National College of Ireland

A Machine Learning Framework for Predicting Crop Production in Support of SDG13 Climate Action

MSc Research Project
Data Analytics

Amrit Laxmanasa Shidling
Student ID: X21198951@student.ncirl.ie

School of Computing
National College of Ireland

Supervisor: Dr. Paul Stynes

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Amrit Laxmanasa Shidling
Student ID:	X21198951@student.ncirl.ie
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr. Paul Stynes
Submission Due Date:	31/01/2023
Project Title:	A Machine Learning Framework for Predicting Crop Production in Support of SDG13 Climate Action
Word Count:	6419
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Amrit Laxmanasa Shidling
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☑
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☑
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☑

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Machine Learning Framework for Predicting Crop Production in Support of SDG13 Climate Action

Amrit Laxmanasa Shidling
X21198951@student.ncirl.ie

Abstract

Drought, a prolonged dry period in the natural climate cycle that will lead to reduced agricultural productivity. Current research utilizes machine learning techniques for drought classification, However, There is a need to develop precise crop production prediction models to address the impact of climate change. Addressing climate change to ensure food security is a challenge. This research proposes an advanced machine-learning framework to encourage better agriculture productivity. The proposed framework combines drought prediction factors with a crop production rate prediction model. A combination of soil moisture, drought information and crop production information consisting of Corn, Oats, Rice, Soybean and Wheat production data are used to train the machine learning regression Model. Data Preprocessing and Correlation Analysis techniques are applied to train four models namely, Linear Regression, XGBoost Regression, Random Forest Regression and Feed Forward Neural Networks. The results of the four models are presented in this research work based on Mean Squared Error(MSE), Mean Absolute Error(MAE) and R^2 (Coefficient of Determination) value. Cross-validation technique like K-folds is applied to assess the performance and generalizability of models. This research shows promise for Random Forest, Feed Forward Neural Network and XGBoost in encouraging farmers to decide on crop selection and agricultural efficiency.

Keywords - Crop Production, Regression, Random Forest, XGBoost, Cross Validation

1 Introduction

Agricultural production is affected directly by climate factors such as precipitation and temperature. These environmental factors control crop health and growth, annual crop production, and yield outcome of the cropping system over time Lehmann (2013) Liang et al. (2017). There is a need for predictive models to address the impacts on crop production due to climate change and worsening drought Mariotti et al. (2013) Schubert et al. (2007). Farmers engaged in anticipating the different crop production rates manually is a challenge. Remote sensing data such as the Normalised Difference Vegetation Index(NDVI) Li et al. (2019) and Satellite Imaginary Data can predict the crop production rate. These techniques have limitations in terms of saturation effect in the case of NDVI Huang et al. (2021) and the cost of collecting high-quality images in terms of Satellite Imaginary Data. Therefore a cost-effective approach is required. Cost and Precision are limitations with Remote Sensing techniques and therefore a machine learning model is required.

The aim of this research is to investigate to what extent a machine learning framework predicts crop production in support of SDG13 (Sustainability Development Goal) climate action.

To address the research question, the following specific sets of research objectives were derived:

1. Investigate the state-of-the-art broadly around Machine learning approaches to predicting and classifying drought conditions based on environmental factors.

2. Design the machine learning framework for crop production rate prediction to analyze patterns of different crops.

3. Implement the machine learning framework for crop production rate prediction.

4. Evaluate the machine learning framework for crop production rate based on MSE, MAE, R^2 and cross-validation.

The aim of this research is to investigate to what extent a machine learning framework Predicts Crop Production in Support of SDG13 Climate Action. The major contribution of this research is a novel advanced machine-learning framework that combines drought prediction factors with a crop production rate prediction model that encourage better agriculture productivity. Soil Moisture and Drought data cover drought temporal information from 2019 to 2022 of Wind Speed, Specific Humidity, Surface Pressure, Frost Point, Wet Bulb Temperature and Precipitation for 49 states of the United States. Crop Production data covers crop information on Rice, Wheat, Corn, Oats and Soybeans which are staples in the American diet. In order to identify the better model this research compares Linear Regression, Random Forest Regression, XGBoost Regression and Feed Forward Neural Network based on MSE, MAE and R^2 value and performs cross-validation for model assessment and validation.

This paper discusses previous related work on crop production rate prediction in 2^{nd} section related work. The research methodology is discussed in 3^{rd} section. 4^{th} section discusses the design components for the machine learning framework. The implementation of this research is explained in 5^{th} section. 6^{th} section of the report presents and discusses the evaluation results. 7^{th} section concludes the research and discusses future work.

2 Related Work

Sustainability Development Goal 13 is directly related to climate actions, addressing this climate change challenge in agriculture is important as mentioned by Zhang et al. (2019). The study by Khalid et al. (2020) indicated that the global need for food is projected to double by the year 2050. Pretty Rani (2023) emphasizes the need to study adaptation and approaches for climate change, including climate-smart agriculture. The research work conducted by Hwang et al. (2021) provided actionable climate policies, resonating with our adaptable machine learning framework approach. Whereas, Medar et al. (2019) and Kalimuthu et al. (2020) talk about the role of agriculture in India's well-being and economy. These studies talk about the importance of efficient crop planning, highlighting factors such as market prices, production rates, and government policies in crop selection. Some studies propose using machine learning for accurate farming decisions.

Given the diverse nature of climatic elements, various indices are employed to evaluate changes, with pressure indices and SST indices being utilized in this study done by Hunt et al. (2018). Drought forecasting methods encompass statistical, physical, and data-driven approaches. While physical-based models are sophisticated and require more resources, data-driven models, including machine learning (ML), offer a less complex and computationally efficient alternative Chaudhari et al. (2021). Nafii et al. (2022) used ML algorithms including ANN and SVM to forecast the annual Standardized Precipitation Evapotranspiration Index (SPEI) and implemented a correlation between the reservoir's annual inflow and the annual SPEI to embed it into hydrological drought forecasting. A research work done in the field of drought prediction by Jiang and Luo (2022), which made use of historical drought data combined with soil moisture data, implemented various machine learning and deep learning algorithms which showed that XGBoost performed best with an accuracy of 73.1% and among deep learning models GRU performed best with an accuracy of 73.3%. However, there is a need to incorporate techniques to assess on crop production.

Khaki and Wang (2019) implemented a Deep Neural Network (DNN) approach on the 2018 Syngenta crop challenge dataset about maize production. Which achieved the Root Mean Squared Error(RMSE) of 12%. RMSE of the average yield and 50% of the standard deviation (SD) for the validation dataset, outperforming methods like Lasso, SNN, and RT. Perfect weather data reduced the RMSE to 11% of the average yield and 46% of the standard deviation. But, work is limited to one crop type. Sellam and Poovammal (2016) analyzes environmental parameters such as Annual Rainfall (AR), Area under Cultivation (AUC), and Food Price Index (FPI) and establish relationships among them. Utilizing Regression Analysis (RA), a multivariate technique, the study examines factors' influence on crop yield over 10 years. Linear Regression (LR) reveals that crop yield is predominantly dependent on AR, with AUC and FPI also playing significant roles. The Linear Regression (LR) results showed that crop yield is primarily dependent on AR ($R^2 = 0.7242$), with AUC and FPI also influencing yield. Specifically, the relationship between AUC and crop yield demonstrated a high correlation ($R^2 = 0.7242$), indicating that as the cultivation area increases, the crop yield also increases.

Colombo-Mendoza et al. (2022) combined the IoT with Data Mining for the prediction of crop production rate especially making use of K-Nearest Neighbors (KNN) and LR. The results of these models are evaluated in terms of RMSE, with LR having an RMSE of 0.122 and KNN of 0.058. In the work by Bondre and Mahagaonkar (2019), SVM and Random Forest algorithms are introduced for predicting crop yield and recommending fertilizers, with SVM proving more effective. Similarly, Nigam et al. (2019) explores various machine learning techniques for crop yield prediction, highlighting their potential to revolutionize agriculture. In another work by Shook et al. (2021), Long Short Term Memory (LSTM) networks predict crop yield by integrating genotype with weather variables. performing better than traditional models, this study achieves better accuracy in predicting crop performance across various climatic conditions. The temporal attention mechanism enhances interpretability, offering valuable insights for farmers.

