into deep learning architectures such as Long Short-Term Memory (LSTM) networks and Transformer models can potentially make the model even better at detecting complex temporal patterns in meteorological and agricultural development data. These architectures have demonstrated an impressive capability to deal with sequential data and can potentially perform better than traditional machine learning methods in detecting long-term patterns in collections of agricultural data.
In addition, expanding the dataset to include multi-location field trials is crucial in making the model strong and generalizable. By incorporating data for different climatic and soil types, the predictive models can be adapted for application on a larger scale, ensuring they are less specific to location and work across a range of farming areas.

Conclusion

This study demonstrates that advanced machine learning methods, namely Gradient Boosting and Random Forest with hyperparameter tuning, yield significant improvements in the forecasting of corn yields. These methods are better at capturing complex interactions among meteorological factors and crop yield than traditional regression models. The findings demonstrate the importance of feature selection because shrinking the dataset size had negligible impacts on high-quality models but a dramatic impact on low-quality models. Limitations such as data availability, weather variability, and model generalizability remain, emphasizing the need for further research. Remote sensing data, deep learning architectures, and multi-regional datasets need to be explored in future studies to enhance model robustness and applicability. Through integration of these advancements, machine learning can potentially play a significant role in improving agricultural forecasting and decision-making.

References

[1] J. Scott Armstrong. 1985. *Long-range forecasting: From crystal ball to computer.* John Wiley & Sons.
[2] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[3] Tianfeng Chai and Roland R Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geoscientific Model Development* 7, 3 (2014), 1247–1250.
[4] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. (2016), 785–794.
[5] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20 (1995), 273–297.
[6] Jerome H Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Annals of statistics* (2001), 1189–1232.
[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Model	MAE (Fresh Ear Yield)	MAPE (Fresh Ear Yield) %	MAE (Dry Ear Yield)	MAPE (Dry Ear Yield) %
RandomForest (GridSearch)	**2.87**	**12.21**	**1.84**	**13.75**
SVM	3.86	15.82	2.61	22.48
SVM (GridSearch)	4.24	17.32	2.34	16.58
XGBoost	3.94	17.40	2.09	15.69
XGBoost (GridSearch)	3.67	15.02	2.56	17.98
Ridge Regression	4.22	20.68	2.70	23.18
SGDRegressor	5.76	28.17	3.51	30.08
SGDRegressor (GridSearch)	4.02	19.57	2.64	22.41
Lasso Regression	4.19	17.12	2.53	17.70
ElasticNet	4.19	17.12	2.53	17.70
Gradient Boosting	3.22	13.32	2.11	15.19

Table 1: Evaluation Metrics for Machine Learning Models Using All Features

Model	MAE (Fresh Ear Yield)	MAPE (Fresh Ear Yield) %	MAE (Dry Ear Yield)	MAPE (Dry Ear Yield) %
RandomForest (GridSearch)	**2.92**	**12.56**	**1.82**	**13.83**
SVM	3.93	16.09	2.66	22.94
SVM (GridSearch)	3.67	15.12	2.45	17.28
XGBoost	3.94	17.40	2.09	15.70
XGBoost (GridSearch)	3.68	15.10	2.58	18.08
Ridge Regression	4.29	21.08	2.74	23.61
SGDRegressor	6.01	29.36	3.60	30.87
SGDRegressor (GridSearch)	3.83	18.81	3.41	29.08
Lasso Regression	4.19	17.12	2.53	17.70
ElasticNet	4.19	17.12	2.20	16.03
Gradient Boosting	3.16	13.13	2.14	15.39

Table 2: Evaluation Metrics for Machine Learning Models Using Reduced Features

Hyperparameter	Description	Values Explored
`n_estimators`	Number of trees in the forest	{100, 200, 300}
`max_depth`	Maximum depth of the trees	{10, 20, None}
`min_samples_split`	Minimum samples required to split an internal node	{2, 5, 10}
`min_samples_leaf`	Minimum samples required at a leaf node	{1, 2, 4}
`max_features`	Number of features considered for the best split	{`"sqrt"`, `"log2"`, 0.8}
`bootstrap`	Sampling with replacement (True/False)	{True, False}

Table 3: Hyperparameter tuning values explored for a Random Forest model