# CROPS AND THE CLIMATE: PROJECT REPORT

Luca Bombelli and Kai Bengston

## 1. Abstract

Climate change poses one of the greatest threats to the planet in world history. On top of increasing extreme weather events and destruction of ecosystems, climate change also threatens the global food supply. This project seeks to develop an optimization tool for the agricultural industry by training a machine learning algorithm on climate data, with the goal of taking regional climate data and predicting crop yields.

## 2. Introduction

The threat to food security is ever increasing, and while in first world countries many people take access to healthy, fresh, and safe food for granted, famine is a widespread issue throughout the world. The agricultural industry is a keystone of both the economy and of world food security, and climate change poses an enormous threat to its stability. By creating a tool that would allow farmers to regional predict the success of certain crops, it could optimize crop production in the face of consistently changing weather patterns.

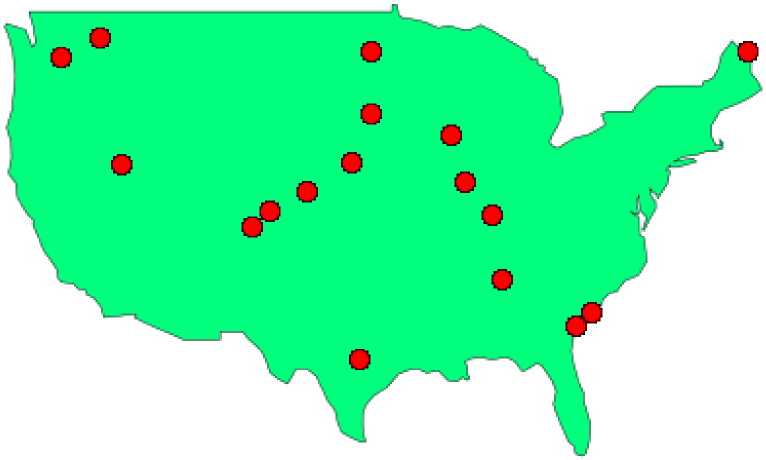
## 3. Problem Statement and Goal

The primary goal of this project is to create and train a machine learning model on an assortment of historical climate and agricultural yield data, so that it can then predict production of a variety of agricultural products based on observed key climate indicators. The tool is intended to be applicable regionally, and provide insight into how the agricultural industry may adapt their practices in order to optimize their efforts and mitigate damage in the face of climate change.

**Roles**
**Luca Bombelli -** Project Lead and Machine Learning Designer, responsible for creating and training machine learning algorithms to predict agricultural yields and testing the accuracy of each of these models.
**Kai Bengston -** Data Scientist and Climate Analyst, responsible for data collection and cleaning, showing correlation between climate factors and agricultural production to identify the existence of a relationship, and analyzing results.

## 4. Data Collection

The first step in the project was data collection. We first went to the National Oceanic and Atmospheric Administration (NOAA) for data on climate indicators from around the country.[1] We picked 11 key[2] climate indicators from 17 weather stations around the United States, measuring from 1990-2022, to train our model. This was to form an aggregate approximation of weather patterns throughout the United States; while this approximation is not accurate as the region being observed is too vast, it was enough climate data for us to train and test our model. We then took crop yield data from the Food and Agriculture Organization[3] (FAO), measuring the total yield of production in the United States of 9 key agricultural products[4] over the same timespan, to give an approximation of food production throughout the country as a whole. We also analyze the relationship between weather patterns and total yield of all measured food products.