In conclusion, the state-of-the-art indicates that different models such as Random Forest, XGBoost Regression, and Deep Learning models are used in drought classification and prediction and crop production prediction. However, there need for experimentation to predict crop production rates in the face of drought conditions. Current research indicates that drought factors combined with historical crop information predict the crop production rate. The state of the art indicates that drought can be predicted using soil

moisture data. However, there is a need to expand the dataset features by incorporating historical crop production data. In addition, there is a need to map these datasets according to geolocation and time. The research proposes and machine-learning framework that combines drought prediction factors with crop production information to encourage the right agricultural decisions in terms of crop section. The advanced machine learning model will be identified from the better-performing model from Linear Regression, Random Forest, XGBoost and Neural Network through experimentation.

3 Methodology

The research methodology consists of six steps namely Data Gathering, Data Integration, Data Preprocessing, Crop-wise Data Segregation, Model Training and Model Evaluation as shown in Figure 1.

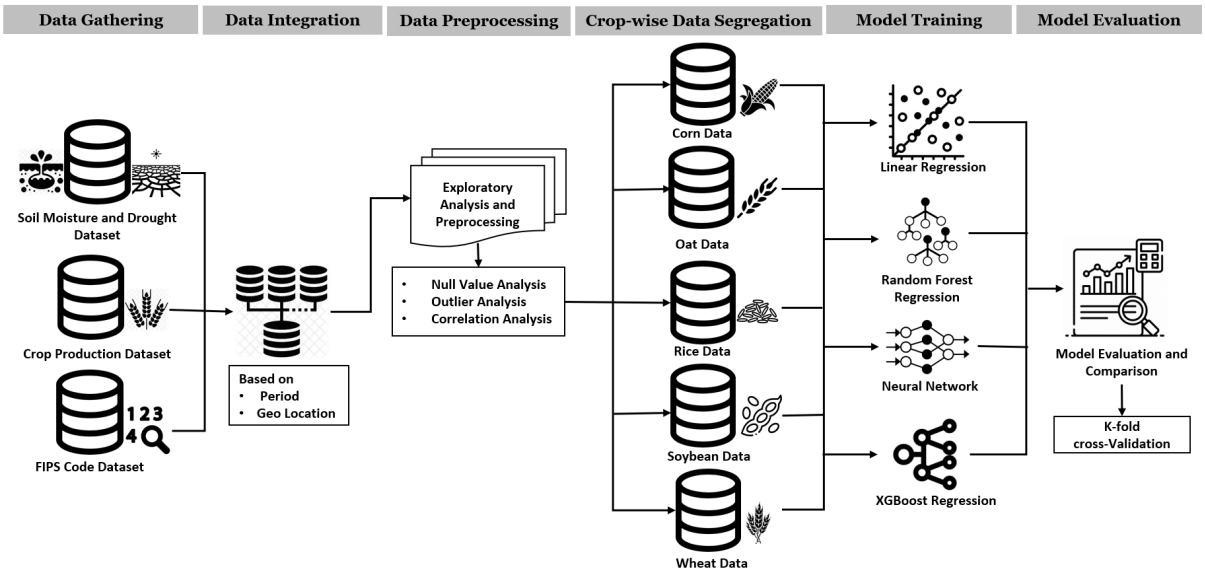


Figure 1: Research Methodology

3.1 Data Gathering:

The first step, *Data Gathering* involves collecting three different datasets soil moisture and drought dataset [1] and historical crop production dataset [2]. Another dataset for mapping state names and FIPS code [3] mapping dataset for data combining.

3.2 Data Integration:

The second step, *Data Integration* involves merging the soil moisture, and drought dataset with the historical crop production dataset. For the consistent data merging combining is done based on the state name and the date year column. As one dataset contains a state name column whereas the other contains FIPS code, Another dataset was used for

[1]https://www.kaggle.com/datasets/cdminix/us-drought-meteorological-data
[2]https://usda.library.cornell.edu/
[3]https://github.com/kjhealy/fips- codes/blob/master/state and county fips master.csv

mapping between FIPS Code and state names for smooth merging of these datasets. The combined dataset contains the fields as shown in Table 1.

Variable	Description	Units
WS10M_MIN	Min Speed of Wind (at 10 Meters)	m/s
QV2M	Specific Humidity (at 2 Meters)	g/kg
T2M_RANGE	Temperature Range (at 2 Meters)	°C
WS10M	Wind Speed (at 10 Meters)	m/s
T2M	Temperature (at 2 Meters)	°C
WS50M _MIN	Min Speed of Wind (at 50 Meters)	m/s
T2M_MAX	Max Temperature (at 2 Meters)	°C
WS50M	Wind Speed (at 50 Meters)	m/s
TS	Earth Skin Temperature	°C
WS50M_RANGE	Wind Speed Range (at 50 Meters)	m/s
WS50M_MAX	Max Speed of Wind (at 50 Meters)	m/s
WS10M_MAX	Max Speed of Wind (at 10 Meters)	m/s
WS10M_RANGE	Wind Speed Range (at 10 Meters)	m/s
PS	Surface Pressure	kPa
T2MDEW	Dew/Frost Point (at 2 Meters)	°C
T2M_MIN	Min Temperature (at 2 Meters)	°C
T2MWET	Wet Bulb Temperature (at 2 Meters)	°C
PRECTOT	Precipitation	mm day^{-1}
state_name	State Name	–
production	Crop Production	in 1,000 bushels

Table 1: Description of Meteorological and Crop Production Variables

3.3 Data Preprocessing:

The third step, *Data Preprocessing* involves mapping state names to FIPS codes, filtering data within a specified date range according to the data availability from all datasets, and organizing yield datasets for Corn, Oat, Rice, Soybean and Wheat. A correlation analysis is performed for identifying and handling strong and weak correlations among numeric columns in the datasets. The threshold for the strong correlation is 0.8 and the weak threshold is 0.2 based on the experiment. The correlation analysis includes a heatmap visualization and identifies columns to drop based on correlation thresholds.

3.4 Crop-wise Data Segregation:

The fourth step, *Crop-wise Data Segregation* involves the separation of preprocessed data crop-wise as Corn, Oat, Rice, Soybean and Wheat Data. The dataset contains the soil moisture, drought and crop production information for multiple states of the United States. The data segregation is performed for performing the model training for each crop type data.

3.5 Model Training:

The fifth step, *Model Training* involves the training of different models. The models were trained with a training dataset (80% of the dataset) and tested with the validation set (20% of the dataset). The machine learning models including Linear Regression from sci-kit-learn's LinearRegression Library, Random Forest Regressor, and Gradient Boosting Regressor from sci-kit-learn's Ensemble library, are employed for predictive analysis. The deep learning mode includes a TensorFlow-based Neural Network. The Neural Network model comprises three layers, an input layer with 128 neurons using the rectified linear unit (ReLU) activation function, a hidden layer with 64 neurons also employing ReLU activation and an output layer designed for regression tasks with a single neuron. The model is compiled using the Adam optimizer and the mean squared error loss function.

3.6 Model Evaluation:

The sixth step, the model evaluation process involves assessing the model's performance using metrics such as Mean Squared Error, Mean Absolute Error, and R^2-value. Additionally, K-fold cross-validation is employed to assess generalization performance, providing an evaluation across different data splits using the Root Mean Squared Error(RMSE) metric. Four models for each crop type are compared and visualised using Python. These evaluations provide the model's performance, interpretability, and resilience in predicting crop production with soil and drought data.

4 Design Specification

The machine learning framework architecture combines drought prediction factors with a crop production rate prediction model as shown in Figure 2. The drought prediction component of the machine learning framework includes gathering temporal soil moisture and drought information, classification model development and model evaluation as discussed in section 4.1. Components of the crop production elements are discussed in section 4.2.

