Machine Learning to Predict Corn Yield in Iowa

Jack Zamary
DATA 201: Introduction to Data Science
Instructor: Dr. Rohaifa Khaldi
Due Date: **12/15/2025**

Abstract

Maize (corn) is one of the most widely cultivated crops globally, with production exceeding one billion tons in 2023 (FAOSTAT). It plays a central role in global food systems as a staple in many diets, primary livestock feed, and a key input in biofuel and pharmaceutical industries. The United States is the world's largest producer of corn, with Iowa consistently ranking among the highest-producing states (USDA). My final project develops six machine learning models to predict corn yield in Iowa using drought indices and environmental variables. It is important to improve the ability to estimate yield under varying conditions, which is essential for supporting agricultural planning, meeting growing food demand, and ensuring livestock feed availability.

Contents

1 Introduction

Corn yield is a central component of the American agricultural system. Many environmental and climatic factors influence annual yield outcomes, with climate variability being among the most significant. To evaluate how these factors relate to corn yield, a set of input features (X) was selected to train a series of machine learning models. The features were Year, Agricultural District (Division), Total Precipitation, Mean Temperature, and the Mean Palmer Drought Severity Index (PDSI).

The environmental features (Total Precipitation, Mean Temperature, and PDSI) were calculated specifically for the corn growing season, which spans from April (4) through September (9). This period is the most critical window during which temperature, rainfall, and drought stress can positively or negatively shape final yield outcomes (Westcott, 1989).

The primary goal of this project is to use these features to predict annual corn yield in Iowa across six machine learning models. The models developed and compared include Linear Regression, Polynomial (Non-Linear) Regression, K-Nearest Neighbors (KNN), Decision Trees, Support Vector Regression (SVR), and Random Forest

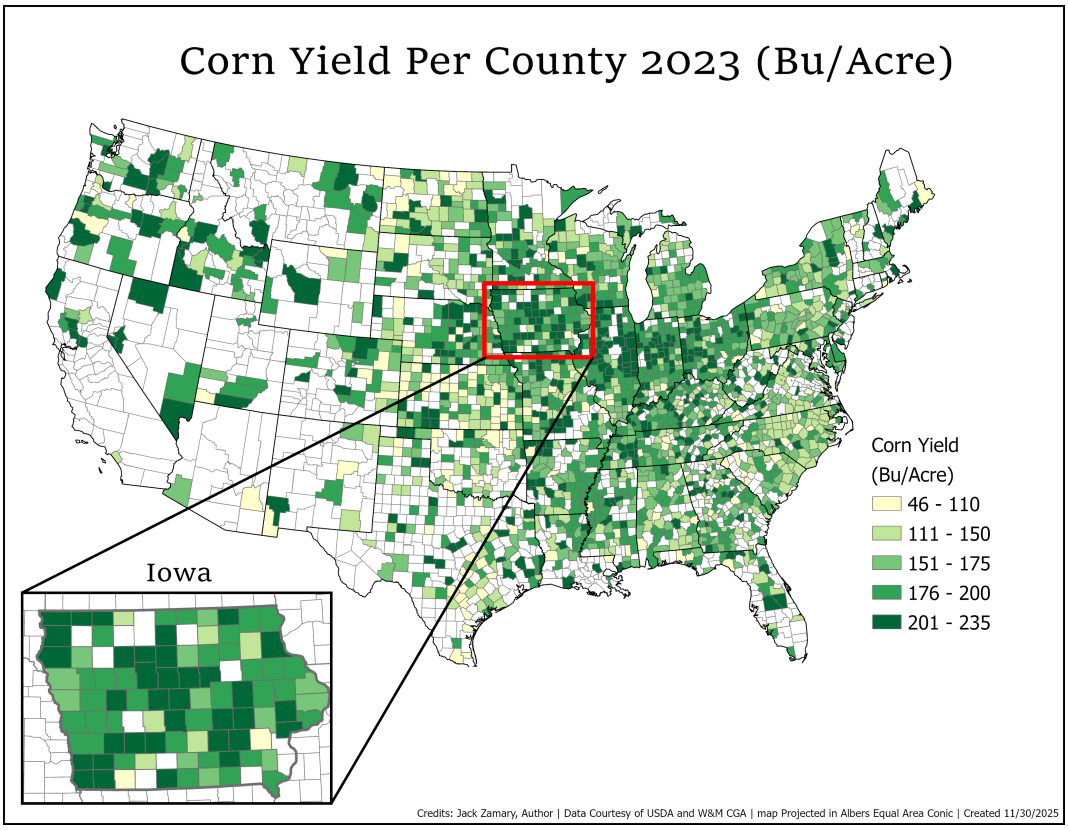

Figure 1: Corn yield by county in Iowa (bushels per acre). Iowa consistently ranks in the top five states for corn production, with a majority of counties yielding over 200 bushels per acre annually, demonstrating its importance as a key study area for crop yield modeling.