Map of Weather Station Locations Used in Project

### Data Remark

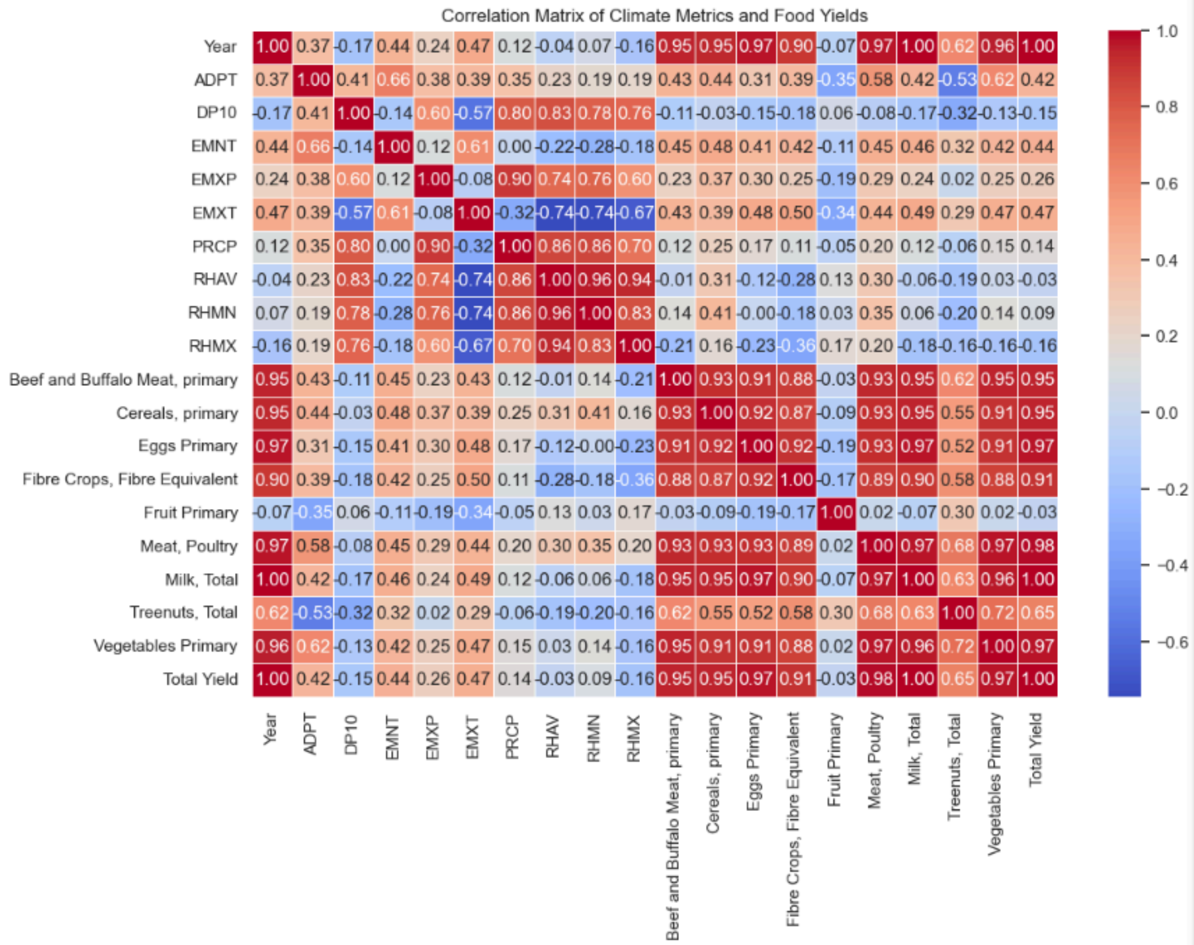
Our data is not a rigorous or accurate approximation of weather patterns or agricultural production for a particular region. Instead we amassed rough approximations of these phenomena, for solely the purpose of training and testing our machine learning model. The purpose of this project is not to draw final conclusions about relationships between the climate and agriculture, but to create an accurate predictive model. With access to more refined data and

[1] NOAA Data Search Portal (https://www.ncei.noaa.gov/access/search/data-search/global-summary-of-the-month?pageNum=1)
[2] Monthly Avg. Dew Point Temp. (ADPT), Num. of Days with greater than .10 inch of Precipitation (DP10), Extreme Minimum Temperature (EMNT), Extreme Maximum Precipitation (EMXP), Extreme Maximum Temperature (EMXT), Total Evaporation (EVAP), Monthly Mean Max. Evaporation Pan Water Temperature (MXPN), Precipitation (PRCP), Monthly Average Relative Humidity (RHAV), Month Average Min. Humidity (RHMN), Monthly Average Max. Humidity (RHMX)
[3] FAOSTAT, FAO Data Search Portal (https://www.fao.org/faostat/en/#data/FBS)
[4] Beef and Buffalo Meat, Cereal Grass, Eggs, Fibre Crops, Fruity Primary, Poultry Meat, Milk, Tree Nuts, and Vegetables

some parameter modifications, our algorithm could be applied on a far smaller and more regional scale for conclusive and informative results.