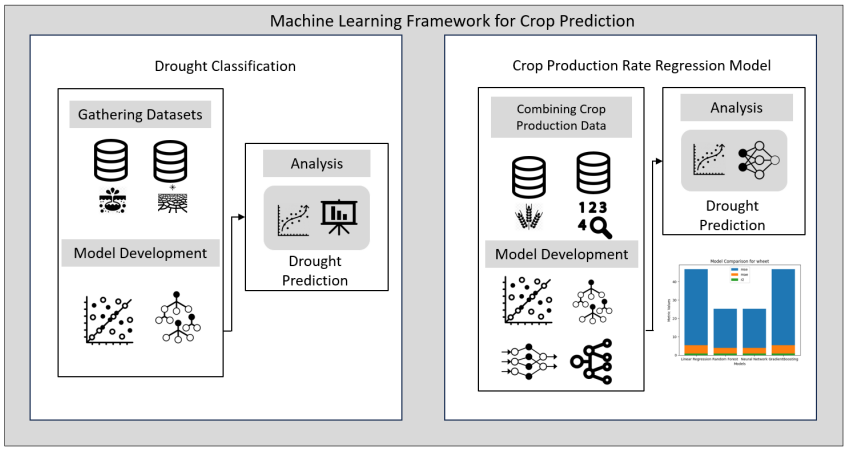


Figure 2: Machine Learning Framework Architecture

4.1 Drought Classification

The drought level classification model is where the combined temporal soil moisture data and the drought information are used to predict and classify the drought level. The drought classification model is based on Random Forests, Decision Trees and KNN. An analysis is carried out to evaluate the performance of models.

4.2 Crop production

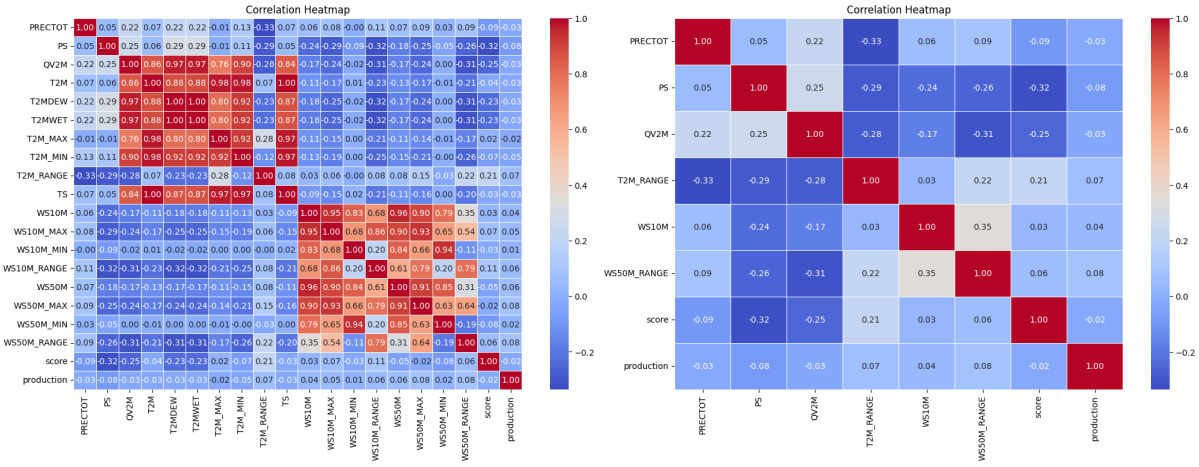
The Crop Production Rate Regression model includes combining historical crop data and FIPS code data, model development, performance analysis of the model and plotting the evaluation results.

Historical crop production data are collected based on the region and period of the drought information. Crop information covers the production data related to Corn, Oats, Rice, Soybean and Wheat. FIP code data is collected to map the state names with code for mapping and combining these data based on state code. The model development section involves training in Linear Regression, Random Forest Regression, Gradient Boost Regression and Neural Networks for segregated data of each crop type.

The analysis Section is where the performance of each model is evaluated based on MSE, MAE and R^2 value. The evaluation values are plotted for comparison to find the best-performing model. The cross-validation technique is applied to find the best reliability and generalisation capacity evaluation of the model.

5 Implementation

The machine learning framework was implemented as a regression model for six types of crops selected. In this section, The steps taken to obtain the research objectives by predicting the crop production data for different states of the United States. Python programming language was used to code all the models in a Jupyter Notebook as it is flexible and able to handle large datasets. Anaconda Python distribution is used in this research work, which is open source. This study used the United States drought dataset from 2019 to 2022, The research relies on crucial datasets: the Soil Moisture and Drought Dataset from NASA LaRC, containing key indicators like wind speed, humidity, temperature, and precipitation for understanding soil conditions; the Historical Crop Production Dataset from the U.S. Department of Agriculture, offering insights into crop yields and production trends over time; and the Country FIP Codes Dataset, sourced from a GitHub repository, for precise alignment of soil, crop, and geographical data. These datasets were imported using the "Pandas" library.



(a) Correlation in Original Data (b) After Handling Correlation

Figure 3: Correlation Matrix

In the implementation, 20% of rows in the dataset were randomly selected and perturbed by adding Gaussian noise to improve the generalisation capacity of the model. This was achieved using NumPy and pandas. The dataset is split into training and testing sets, one-hot encoding categorical variables. The StandardScaler from the Scikit-learn library is used for standardizing numerical features to avoid variable bias. The correlation analysis is conducted for the identification and management of multicollinearity issues in a dataset. Heatmap visualisation is plotted using the Seaborn library as shown in Figure 3. The thresholds for strong and weak correlations are set for handling excessive correlation (strong correlation) or negligible correlation (weak correlation) between features.

The machine learning model Linear Regression from sci-kit-learn's LinearRegression Library, Random Forest Regressor and Gradient Boosting Regressors from sci-kit-learn's Ensemble library are employed for predictive analysis. For neural network modelling TensorFlow Sequential model with two hidden layers (128 and 64 neurons, respectively) and a linear output layer is used. The model is compiled with the Adam optimizer and mean squared error loss. Training occurs over 50 epochs with a batch size of 32. Other machine learning models, such as RandomForest and GradientBoosting are obtained from the models sci-kit-learn and trained on the preprocessed data. In assessing the model's performance, metrics like MSE, MAE, and R^2 are for evaluation. The cross-validation process ensured the model's robustness through k-fold validation, measuring RMSE as a key metric. Leveraging sci-kit-learn's evaluation tools and feature importance analysis, understanding of model's accuracy, and generalization. 'Permutation Importance' is evaluated for analysing the impact of each variable on model performance by using Sci-kit-learn's 'Inspection' package.

6 Results and Discussion

The aim of this experiment is to perform multiple experiments and compare Linear Regression, XGBoost Regression, Random Forest Regression and Feed Forward Neural Network regression model with the combined crop datasets. Through the use of advanced machine learning techniques, four models were trained to predict the crop production

rate for Corn, Oats, Rice, Soybean and Wheat crops.

6.1 Experiment 1: Exploratory Data Analysis

The aim of this experiment is to analyse the pattern followed by soil and drought data like distribution of the drought level according to year, month and days. Correlation between temporal variables with respect to drought score and the variation in the crop production across states.

The correlation between Humidity at 2 meters and Temperature at 2 meters with Drought Score is illustrated in Figure 4. The dark colour points indicate a high drought score. It can be observed that there is a relation between these indicators, The drought score tends to be high when both humidity and temperature have a high value. There are a few instances where the drought score is high for very low temperatures and low humidity.