2 Data Description and Preprocessing

2.1 Data Sources and Assembly

Annual corn yield data were obtained from the United States Department of Agriculture (USDA) National Agricultural Statistics Service Quick Stats database. Environmental variables were sourced from the National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Information using the Climate at a Glance Divisional Time Series tool. Data were collected for all nine agricultural districts and climate divisions within Iowa, which are geographically aligned.

Prior to exploratory analysis, all downloaded CSV files were consolidated into a single dataframe. Climate data were filtered to include only the growing-season months, and precipitation, temperature, and Palmer Drought Severity Index (PDSI) values were aggregated using the sum, mean, and mean, respectively. Columns were renamed for clarity, and the processed climate dataset was merged with the corn yield dataset using the Agricultural District identifier. County-level yield values were summed to produce annual district-level yield values. The final dataset was inspected for missing values and structural inconsistencies.

2.2 Exploratory Data Analysis

Once the dataset was cleaned and structured, exploratory data analysis (EDA) was performed. Matplotlib and Seaborn were used to visualize long-term yield trends across Iowa's agricultural districts (2000–2024) and to generate a correlation matrix illustrating relationships among all features (X). Additional scatter plots were produced to further examine the association between corn yield and each environmental predictor.

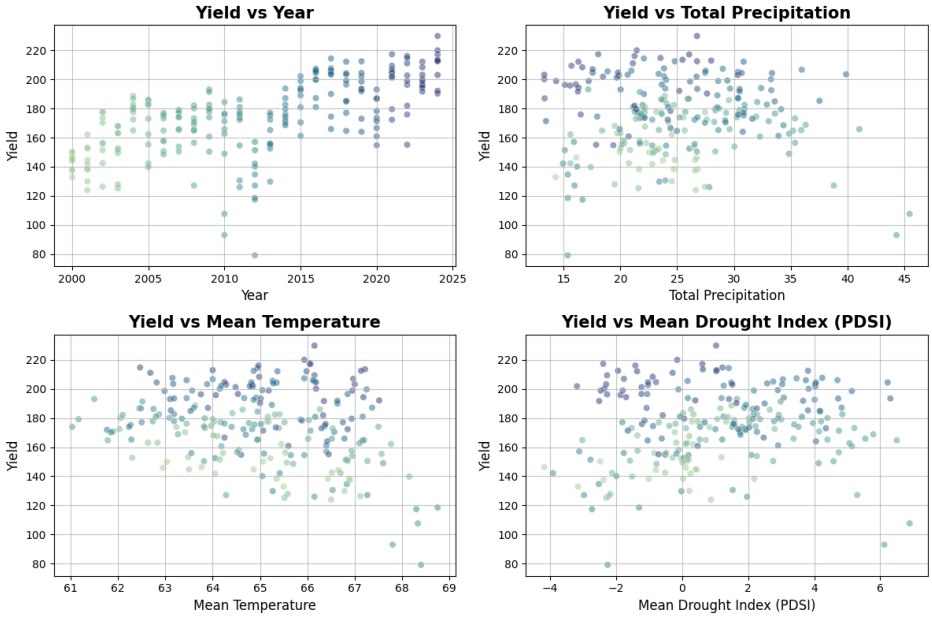

Figure 2: Year, Precipitation, Temperature, and PDSI in correlation to yield (Bu/Acre). The data spread indicates no extreme outliers and some correlation.