## 5. Correlation Analysis

We ran a simple correlation analysis between each of our climate factors and each of our agricultural yields.  The goal was twofold; first to support the existence of strong relationships between climate factors and agricultural yields over time, and second to highlight especially strong effects that individual factors had.  Our hope was the metrics with strong correlations would also prove to be important metrics in training our predictive model.  Our correlation matrix is depicted below.  We found that the average daily precipitation, extreme maximum temperature, and extreme minimum temperature had a particularly strong effect on much of the agricultural yields



| | Year | ADPT | DP10 | EMNT | EMXP | EMXT | PRCP | RHAV | RHMN | RHMX | Beef and Buffalo Meat, primary | Cereals, primary | Eggs Primary | Fibre Crops, Fibre Equivalent | Fruit Primary | Meat, Poultry | Milk, Total | Treenuts, Total | Vegetables Primary | Total Yield |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | 1.00 | 0.37 | -0.17 | 0.44 | 0.24 | 0.47 | 0.12 | -0.04 | 0.07 | -0.16 | 0.95 | 0.95 | 0.97 | 0.90 | -0.07 | 0.97 | 1.00 | 0.62 | 0.96 | 1.00 |
| ADPT | 0.37 | 1.00 | 0.41 | 0.66 | 0.38 | 0.39 | 0.35 | 0.23 | 0.19 | 0.19 | 0.43 | 0.44 | 0.31 | 0.39 | -0.35 | 0.58 | 0.42 | -0.53 | 0.62 | 0.42 |
| DP10 | -0.17 | 0.41 | 1.00 | -0.14 | 0.60 | -0.57 | 0.80 | 0.83 | 0.78 | 0.76 | -0.11 | -0.03 | -0.15 | -0.18 | 0.06 | -0.08 | -0.17 | -0.32 | -0.13 | -0.15 |
| EMNT | 0.44 | 0.66 | -0.14 | 1.00 | 0.12 | 0.61 | 0.00 | -0.22 | -0.28 | -0.18 | 0.45 | 0.48 | 0.41 | 0.42 | -0.11 | 0.45 | 0.46 | 0.32 | 0.42 | 0.44 |
| EMXP | 0.24 | 0.38 | 0.60 | 0.12 | 1.00 | -0.08 | 0.90 | 0.74 | 0.76 | 0.60 | 0.23 | 0.37 | 0.30 | 0.25 | -0.19 | 0.29 | 0.24 | 0.02 | 0.25 | 0.26 |
| EMXT | 0.47 | 0.39 | -0.57 | 0.61 | -0.08 | 1.00 | -0.32 | -0.74 | -0.74 | -0.67 | 0.43 | 0.39 | 0.48 | 0.50 | -0.34 | 0.44 | 0.49 | 0.29 | 0.47 | 0.47 |
| PRCP | 0.12 | 0.35 | 0.80 | 0.00 | 0.90 | -0.32 | 1.00 | 0.86 | 0.86 | 0.70 | 0.12 | 0.25 | 0.17 | 0.11 | -0.05 | 0.20 | 0.12 | -0.06 | 0.15 | 0.14 |