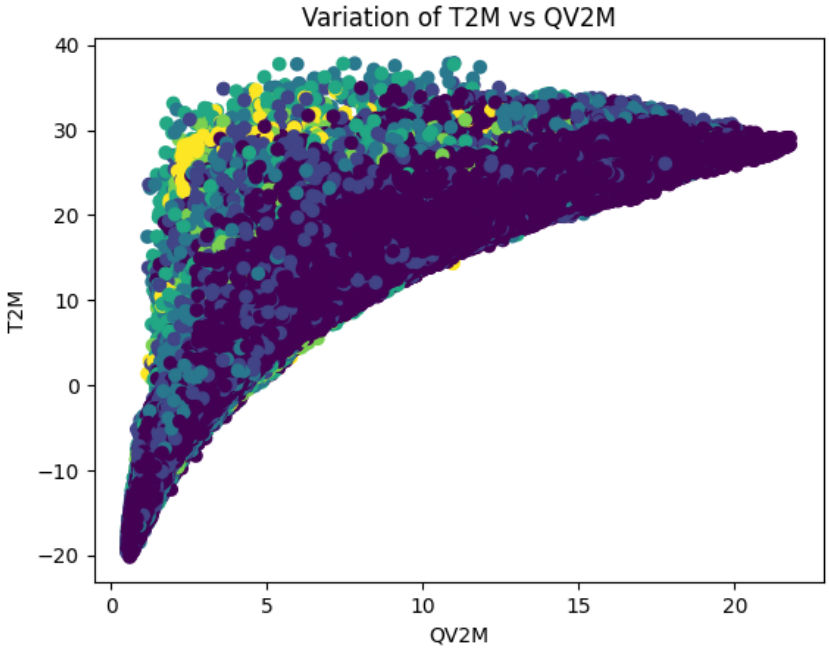


Figure 4: Correlation Between Specific Humidity (QV2M) and Temperature (T2M) with Drought Scores

Variation in crop production rate across different states of the US is analysed by box plot. Variation in Oat production is illustrated in Figure 5. The boxplot shows that there is high oat production in states like North Dakota, South Dakota, Oregon and so on.

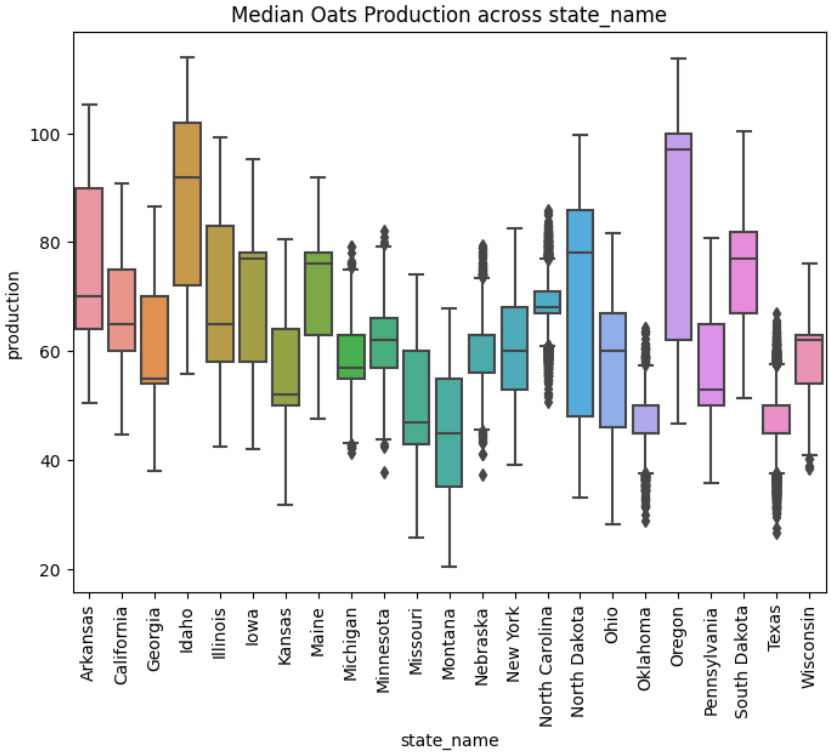


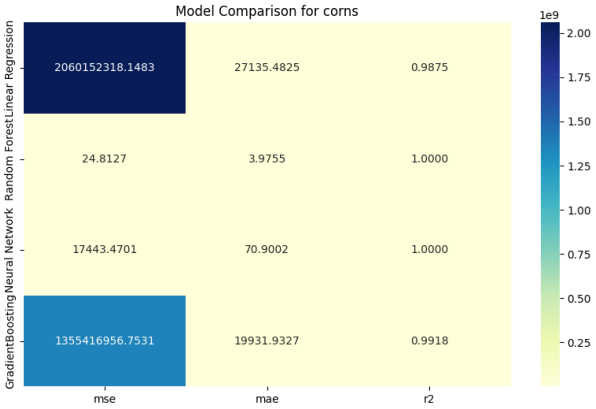
Figure 5: Variation in Oats Production Across States

6.2 Experiment 2: Replication of State of the Art

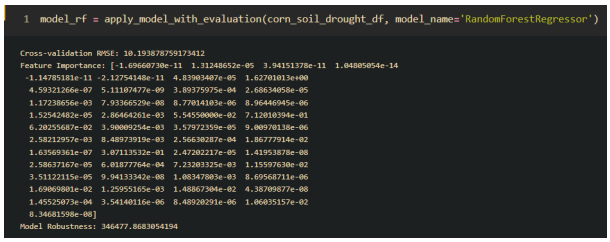
The aim of this experiment is to replicate the state-of-the-art Jiang and Luo (2022) that includes the preprocessing of the drought and soil moisture data and building various models for the purpose of drought prediction and classification. A Random Forest classifier with 20 estimators and a maximum depth of 70 was employed. The model achieved notable performance metrics on the test set, with an accuracy of 80.02%, precision of 78.22%, recall of 80.02%, F1 score of 78.61%, and a Cohen Kappa score of 57.19%. These results indicate that the Random Forest approach accurately classifies drought conditions. The second model, the K-Nearest Neighbors (KNN) classifier performed well, with an accuracy of 78.58%, precision of 76.89%, recall of 78.58%, F1 score of 77.46%, and a Cohen Kappa score of 55.19%. These results showed that the Random Forest Classifier performed better than KNN.

6.3 Experiment 3: Crop Production Rate Prediction for Corn

The aim of this experiment is to find a best-performing model for Corn production rate prediction. To perform this experiment a change in dataset is performed by including data related to Corn. The evaluation results for corn production prediction showed better results for the Random Forest (RF) model as shown in Figure 6. With an impressively low MSE of 24.81, MAE of 3.97(Indicating precise prediction) and R^2 value of 1.0 indicating a perfect fit to the data.



(a) Comparison of model performance



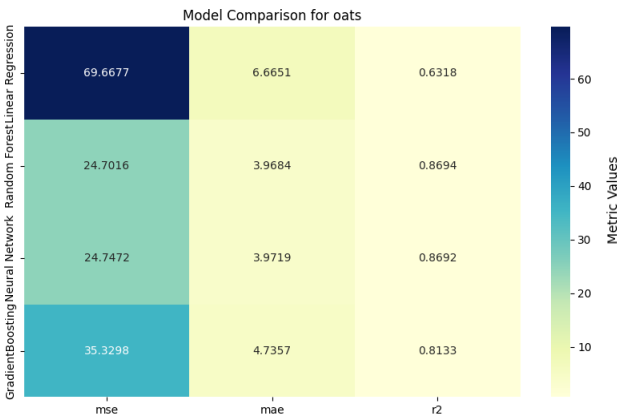
(b) K-Fold Cross Validation Result

Figure 6: Model Evaluations for Corn Production Rate Prediction

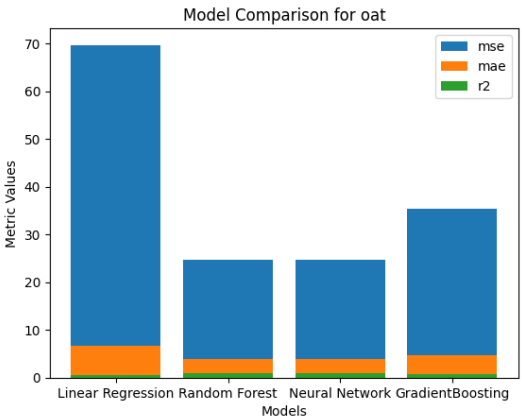
The corn production prediction model, assessed through cross-validation showed an RMSE of 10.19, indicating overall precision as shown in the Figure 6b. Feature importance analysis showed the influence of specific features such as precipitation and temperature.