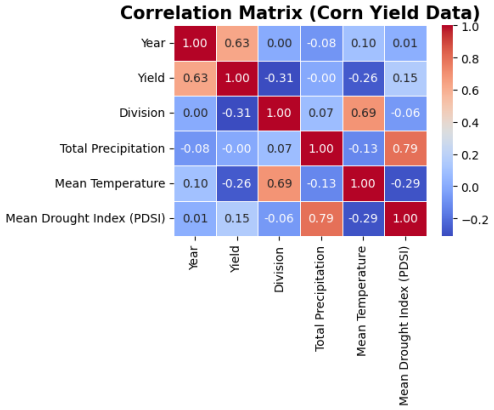

Figure 3: Correlation matrix to represent the relationship between the data. Due to large variability in climate values, there are not too many outstandingly high values.

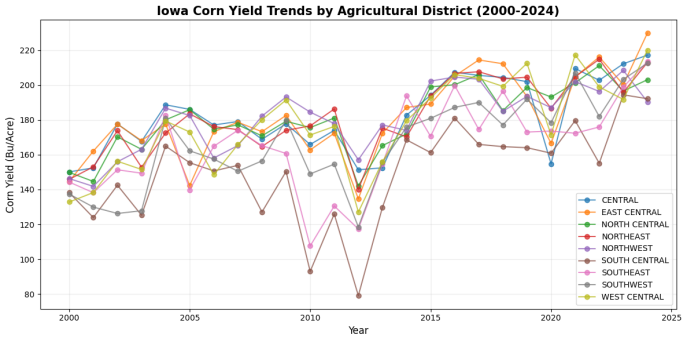

Figure 4: Corn Yield (Bu/Acre) by Iowa's nine Agricultural Districts from 2000 to 2024. Overall positive trend can be attributed to an increase in demand and technological advances.

2.3 Train, Validation, Test Split

Data were subsequently partitioned into training, validation, and testing subsets using a 70:15:15 split, a proportion commonly used in practice and in course materials to balance model learning and unbiased evaluation. This resulted in 162 training samples, 29 validation samples, and 34 testing samples, for a total of 225 observations. All features (X) were standardized prior to model development using the standard scalar method to ensure that features with different units and scales contributed appropriately to the learning process.

3 Methods

3.1 Modeling Approach

The models trained in this project were selected based on those introduced in class, and each was implemented in its regression form because the target variable, corn yield (Bu/Acre), is continuous. The chosen predictor variables are appropriate, as they capture environmental conditions known to influence crop productivity. For example, drought conditions that are reflected by high PDSI values and low precipitation values often lead to reduced yields. The combination of environmental factors (precipitation, temperature, and PDSI), physical factors (agricultural district), and temporal information (year) enables the models to identify and learn trends across space, time, and climate conditions.

The following models were chosen to evaluate the data. The rationale for choosing each model is listed below.

- **Linear Regression**
  This model serves as the baseline, as it fits a simple linear relationship to the data. Ridge and lasso regularization techniques were also trained on the data to prevent overfitting by adding a penalty to the model's cost function.

- **Non-Linear Regression**
  This model captures nonlinear relationships that a standard linear model cannot. It is particularly useful for modeling diminishing returns in variables such as precipitation or temperature.
- **K-Nearest Neighbors (KNN) Regressor**
  Because the dataset is relatively small, KNN is well-suited for identifying yield patterns based on local similarities in the data.
- **Decision Tree Regressor**
  This model naturally handles interactions between variables and offers interpretability by showing which features drive splits in predicted yield.
- **Support Vector Regressor (SVR)**
  SVR is effective for smaller datasets and can model complex, non-linear patterns using kernel functions. It also performs well in the presence of noise, which characterizes much of the climate and yield data.
- **Random Forest Regressor**
  As a powerful ensemble method, Random Forest reduces overfitting and improves predictive performance by averaging many decision trees. Because of the high degree of interaction among environmental variables in this dataset, this model is particularly appropriate.

3.2 Hyperparameter Tuning

Hyperparameter optimization (HPO) was performed for all models except the standard linear regression (though HPO was applied to the regularization strengths in Lasso and Ridge). Grid search was used to evaluate a range of candidate hyperparameters and identify those that achieved the highest validation performance while maintaining a small gap between validation and testing scores to reduce overfitting. The parameters tuned included alpha values, K-values, polynomial degree, decision tree depth, C-values, and the number of trees in ensemble models. A random state of 42 was used when applicable, and validation procedures were used to ensure reproducibility, consistent with class practice.