| RHAV | -0.04 | 0.23 | 0.83 | -0.22 | 0.74 | -0.74 | 0.86 | 1.00 | 0.96 | 0.94 | -0.01 | 0.31 | -0.12 | -0.28 | 0.13 | 0.30 | -0.06 | -0.19 | 0.03 | -0.03 |
| RHMN | 0.07 | 0.19 | 0.78 | -0.28 | 0.76 | -0.74 | 0.86 | 0.96 | 1.00 | 0.83 | 0.14 | 0.41 | -0.00 | -0.18 | 0.03 | 0.35 | 0.06 | -0.20 | 0.14 | 0.09 |
| RHMX | -0.16 | 0.19 | 0.76 | -0.18 | 0.60 | -0.67 | 0.70 | 0.94 | 0.83 | 1.00 | -0.21 | 0.16 | -0.23 | -0.36 | 0.17 | 0.20 | -0.18 | -0.16 | -0.16 | -0.16 |
| Beef and Buffalo Meat, primary | 0.95 | 0.43 | -0.11 | 0.45 | 0.23 | 0.43 | 0.12 | -0.01 | 0.14 | -0.21 | 1.00 | 0.93 | 0.91 | 0.88 | -0.03 | 0.93 | 0.95 | 0.62 | 0.95 | 0.95 |
| Cereals, primary | 0.95 | 0.44 | -0.03 | 0.48 | 0.37 | 0.39 | 0.25 | 0.31 | 0.41 | 0.16 | 0.93 | 1.00 | 0.92 | 0.87 | -0.09 | 0.93 | 0.95 | 0.55 | 0.91 | 0.95 |
| Eggs Primary | 0.97 | 0.31 | -0.15 | 0.41 | 0.30 | 0.48 | 0.17 | -0.12 | -0.00 | -0.23 | 0.91 | 0.92 | 1.00 | 0.92 | -0.19 | 0.93 | 0.97 | 0.52 | 0.91 | 0.97 |
| Fibre Crops, Fibre Equivalent | 0.90 | 0.39 | -0.18 | 0.42 | 0.25 | 0.50 | 0.11 | -0.28 | -0.18 | -0.36 | 0.88 | 0.87 | 0.92 | 1.00 | -0.17 | 0.89 | 0.90 | 0.58 | 0.88 | 0.91 |
| Fruit Primary | -0.07 | -0.35 | 0.06 | -0.11 | -0.19 | -0.34 | -0.05 | 0.13 | 0.03 | 0.17 | -0.03 | -0.09 | -0.19 | -0.17 | 1.00 | 0.02 | -0.07 | 0.30 | 0.02 | -0.03 |
| Meat, Poultry | 0.97 | 0.58 | -0.08 | 0.45 | 0.29 | 0.44 | 0.20 | 0.30 | 0.35 | 0.20 | 0.93 | 0.93 | 0.93 | 0.89 | 0.02 | 1.00 | 0.97 | 0.68 | 0.97 | 0.98 |
| Milk, Total | 1.00 | 0.42 | -0.17 | 0.46 | 0.24 | 0.49 | 0.12 | -0.06 | 0.06 | -0.18 | 0.95 | 0.95 | 0.97 | 0.90 | -0.07 | 0.97 | 1.00 | 0.63 | 0.96 | 1.00 |
| Treenuts, Total | 0.62 | -0.53 | -0.32 | 0.32 | 0.02 | 0.29 | -0.06 | -0.19 | -0.20 | -0.16 | 0.62 | 0.55 | 0.52 | 0.58 | 0.30 | 0.68 | 0.63 | 1.00 | 0.72 | 0.65 |
| Vegetables Primary | 0.96 | 0.62 | -0.13 | 0.42 | 0.25 | 0.47 | 0.15 | 0.03 | 0.14 | -0.16 | 0.95 | 0.91 | 0.91 | 0.88 | 0.02 | 0.97 | 0.96 | 0.72 | 1.00 | 0.97 |
| Total Yield | 1.00 | 0.42 | -0.15 | 0.44 | 0.26 | 0.47 | 0.14 | -0.03 | 0.09 | -0.16 | 0.95 | 0.95 | 0.97 | 0.91 | -0.03 | 0.98 | 1.00 | 0.65 | 0.97 | 1.00 |