6.4 Experiment 4: Crop Production Rate Prediction for Oat

The aim of this experiment is to find the best-performing model for Oat production rate prediction. To perform this experiment a change in dataset is performed by including data related to Oat. The oats production prediction models are evaluated with mean squared error (MSE), mean absolute error (MAE), and R^2 metrics as shown in Figure 7. The linear regression model (LR) achieved an MSE of 69.67, MAE of 6.67, and R^2 of 0.63. The gradient boosting model (GB) improved the performance with an MSE of 35.33, MAE of 4.74, and R^2 of 0.81. The random forest model (RF) demonstrated even better results with an MSE of 24.70, MAE of 3.97, and R^2 of 0.87. Similarly, the neural network model (NN) closely matched the RF model with an MSE of 24.75, MAE of 3.97, and R^2 of 0.87.



(a) Comparison of model performance



(b) Comparison of model performance

Figure 7: Comparison of model performance for Oat Production

The model demonstrates robustness, as evidenced by a Cross-validation Root Mean Squared Error (RMSE) with perfect 0.0. The overall assessment indicates the Random Forest model performs best among the models .

6.5 Experiment 5: Crop Production Rate Prediction for Rice

The aim of this experiment is to find a best-performing model for Rice production rate prediction. To perform this experiment a change in dataset is performed by including data related to Rice. The evaluation results for the rice prediction model are summarized for four different algorithms namely, Linear Regression, Random Forest, Neural Network, and Gradient Boosting in Figure 8. The Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2-value metrics are provided for each algorithm. For Linear Regression, the MSE is 135,816.34, MAE is 291.31, and R^2 is 0.34. Random Forest and Neural Network algorithms also exhibit strong results, with MSE values of 25.22 and 26.17, MAE values of 4.01 and 4.08, and R^2 values of 0.9999 and 0.9999 and The Gradient Boosting algorithm shows superior performance with an MSE of 59.80, an MAE of 6.21, and R^2 of 0.9997, respectively. Random Forest is best in terms of accuracy and overall performance.

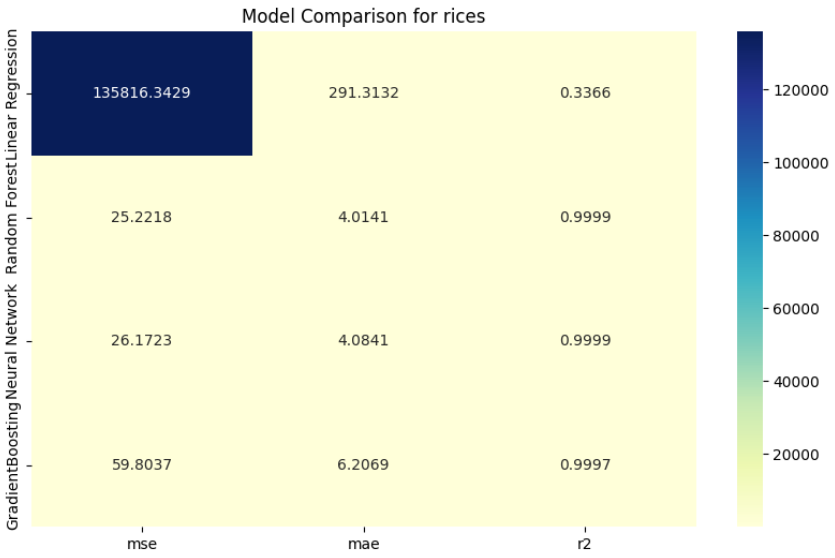


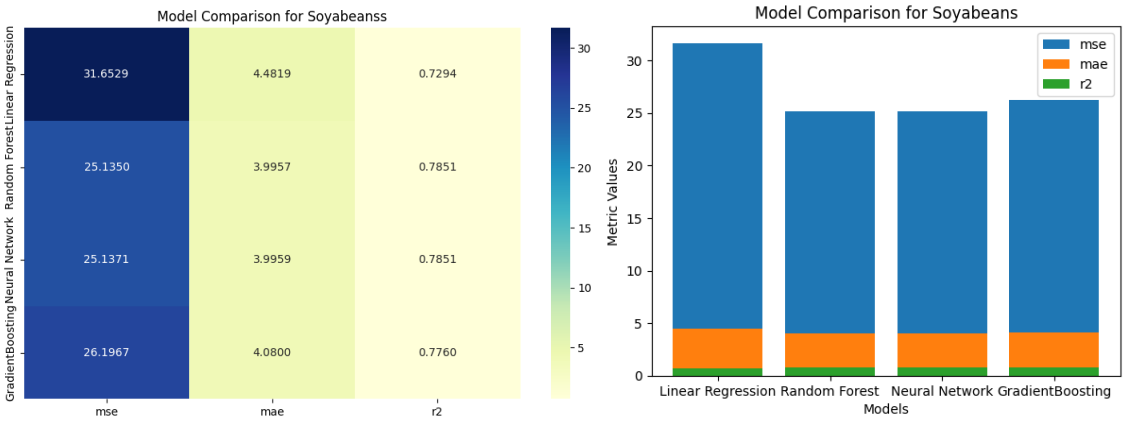
Figure 8: Model Evaluations for Rice Production Prediction

The Random Forest model, applied to a dataset on rice production with soil and drought features, exhibited high accuracy and consistency in cross-validation, reflected by a near-zero RMSE. Key impactful features were identified, and a robustness test demonstrated the model's resilience with an RMSE of 164.54, Indicating the Random Forest model's strong and consistent performance in accurately predicting rice production.

6.6 Experiment 6: Crop Production Rate Prediction for Soybean

The soybean production prediction models were evaluated using different algorithms as shown in Figure 9. The Linear Regression model achieved a Mean Squared Error (MSE) of 31.65, with a Mean Absolute Error (MAE) of 4.48 and an R^2 value of 0.73. The Gradient Boosting model resulted in an MSE of 26.20, MAE of 4.08, and R^2 of 0.78.

The Random Forest model exhibited an MSE of 25.13, MAE of 3.99, and R^2 of 0.79. The Neural Network model showed similar performance, with an MSE of 25.14, MAE of 3.99, and R^2 of 0.79. These findings indicate that the Random Forest and Neural Network models perform consistently well in predicting soybean production based on soil and drought data.



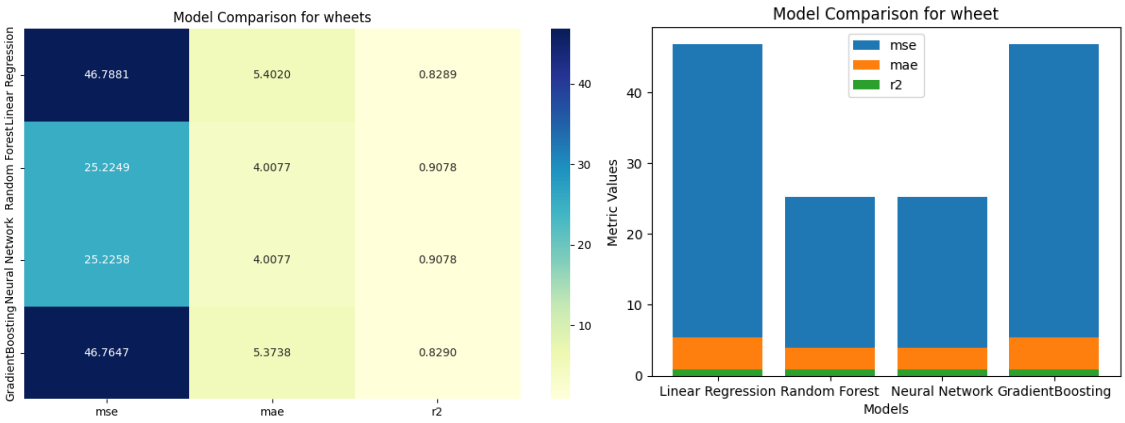
(a) Comparison of model performance (b) Comparison of model performance

Figure 9: Comparison of model performance for Soybean Production

The soybean production prediction model, specifically the Random Forest algorithm, showed good performance with a Mean Squared Error (MSE) of 25.13, a Mean Absolute Error (MAE) of 3.99, and a R^2 value of 0.79. Cross-validation results indicated a perfect fit, yielding an RMSE of 0.0. The model demonstrated robustness with minimal sensitivity to perturbations(noise), as evidenced by a low perturbed RMSE of 10.19. Overall, these results highlight the Random Forest model's accuracy, interpretability, and stability in predicting soybean production based on soil and drought data.