Table 1: Model Performance Comparison

Model	**Val.** R^2	**Train** R^2	**Val. RMSE**	**Train RMSE**	**Val. MAE**	**Train MAE**
Linear Regression	0.502	0.488	19.482	18.307	15.001	13.935
Polynomial Regression	0.672	0.656	15.817	15.016	12.208	11.727
K-Nearest Neighbors	0.636	0.761	16.672	12.500	12.370	9.523
Decision Tree	0.799	0.819	12.390	10.878	9.462	7.967
Support Vector Machine	0.674	0.753	15.763	12.722	12.674	8.666
Random Forest	0.856	0.947	10.466	5.873	8.550	4.292

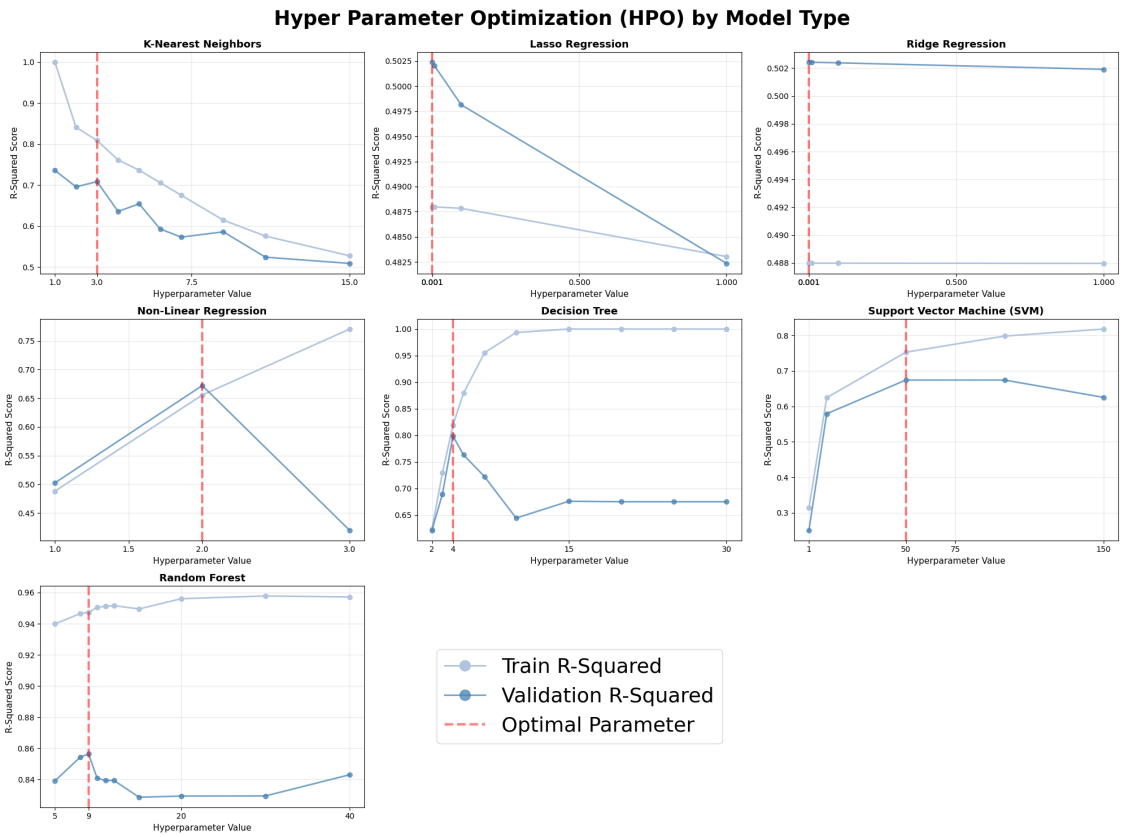

Figure 5: The graphs depict grid search hyperparameter optimization, which was used across all models. Training R-Squared (light blue) and validation R-Squared (dark blue) are shown for each tested hyperparameter value. The optimal configuration (red line) was selected by maximizing validation performance, balancing model complexity with generalization ability.

Hyperparameters were optimized by maximizing validation R-Squared values along with looking at RMSE and MAE. Training performance was also monitored to prevent cases of overfitting and underfitting the data. Models showing close alignment between training and validation scores along with high validation scores were selected as the final models, ensuring generalization and reproducibility. All final hyperparameters were tested and compared against other results to determine the three final models that best fit the data.