Correlation matrix between Datasets

## 6. Machine Learning Model

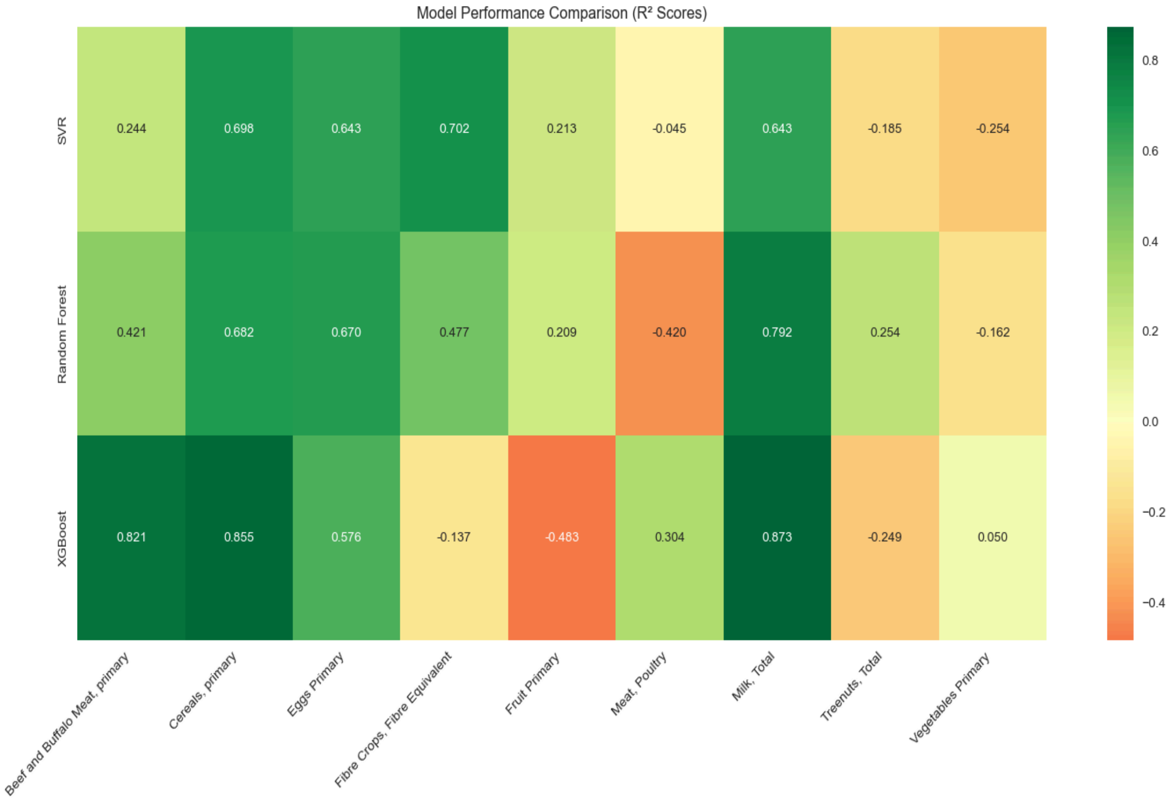
The machine learning implementation employs a structured class-based design through the ClimateModelTrainer, which streamlines data processing and model evaluation across multiple algorithms. This approach ensures consistent data handling and creates a flexible framework for future analysis. We applied 3 machine learning models to the data, to test accuracy and predictive nature, as well as weigh the importance of individual metrics. These models were a Support Vector Regression model (SVR), a Random Forest model, and an XGBoost model.

The models were trained on twelve climate-related features: average dew point temperature (ADPT), days with precipitation ≥ 1.0 inch (DP10), extreme minimum temperature (EMNT), extreme maximum precipitation (EMXP), extreme maximum temperature (EMXT), evaporation (EVAP), maximum daily precipitation (MXPN), precipitation (PRCP), average relative humidity (RHAV), minimum relative humidity (RHMN), maximum relative humidity (RHMX), and total yield. These features were used to predict nine target variables representing different food production categories: beef and buffalo meat, cereals, eggs, fiber crops, fruit, poultry meat, milk, tree nuts, and vegetables. Each target variable was measured in thousands of tonnes of production. The data was standardized using StandardScaler to ensure consistent scaling across all features, and a train-test split of 80-20 was implemented to evaluate model performance. The feature set encompasses a comprehensive range of climate metrics, capturing both extreme weather events and average conditions, while the target variables represent key agricultural outputs across different food sectors. This diverse set of features and targets allows for a thorough analysis of climate-food production relationships across different agricultural categories.

**Parameters**

Each model's hyperparameters were carefully optimized for performance. The SVR model employs an RBF kernel to capture non-linear relationships, with its complexity parameter C gradually increasing from 0.1 to 5.0 over 50 epochs. The Random Forest builds incrementally, adding 5 trees per epoch until reaching 250 trees, while maintaining unrestricted tree depth to capture complex patterns. XGBoost operates with a conservative learning rate of 0.1 and implements early stopping to prevent overfitting, running for a maximum of 50 epochs with continuous validation.To analyze the efficiencies of the three models, we ran performance analysis comparing the variance for each target using each model, and feature analysis on the RandomForest and XGBoost Models to determine the most important features for each model.