6.7 Experiment 7: Crop Production Rate Prediction for Wheat

The aim of this experiment is to find a best-performing model for Wheat production rate prediction. To perform this experiment a change in dataset is performed by including data related to Wheat. The wheat production prediction models were evaluated as shown in Figure 10. Linear Regression and Gradient Boosting showed moderate accuracy with MSE of 46.79 and MAE of 5.40 and 5.37. In contrast, Random Forest and Neural Network performed better with lower MSE and MAE of 25.22, 4.01 and 25.23, 4.01 respectively. and a high R^2 value of 0.91. Overall, Random Forest and Neural Network showed good results for Wheat production rate prediction.



(a) Comparison of model performance (b) Comparison of model performance

Figure 10: Comparison of model performance for Wheat Production

In K-Fold Cross Validation, the Random Forest model showed good accuracy with a low RMSE of 1.59e-13 in. Key features, including state-wise data, soil characteristics, and drought scores, significantly influence predictions. The model's stability is confirmed through a robustness test, revealing a minimal RMSE of 6.94.

7 Conclusion and Future Work

The aim of this research was to encourage farmers to make decisions on crop selection and enhance agricultural efficiency. This research proposes a machine learning frame that combines the drought prediction aspect with crop production information. Results demonstrate that the Random Forest Regressor performs best, showing the least MSE, MAE and high R^2 value. Further cross-validation such as K-fold demonstrated that Random Forest is a good choice in terms of RMSE. The other models like Neural Networks and Gradient Boost performed better than Linear Regression in terms of MSE, MAE and R^2. Feature important assessment is performed to find key contributors for each crop type. A limitation of this study is the insufficient historical crop production data, to capture long-term trends and patterns.

This research can improve agriculture practices, especially in countries with sustainability challenges. Optimising the model can improve the work. Furthermore, extensive research can be carried out on this work by developing a time series-based model for data drought and crop datasets of different regions. time series model developed on data of extended period enable for better accurate prediction. Another plan will develop visual images along with the location maps that will provide effective prediction of crop production rate. This will provide farmers with a chance to improve agricultural decisions.

8 Acknowledgement

I heartfully want to thank my research supervisor, Dr. Paul Stynes for constantly motivating and supporting me. Continuous assistance and guidance from my supervisor helped me a lot in the research work. It is my pleasure to express my sincere gratitude to my family for their support, understanding, and love. Lastly, I would like to thank the college for providing the best knowledge in Data Analytics.

9 Question and Answers

1. **What is the motivation for selecting 0.8 and 0.2 as the threshold for the strong correlation and the weak threshold in methodology**

   The reason for implementing the correlation matrix is for the purpose of selecting the important features. As the variables having a strong positive correlation may lead to multi-collinearity in the regression model. Multicolinearity will lead to a higher standard error. On the other hand variables with strong negative correlation will contribute little information or noise to the model. This is generally performed for the make model simple and hence leads to the generalisation.

   - **Strong Correlation (0.8):**
     The reason I chose the 0.8 for strong correlation is based on the observation and experiment. As there are some variables like Max Wind Speed at 10 Meters (WS10M) and Max Wind Speed at 50 Meters (WS50M), Both indicate the wind speed and the values of both don't vary much. They form a strong correlation. I wanted to avoid such repetitive features, and from my observation, they mostly lie in the range of 0.85 to 1.0.

   - **Weak Correlation (0.2):**
     I considered observations and experiments for weak correlation thresholds. For instance, variables like wind speed at 10 meters (WS10M) and pressure(PS) often show weak correlations. Setting the threshold at 0.2 helps avoid including features with such weak correlations, which will make the model less generic. Generally, feature correlations in the range of 0.0 to 0.2 were observed in the context of the considered dataset.

2. **In figure 4 what is the relationship between Specific Humidity (QV2M) and Temperature (T2M).**

   Figure 11 illustrates the relationship between specific humidity ($QV2M$) and temperature at 2 meters above the ground ($T2M$). Each point in the scatter plot corresponds to a specific observation, with the x-axis representing specific humidity values and the y-axis representing temperature values. The colour of each point is determined by the associated 'drought score' values.

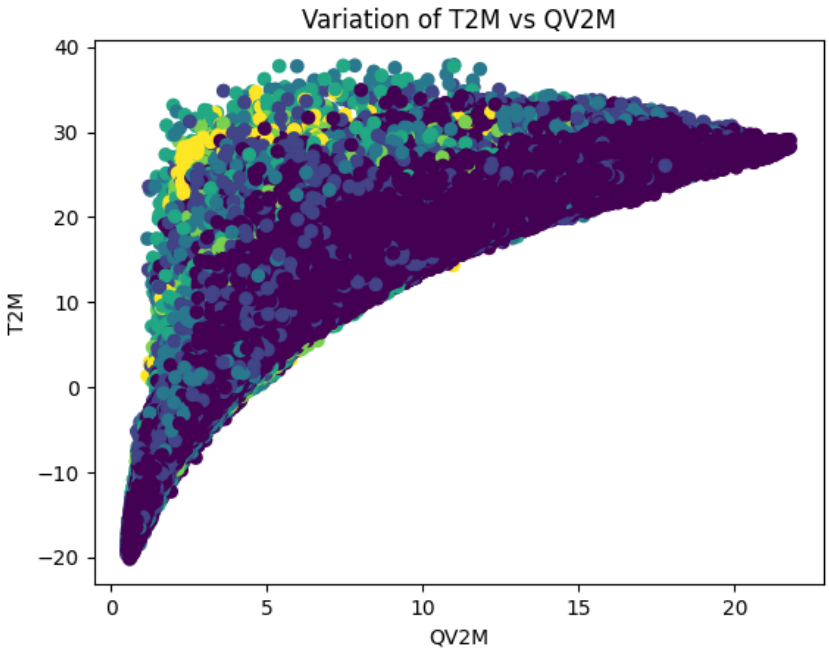


Figure 11: Correlation Between Specific Humidity (QV2M) and Temperature (T2M) with Drought Scores

The scatter plot provides a visual representation of drought conditions concerning changes in temperature and specific humidity, and the colour-coded scores indicate potential patterns within the dataset.

From the plot, it can be said that there is a high concentration of high drought (dark blue points) at high temperatures and high humidity especially temperatures above and around 15 °C and humidity above 20 g/kg, On the other hand, drought intensity is low when there is a moderate temperature and humidity which is indicated by yellow and light blue dots. Also, there are instances where there is moderate humidity of around 3 to 5 and temperatures in the range of 20 °C to 35 °C and shows less drought conditions as indicated by yellow and light blue points.

3. In experiment 2 can you discuss how these results relate to the state of the art?

To answer this question, I would start with a short background. The state-of-the-art domain primarily focuses on predicting droughts using weather and soil data. There are 6 different drought levels defined by U.S.Drought Monitor[4] based on the intensity of drought, namely D0, D1, D2, D3, D4 and No Drought.

Experiment 2 is nothing but the reproducing the state-of-the-art work. The results showed that the Random Forest model achieved good performance metrics on the test set, with an accuracy of 80.02%, precision of 78.22%, recall of 80.02% and F1 score of 78.61%. The second model, the K-Nearest Neighbors (KNN) classifier model showed an accuracy of 78.58%, precision of 76.89%, recall of 78.58%, and F1 score of 77.46%. These results showed that the Random Forest Classifier performed better than KNN.