Table 2: Hyperparameter Optimization Results

Model	**Hyperparameter Tuned**	**Optimal Value**
Linear Regression	Alphas (Ridge/Lasso)	0.001
Non Linear Regression	Polynomial Degree	2
K-Nearest Neighbors	Number of Neighbors (k)	3
Decision Tree	Max Depth	4
Support Vector Machine	Regularization (C)	50
Random Forest	Number of Estimators	9

4 Results

4.1 Model Performance

The final models selected were K-Nearest Neighbors, Decision Tree, and Random Forest. These models performed the best on the validation set, as they captured nonlinear relationships more effectively. R-squared was chosen as the validation metric of choice, as it measures the proportion of variance in the target explained by the model, independent of the target's units and scale. Environmental data often contain noise, complex interactions, and high variability, which makes flexible models particularly well-suited for this task. Decision Trees identify key structure in the data by partitioning features into meaningful regions, KNN smooths local noise by averaging nearby observations, and Random Forests further reduce overfitting by aggregating predictions across many decorrelated trees. Together, these properties explain why these three algorithms achieved the highest validation performance.

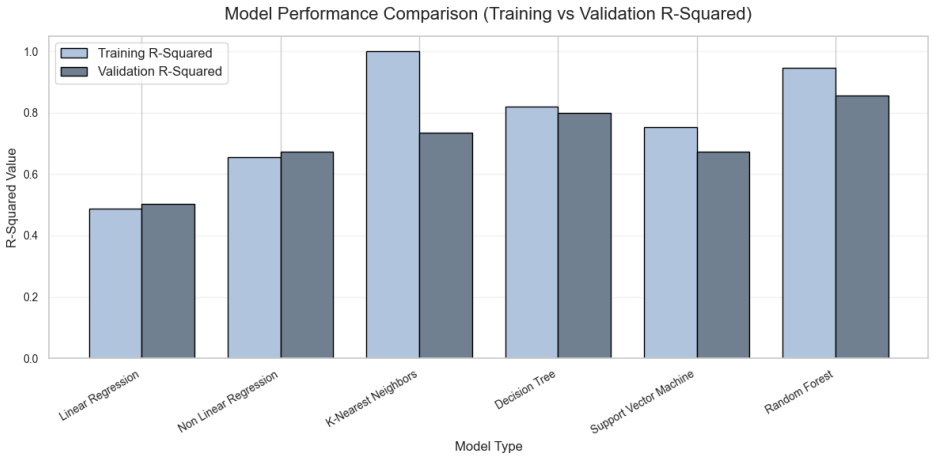

Figure 6: The light blue lines represent training R-Squared scores, and the dark blue lines represent validation R-Squared scores for each model. KNN, Decision Tree, and Random Forest have the highest validation scores.

Among the three final models, the Random Forest achieved the highest R-squared value at 0.769. This was more than 0.2 higher than both KNN and the Decision Tree. Random Forests typically outperform single trees because they reduce overfitting by training many decorrelated trees through bootstrap sampling and averaging their predictions. This ensemble structure lowers variance and helps the model generalize better, which is especially beneficial for noisy or complex environmental datasets.

4.2 Model Interpretation

Out of the final models, both KNN and the Decision Tree performed similarly to the baseline linear regression model. Their moderate performance suggests that while they can capture some nonlinear patterns, they remain sensitive to noise and variability in the dataset. The Random Forest model performed substantially better, achieving a much higher R-squared

Table 3: Performance of Top Three Models

Model	Testing R-Squared
K-Nearest Neighbors	0.499
Decision Tree	0.501
Random Forest	0.769

on the test set and producing predictions that closely matched the true values. This improvement highlights Random Forest's strength on medium-sized datasets with noisy, environmentally driven patterns. Being an ensemble learning method, it managed to fit the data well and derive better-aligned results than simpler models.