## Performance Analysis

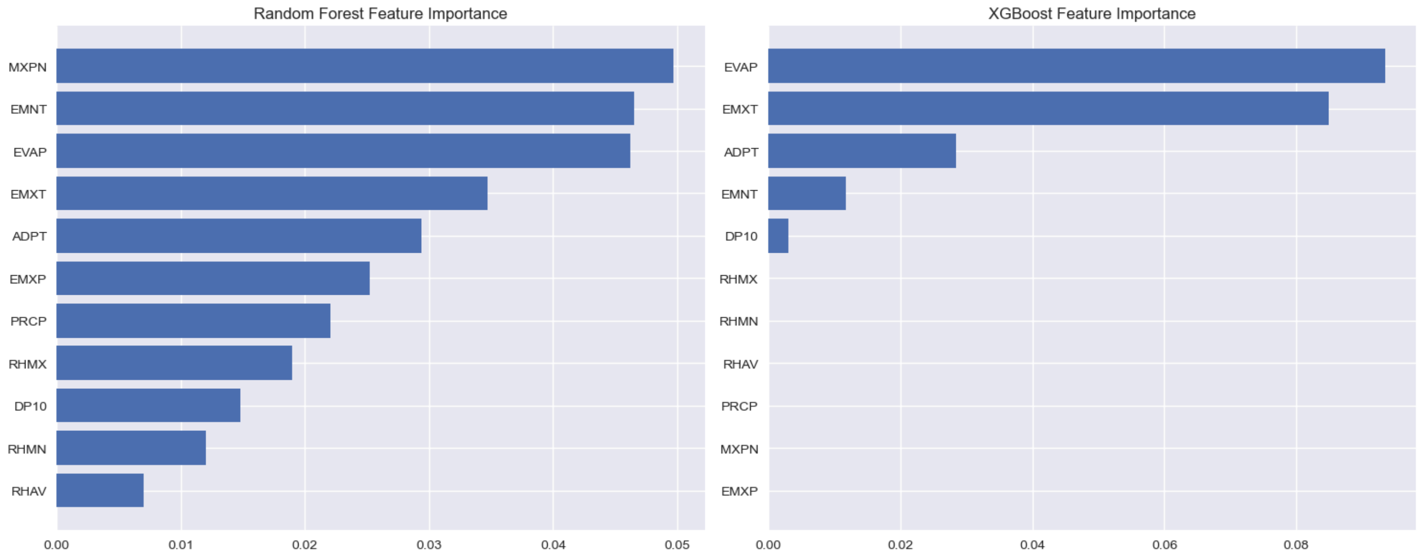


Performance Heatmap

The performance analysis reveals XGBoost as the superior performer, achieving exceptional results for milk production ($R^2 = 0.873$), cereals ($R^2 = 0.855$), and beef/buffalo meat production ($R^2 = 0.821$). Its training curves demonstrate rapid initial convergence, typically stabilizing within 20 epochs while maintaining sustained performance improvements. Random Forest delivers reliable performance with $R^2$ scores reaching 0.792 for milk production and 0.682 for cereals, showing particular strength in handling challenging categories with fewer negative $R^2$ scores. The SVR model, while showing lower peak performance, maintained consistency across several categories, notably in fiber crops ($R^2 = 0.702$) and cereals ($R^2 = 0.698$), though requiring longer training periods to reach optimal performance.

## Feature Analysis

Feature importance analysis reveals intriguing differences between the ensemble methods. Random Forest identifies maximum precipitation (MXPN) and extreme minimum temperature (EMNT) as primary predictors, while XGBoost emphasizes evaporation (EVAP) and extreme maximum temperature (EMXT). This divergence in feature prioritization suggests these models capture different aspects of the climate-production relationship, potentially offering complementary insights into agricultural prediction.