[4]https://droughtmonitor.unl.edu/About/AbouttheData/DroughtClassification.aspx

These results in experiment 2 are related to the state-of-the-art in a way that, given the values of the soil moisture data and environmental data like Min Speed of Wind (at 10 Meters), Specific Humidity (at 2 Meters), Temperature Range (at 2 Meters), Wind Speed (at 10 Meters), Temperature (at 2 Meters), Max Temperature (at 2 Meters), Wind Speed (at 50 Meters), Earths Skin Temperature, Wind Speed Range (at 50 Meters), Max Speed of Wind (at 50 Meters), Max Speed of Wind (at 10 Meters), Wind Speed Range (at 10 Meters), Surface Pressure, Dew/FrostsPoint (at 2 Meters), Min Temperatures (at 2 Meters), Wet Bulb Temperature (at 2 Meters) and Precipitations, How effectively is the model able to predict the drought level?

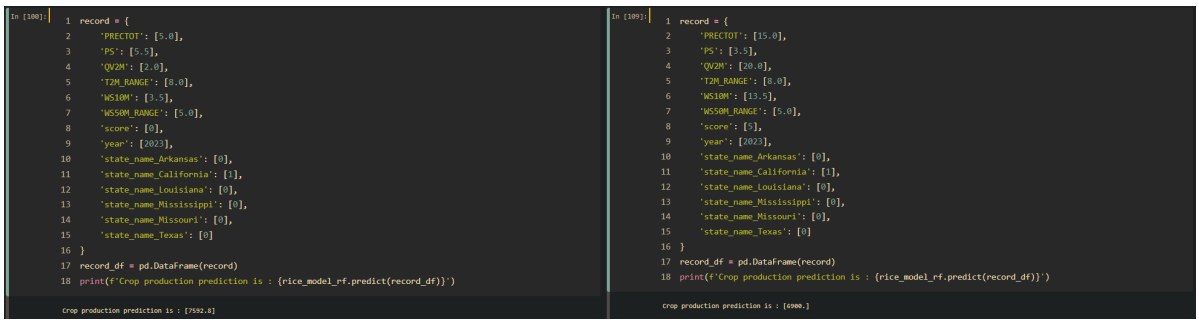
In comparing the experiment's results to state-of-the-art work, It is observed that I achieved almost similar accuracy and precision, the main reason for the small variation is because of the consideration of the limited dataset (around 1 million records) due to the space constraint (which I discussed during the project phase).

4. Discuss the results in the context of answering the research question and comparison to state of the art.

The aim of my research is to investigate to what extent a machine learning framework predicts crop production in support of SDG13 (Sustainability Development Goal) climate action.

To answer the above question based on the research question, I wrote code to predict the production of the crop under different environmental conditions as shown in figure 12.

The first scenario, characterized by moderate precipitation, average surface pressure, and other balanced factors has normal environmental conditions like Low humidity of 2 g/kg, Mild temperature range of 8 °C, Moderate wind speed at 10m of 5 m/s, score of 0.0 which indicates no drought condition the model predicted a crop production of 7592.8 (measure in 1000 bushels) as shown in figure 12a. On the other hand, the second scenario shown in figure 12b, featuring high precipitation of 25.0 mm per day, surface pressure of 25.0 kPa, humidity of 8.0 g/kg, temperature range 30.0 °C, High wind speed at 10m of 23.5 m/s, at drought condition (indicated by score 5.0) predicted a low crop production of 6900.0 (measure in 1000 bushels).



(a) Rice Production at Normal Condition (b) Rice Production at Abnormal Condition

Figure 12: Rice Production Prediction at Different Conditions

By analysing the results of project work, these predictions offer valuable insights for precision agriculture. Farmers can tailor cultivation practices based on specific

environmental conditions in each state, leading to optimized resource use and risk mitigation. Additionally, the MSE and MAE of the Random Forest model used for the above prediction are very low (25.09 and 3 respectively), which indicates that predictions are not much different than actual values while testing on test data.

To, Address the second part of the question which is, How this is related to the state-of-the-art. The state-of-the-art work is "drought prediction" whereas my research is an extension of that work. This research uses all the environmental factors along with the drought score for the prediction of crop production. As we can see in Figure 12 the variation in the production rate of Rice crops is based on the different factors.

5. **Can you clarify the research aims in the context of drought? The research aim is making predictions based on historical data and as such time series analysis is relevant. Can you discuss your motivation for not using time series?**

   The aim of this research is to investigate to what extent machine learning frameworks predict crop production in support of SDG13 climate action. However, In terms of the drought, The primary research aim is to investigate and understand the impact of environmental conditions, specifically drought, on crops like corn, rice, oat, soybean and wheat production. The study seeks to explore how changing climatic factors contribute to fluctuations in yields of these crops across different states and years.

   The primary reason for the selection of the model is, that I had to merge two distinct datasets: one containing soil moisture data and another comprising historical crop production records. The challenge was the fact that the historical crop production dataset covered only a short time span of three years for each crop category (Corn, Oat, Soybean, Rice, and Wheat) which is obtained from the U.S. Department of Agriculture(USDA).

   Time series analysis often requires a more extensive and temporally dense dataset to capture and model the temporal patterns and trends effectively. In my case, the limited three-year duration for each crop might lead to an overfitted model or, conversely, an insufficient representation of the true temporal nature.

   By using machine learning approaches like Radom Forest and XGBoost and deep learning approaches like feed-forward neural networks models were able to capture the complex relationship within the data without being overfitted. These techniques are well-suited for situations where the temporal aspect might be limited or challenging to model explicitly.

   In summary, the decision to choose the other models over the time series analysis was driven by the specific constraints of my data, where the historical crop production data covered only specific years for each crop. I chose machine learning methods as an alternative to effectively capture patterns and make predictions in a scenario with limited temporal data.

6. **Your data set contains temporal data. Can you describe how you checked for time-based dependencies and how your models capture those dependencies?**

As mentioned in the previous answer, I had to merge 2 different datasets, one containing soil moisture data and another comprising historical crop production records. The challenge was the fact that the historical crop production dataset covered only a short time span of three years for each crop category (Corn, Oat, Soybean, Rice, and Wheat) which is obtained from the U.S. Department of Agriculture(USDA). Time series analysis often requires a more extensive and temporally dense dataset to capture and model the inherent temporal patterns and trends effectively, Otherwise model is mostly prone to overfitting. However, I used the 'year' feature for the models which is correlated to the dependent variable (production). This will capture the time-based dependency in the Neural Network, Random Forest.

7. Which correlation measure did you use to perform feature selection and can you describe how you be certain it is the most suitable for your data?

I utilised the Pearson correlation coefficient as the correlation measure for feature selection for this analysis, as implemented by the Pandas `df.corr()` method. Following are some of the reasons for choosing this to ensure its suitability for my data:

(a) **Type of Data:** Pearson correlation is well-suited for assessing linear relationships between continuous variables. Given that the dataset primarily consists of numeric variables which are continuous in nature like temperature, pressure, precipitation, etc, the Pearson correlation coefficient was a natural choice.

(b) **Distribution and Linearity:** The heatmap visualization generated by the code helps to assess both the distribution and linearity of the relationships. Pearson correlation is appropriate when examining linear associations between different variables in the dataset.

(c) **Robustness to Outliers:** Pearson correlation is sensitive to outliers, However, the visualization produced by the code helps for a visual inspection of potential outliers and their impact on the correlation structure.

(d) **Practical Considerations:** The choice of Pearson correlation is practical for its interpretability. The heatmap provides a clear visual representation of the strength and direction (directly or inversely proportional) of correlations as shown in the correlation matrices in the code.

8. One of the models you produced was an Ordinary Least Squares regression model. How did you ensure that the assumptions that underlie this model hold?

In the context of the Ordinary Least Squares (OLS) regression model, the primary assumption I focused on was the absence of perfect multicollinearity. This assumption specifies that independent variables should not be perfectly correlated, as it can lead to unstable estimates of the regression coefficients. So, I plotted the correlation and performed the necessary steps to avoid very strong and very weak correlations among the features.

However, the OLS model was mainly used as a basic or starting model for the experiments. The performance of this model did not show a significant contribution compared to other models like Random Forest, Gradient Boost and Feed Forward Neural Network which I have discussed in the result section of my report.

9. How did you determine that the MAE and MSE scores you obtained are indicative of precise prediction?

As the crop production prediction model is a regression model MAE (Mean Absolute Error) and MSE (Mean Squared Error) are the best choices for the evaluation. In the context of crop production prediction, the MAE and MSE scores provided a clear view of the precision and accuracy.