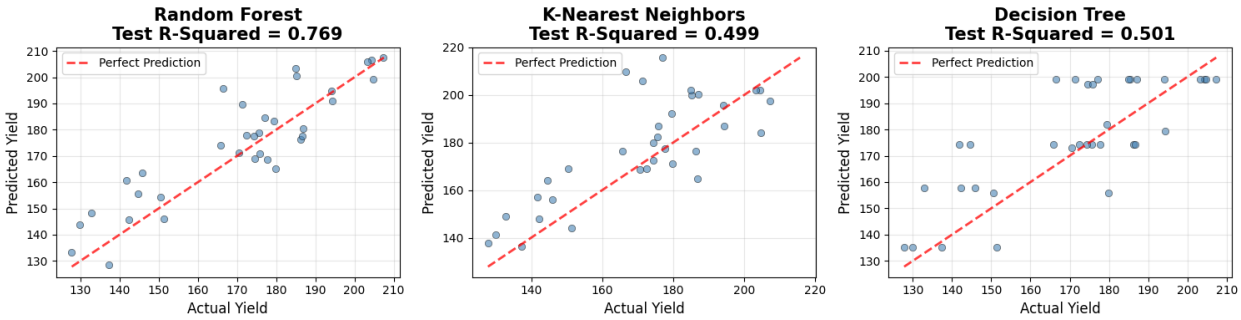

Figure 7: The three final machine learning models were used to generate predictions on the test dataset. Among them, the Random Forest model performed the best, achieving an R-squared value of 0.769, as shown by the close alignment between the predicted and actual yield values.

Both the KNN and Decision Tree models underperformed relative to expectations, largely due to overfitting. Choosing a value of three for k (k=3) might have been too small. This could be resolved by increasing the amount of training data or choosing a larger, more stable k value, which would likely improve performance.

Similarly, the Decision Tree model performed poorly, as it is prone to overfitting. With this dataset being smaller and noisier, overfitting is likely to occur as the model begins to split too specifically based on patterns in the training data. With more data or additional regularization (e.g., limiting depth, pruning), the model could better generalize and identify meaningful structure in environmental variability. This methods slightly falls outside the scope of the class but would be interesting to consider and compare with the Random Forest model.

Lastly, it was expected that the Random Forest model would perform the best. It resolves many of the issue that occur within a standard Decision Tree model by building out multiple trees. Also a form of ensemble learning, its ability to average the final prediction of all the individual decision trees within the forest is quite powerful.

5 Discussion and Conclusion

When working with complex and highly variable environmental factors in a relatively small dataset, the simpler models used struggled to generalize the input features. The combination of noisy predictors and limited samples made it difficult for many of the tested models to learn stable patterns, leading them to overfit the training set. While testing, R-Squared values remained decent, and overfitting became evident during testing, where non-ensemble models were unable to predict yield values with high accuracy.

Given these challenges, it is unsurprising that the Random Forest model performed the best. As an ensemble approach, it reduces variance by averaging predictions across many trees, allowing it to handle noisy, nonlinear, and interacting environmental variables far more effectively than the standalone models. This allowed it to achieve the highest R-squared scores across training, validation, and testing.

Corn has played a central role in U.S. agriculture for centuries, and understanding its production dynamics continues to be of importance. Beyond food, corn contributes to a wide range of industries, including paper, plastics, adhesives, inks, cosmetics, and even fireworks (Mathew Wills). Understanding yield patterns is crucial for economic stability, supply-chain planning, and environmental management.

Looking ahead, model performance would likely improve with a larger dataset and a stronger temporal component (increasing the years, hence, increasing the data overall), enabling better detection of long-term climatic trends. Adding more features, such as soil characteristics, nutrient levels, or pH, would also be beneficial, helping models to further refine predictive power. Overall in the context of this class and final project, the data collected was sufficient, and it was beneficial to explore various implications of model parameters and tune various models to maximize performance.

6 References

1. USDA National Agricultural Statistics Service. (2024). Quick Stats Database. Retrieved from `https://quickstats.nass.usda.gov/`
2. NOAA National Centers for Environmental Information. (2024). Climate at a Glance: Divisional Time Series. Retrieved from `https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/`
3. Westcott, P. C., & Westcott, P. C. (1989). An Analysis of Factors Influencing Corn Yields. Unknown. https://doi.org/10.22004/AG.ECON.151701
4. Westcott, P. C., & Jewison, M. (1997). *Weather Effects on Expected Corn and Soybean Yields*. USDA Economic Research Service.
5. Wills, Mathew. JSTOR Daily, 7 Sept. 2020, `https://daily.jstor.org/corn-is-everywhere/`. Accessed 2025.