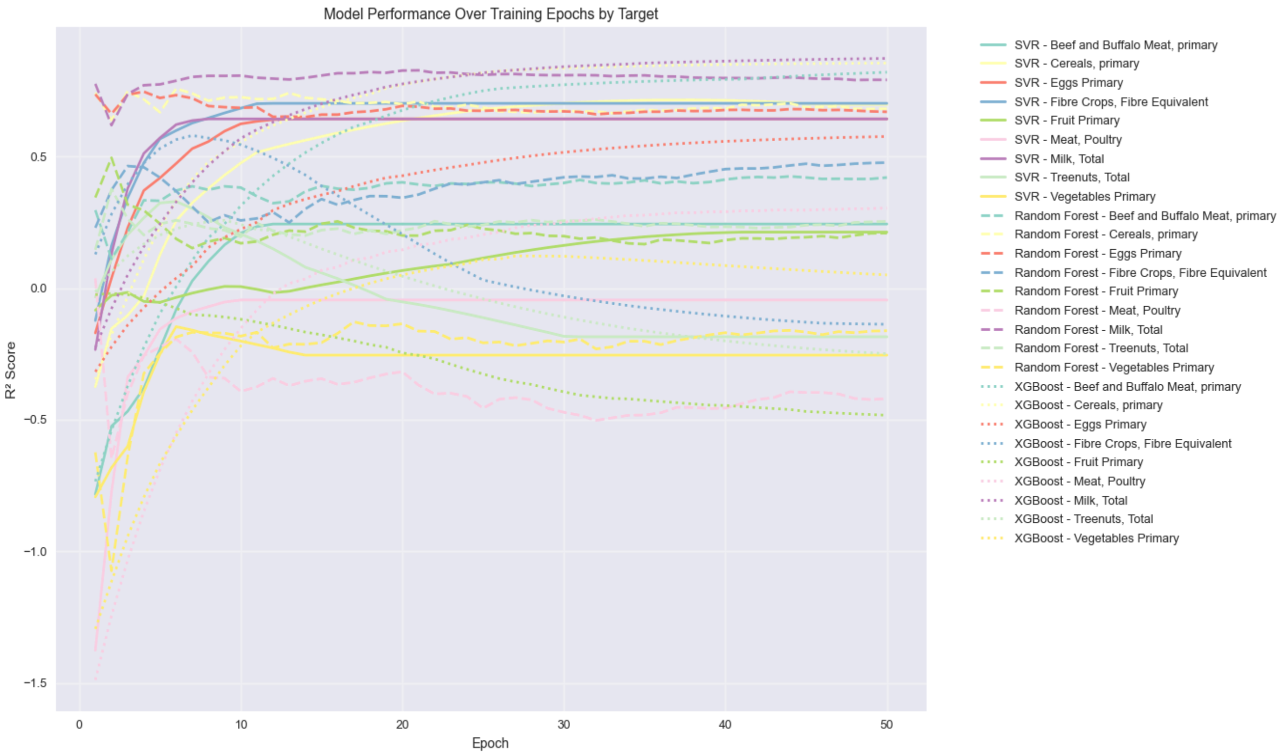


Display of the Importance of Individual Metrics while training model

Several food categories proved particularly challenging to predict. Vegetables, tree nuts, and fruit production showed poor predictability across all models, with negative R² scores indicating significant modeling challenges. The heatmap visualization clearly illustrates this performance disparity, showing strong performance clusters in dairy and cereal categories while highlighting consistent struggles with vegetables, tree nuts, and fruits. This suggests that factors beyond climate variables likely influence these production categories.

**Learning Patterns**

The training progression shows distinct learning patterns: XGBoost achieves rapid optimization, Random Forest maintains steady improvement with minimal volatility, and SVR demonstrates gradual but consistent progress. These characteristics suggest different optimal use cases for each model depending on the specific prediction requirements and data characteristics. While climate variables can effectively predict certain agricultural outputs, some food categories require additional predictive factors for accurate modeling. The analysis provides valuable insights for agricultural planning and climate impact assessment, while highlighting areas needing further investigation.