To explain it based on one of the experiments where wheat production using the Random Forest Model.

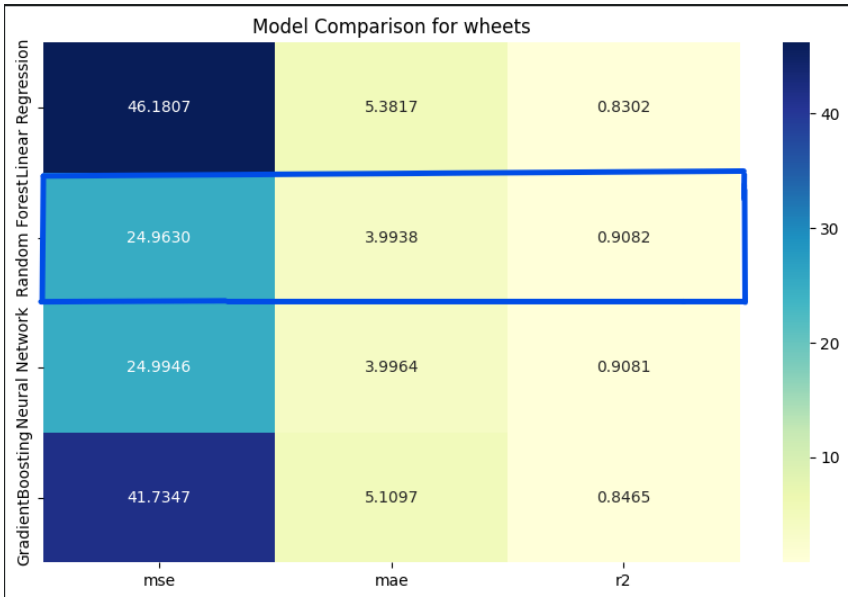


Figure 13: MAE, MSE and R^2 value for wheat production model evaluation

- Mean Absolute Error (MAE):

  The MAE represents the average absolute difference between the predicted and the actual values. In my case, the MAE of approximately 3.99 indicates that, on average, the model's predictions deviate by around 3.99 units from the actual Wheat Production values. A lower MAE indicates better precision, and in this example, the relatively small MAE suggests that the model tends to make accurate predictions with limited error.

- Mean Squared Error (MSE):

  The MSE, calculated as 24.96, measures the average squared difference between predicted and actual values. Since the errors are squared, larger errors have a greater impact on the MSE. A lower MSE indicates less variance in the errors, and in this case, the MSE value of 24.96 suggests that, on average, the squared differences between predicted and actual values are relatively small.

- R-squared (R2):
  Additionally, the R-squared value of approximately 0.91. It represents the proportion of the variance in the dependent variable (Wheat Production) that is predictable from the independent variables (features) in the model. A higher R-squared value, closer to 1, indicates that the model explains a significant portion of the variability in Wheat Production.

So, As the models implemented are the regression models, MSE and MAE are the best choices for the indication of accuracy and precision. And R^2 value will give information on how well the model captures the nature of data.

References

Bondre, D. A. and Mahagaonkar, S. (2019). Prediction of crop yield and fertilizer recommendation using machine learning algorithms, *International Journal of Engineering Applied Sciences and Technology* **4**(5): 371–376.

Chaudhari, S., Sardar, V., Rahul, D., Chandan, M., Shivakale, M. S. and Harini, K. (2021). Performance analysis of cnn, alexnet and vggnet models for drought prediction using satellite images, *2021 Asian Conference on Innovation in Technology (ASIANCON)*, IEEE, pp. 1–6.

Colombo-Mendoza, L. O., Paredes-Valverde, M. A., Salas-Zárate, M. d. P. and Valencia-García, R. (2022). Internet of things-driven data mining for smart crop production prediction in the peasant farming domain, *Applied Sciences* **12**(4): 1940.

Huang, S., Tang, L., Hupy, J. P., Wang, Y. and Shao, G. (2021). A commentary review on the use of normalized difference vegetation index (ndvi) in the era of popular remote sensing, *Journal of Forestry Research* **32**(1): 1–6.

Hunt, K. M., Turner, A. G. and Shaffrey, L. C. (2018). The evolution, seasonality and impacts of western disturbances, *Quarterly Journal of the Royal Meteorological Society* **144**(710): 278–290.

Hwang, H., An, S., Lee, E., Han, S. and Lee, C.-h. (2021). Cross-societal analysis of climate change awareness and its relation to sdg 13: A knowledge synthesis from text mining, *Sustainability* **13**(10): 5596.

Jiang, W. and Luo, J. (2022). An evaluation of machine learning and deep learning models for drought prediction using weather data, *Journal of Intelligent & Fuzzy Systems* **43**(3): 3611–3626.

Kalimuthu, M., Vaishnavi, P. and Kishore, M. (2020). Crop prediction using machine learning, *2020 third international conference on smart systems and inventive technology (ICSSIT)*, IEEE, pp. 926–932.

Khaki, S. and Wang, L. (2019). Crop yield prediction using deep neural networks, *Frontiers in plant science* **10**: 621.

Khalid, S. et al. (2020). Agronomy-food security-climate change and the sustainable development goals, *Agronomy-Climate Change & Food Security*, IntechOpen.

Lehmann, N. (2013). *How climate change impacts on local cropping systems: A bioeconomic simulation study for western Switzerland*, PhD thesis, ETH Zurich.

Li, C., Li, H., Li, J., Lei, Y., Li, C., Manevski, K. and Shen, Y. (2019). Using ndvi percentiles to monitor real-time crop growth, *Computers and Electronics in Agriculture* **162**: 357–363.
URL: *https://www.sciencedirect.com/science/article/pii/S0168169918318337*

Liang, X.-Z., Wu, Y., Chambers, R. G., Schmoldt, D. L., Gao, W., Liu, C., Liu, Y.-A., Sun, C. and Kennedy, J. A. (2017). Determining climate effects on us total agricultural productivity, *Proceedings of the National Academy of Sciences* **114**(12): E2285–E2292.

Mariotti, A., Schubert, S., Mo, K., Peters-Lidard, C., Wood, A., Pulwarty, R., Huang, J. and Barrie, D. (2013). Advancing drought understanding, monitoring, and prediction, *Bulletin of the American Meteorological Society* **94**(12): ES186–ES188.

Medar, R., Rajpurohit, V. S. and Shweta, S. (2019). Crop yield prediction using machine learning techniques, *2019 IEEE 5th international conference for convergence in technology (I2CT)*, IEEE, pp. 1–5.

Nafii, A., Taleb, A., Mesbahi, M. E., Ezzaouini, M. A. and Bilali, A. E. (2022). Early forecasting hydrological and agricultural droughts in the bouregreg basin using a machine learning approach, *Water* **15**(1): 122–122.

Nigam, A., Garg, S., Agrawal, A. and Agrawal, P. (2019). Crop yield prediction using machine learning algorithms, *2019 Fifth International Conference on Image Information Processing (ICIIP)*, IEEE, pp. 125–130.

Pretty Rani, R. R. (2023). Climate change and its impact on food security, *International Journal of Enviornment and Climate Change* **13**(3): 104–108.

Schubert, S., Koster, R., Hoerling, M., Seager, R., Lettenmaier, D., Kumar, A. and Gutzler, D. (2007). Predicting drought on seasonal-to-decadal time scales, *Bulletin of the American Meteorological Society* **88**(10): 1625–1630.

Sellam, V. and Poovammal, E. (2016). Prediction of crop yield using regression analysis, *Indian Journal of Science and Technology* **9**(38): 1–5.

Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S. and Singh, A. K. (2021). Crop yield prediction integrating genotype and weather variables using deep learning, *Plos one* **16**(6): e0252402.

Zhang, X., Chen, N., Sheng, H., Ip, C., Yang, L., Chen, Y., Sang, Z., Tadesse, T., Lim, T. P. Y., Rajabifard, A. et al. (2019). Urban drought challenge to 2030 sustainable development goals, *Science of the Total Environment* **693**: 133536.