Visualization of Test Accuracy Over Epochs

## 7. Revised Model

The reduced feature model demonstrates interesting performance characteristics compared to the full feature implementation. Feature selection was guided by the feature importance analysis from the original model, which identified MXPN (maximum precipitation), EMNT (extreme minimum temperature), EVAP (evaporation), EMXT (extreme maximum temperature), and ADPT (average dew point temperature) as the most influential predictors.  The reduced model shows notable improvements in several key areas. XGBoost achieved exceptional performance with milk production ($R^2 = 0.9422$ compared to 0.873 in the full model) and maintained strong performance with cereals ($R^2 = 0.8065$ compared to 0.855). The model also showed marked improvement in egg prediction ($R^2 = 0.8192$ compared to 0.576) and beef/buffalo meat ($R^2 = 0.7539$ compared to 0.821). Random Forest maintained consistent performance with the reduced feature set, achieving an average $R^2$ of 0.3361 compared to 0.3247 in the full model. It performed particularly well with milk production ($R^2 = 0.8214$) and cereals ($R^2 = 0.7170$), though showing slight degradation from the full model in some categories. SVR demonstrated mixed results with the reduced features, achieving an average $R^2$ of 0.2243 compared to 0.2955 in the full model. However, it showed remarkable improvement in cereals prediction ($R^2 = 0.8891$ compared to 0.698) while maintaining similar performance in eggs production ($R^2 = 0.6771$ compared to 0.643).

The reduced feature set's comparable or improved performance suggests that the original model may have included redundant or noisy features. This is particularly evident in the prediction of milk production and cereals, which maintained strong performance across all three algorithms despite the reduced input dimensions. However, challenges persist in predicting vegetables, tree nuts, and fruits, indicating that these categories may require additional or different predictive factors beyond the selected climate variables. These results demonstrate that careful feature selection based on importance analysis can maintain or enhance model performance while reducing computational complexity. The success of the reduced model validates the feature importance analysis from the original implementation and suggests that focused feature selection may be preferable for practical applications in agricultural prediction.

## 8. Results

The machine learning models demonstrated varying levels of success across different food production categories, as evidenced by the performance metrics and visualizations. The training progression graph shows distinct learning patterns for each model and target combination, with most models achieving stability around epoch 20. XGBoost emerged as the strongest performer overall, showing exceptional results for several key categories. It achieved the highest $R^2$ scores for milk production (0.873), cereals (0.855), and beef/buffalo meat (0.821). The model's learning curves demonstrate rapid initial convergence followed by sustained performance improvements. Notably, XGBoost's feature importance analysis highlighted evaporation (EVAP) and extreme maximum temperature (EMXT) as the most significant predictors. Random Forest demonstrated reliable and consistent performance across categories, with particularly strong results in milk production ($R^2 = 0.792$) and cereals ($R^2 = 0.682$). The Random Forest feature importance analysis revealed different priorities compared to XGBoost, emphasizing maximum precipitation (MXPN) and extreme minimum temperature (EMNT) as key predictors. This model showed more stable learning curves with gradual improvement over training epochs. SVR showed moderate performance, with its strongest predictions in fiber crops ($R^2 = 0.702$) and cereals ($R^2 = 0.698$). The heatmap visualization reveals clear performance patterns across all models, with consistently strong predictions for dairy and cereal products, shown in dark green, contrasting with poorer performance in vegetables and treenuts, displayed in yellow and orange. All models struggled with certain categories, particularly vegetables, tree nuts, and fruits, often producing negative $R^2$ scores. This consistent pattern suggests that these agricultural products may be influenced by factors beyond the climate variables included in our analysis. The heatmap clearly illustrates this performance disparity, with distinct clusters of high-performing predictions in dairy and cereal categories contrasting with consistently poor performance in others. The feature importance analyses from both ensemble methods provide valuable insights into the climate-agriculture relationship, though they emphasize different aspects. This divergence in feature prioritization suggests complex interactions between climate variables and food production, with different models capturing distinct aspects of these relationships.

## 9. Limitations

The most prominent limitation that our project faced was access to data, for both training the model and having holistic approximations of crop production and weather patterns. Our project also does not take into account varying geographic factors such as soil, sunlight, or regional temperature. Future progress on this project would include reaching out to regional farmers and weather stations, and amassing very localized yields as well as weather data, to create predictions that are actionable and incorporate regional specifics of agriculture. Other improvements would be utilizing the CropNet data collection project that has an incredibly extensive archive of agricultural and weather data as well as satellite imaging.

## 10. Conclusion

Mitigating the damage done to the world's food supply is a critical step in the global fight against climate change. As weather behaviors become more intense, and historical weather patterns change rapidly, tools that allow the agricultural industry to adjust accordingly will be instrumental in preserving food stability. This project created a machine learning tool that could help farmers and members of the agricultural industry optimize food production, and in doing so showed that the relationship between crops and the climate is not only one that can be modeled, but one that can be predicted.