Final Project Report for ENGG2112

Harvesting Hope: Onion Growth Prediction to Alleviate the Global Food and Water Crisis

Tutorial 3, Group 4

Jonathan Tapia, 520473410
Lachlan Jones, 520475528
Syed Ahmad Sabaat, 510415790
Iris Shi, 510415790

FACULTY OF ENGINEERING

November 3, 2023

Executive Summary

The purpose of this report is to use a Linear Regression and XGBoost regressor based Machine Learning Models to predict onion crop yield for a given location over a one year time period using historical weather data. It focuses on onion cultivation as a widely prevalent crop and serves to aid farmers in maximising the efficiency and output of their harvests. Ultimately, this report aims to combat the overwhelming issues of world hunger and water scarcity by increasing efficiency of resource use and food accessibility. It was found that the combined effect of the chosen features had a strong correlation to yield with an R^2 value of 0.63. Whilst temperature and dew had the highest individual impacts, wind speed, humidity and cloud cover appeared to be uncorrelated. The highest individual R^2 achieved was only 0.24, this is believed to be because onion yield requires a range of certain conditions which cannot be provided by only one weather phenomenon.

Table of Contents

1.0 Project overview and objectives

1.1 Problem Statement

Our current practices in crop growth are destroying our planet, with excess water usage and insufficient output of food to distribute to the areas that need it most, leaving the world in a food crisis and an epidemic of water scarcity. If these issues continue to worsen, the agricultural sector will cripple, outputs of crops will plummet and water availability will be diminished almost entirely. Our project stands to work against these devastating trends by educating stakeholders on the key weather-based factors influencing water usage and crop yield, so that it can be implemented on a global scale.

1.2 Background and Motivation

This project is focused on the influence of weather on crop yield and how we can optimise this to assist farmers to achieve greater efficiency of water usage and increase production to alleviate the global food crisis. Specifically, we centred our project on onion cultivation, as it is heavily dependent on weather conditions and mimics the holistic requirements for crop growth. Onions are the 2nd most grown and consumed vegetable in the world with a staggering 93 million tons of total growth in the year 2020 alone. The countries of largest onion output are China with over 24 million tons followed by India achieving 21 million tons [1]. Onions are a staple in almost any diet, containing medicinal properties such as vitamin C, vitamin B6, and folate. It also contains the flavonoid quercetin, which is known to prohibit the creation of cancer-causing elements.

Our motivation lies in two areas: The first being World Hunger and the second being Water Scarcity. These issues continue to plague the world and are steadily worsening. For perspective, 828 million people go hungry every day [2]. Additionally, at the current rate of water usage, more than 5 billion people will face water shortages by the year 2050 [3]. Agriculture accounts for 70% of worldwide water usage and thus we believe focusing on crop growth is the most vital approach to attacking these global issues [4].

1.3. Project Objectives

The initial aim of this project was to build a linear regression based Machine Learning model using a dataset on global onion crop yield figures for the years 1982 - 2021 merged with a dataset with historical weather data for those same years. This initial model assumed linear trends for variables but we determined that the results obtained would not necessarily be accurate as the features themselves were of non-linear nature. We adjusted for this by adopting the XGBoost Regressor model due to its resistance against non-linearity. In the short-term, this algorithm aims to:

- Determine the impact of combined weather conditions on onion yield to confirm their correlation.
- Following this, it will determine the most influential weather parameters on onion yield.
- Using these results, specific countries can be deployed into the algorithm to determine their suitability for crop growth based on their past crop yield and historical weather conditions.

These can then be put into practice to support our long-term objectives of addressing global water scarcity and hunger. Understandably, we will not be able to eradicate these issues entirely, but by achieving these short-term objectives, we will be able to educate key stakeholders on the strategies required to combat these issues:

Water Scarcity:

- By informing the key stakeholders of the suitability of global regions for crop growth, emphasis can be placed on the most efficient regions with higher precipitation and dew, requiring less irrigation, increasing access to water which can be reallocated to water scarce regions

World Hunger:

- These lower irrigation costs will decrease the cost of onions themselves, improving affordability, which can then specifically target those of lower socioeconomic status.
- Additionally, these regions with greater efficiency will generate higher yield, increasing food availability, so that it can be distributed to areas in greater need.

2.0 Methodology

2.1 Data Loading and preprocessing

The project collected historical crop yield data for onions and comprehensive weather data from two major data banks. Namely, the onion yield data set was retrieved from The Food and Agricultural Organisations of the United Nations (FAOSTAT) and the weather data was retrieved from VisualCrossing, a major historical weather data base. These datasets included information such as temperature, humidity, precipitation, dew, wind speed and cloud coverage over several years, alongside corresponding onion crop yield records. The data was cleaned, preprocessed, and organised for analysis. To elaborate on this process, the data was cleaned by ensuring countries with only full entries from 1982-2021 were passed, as NaN values must not be present as the XGBoost regressor assumes such. The data was also ordered alphabetically and by ascending year, this is critical as we were using two datasets meaning without matching corresponding rows we would have an unmatched dataframe. For the weather data we had daily weather reports for the corresponding countries, however the yield was only yearly. Hence, we took the mean weather conditions for each year and equated those to each corresponding yield for that year. Finally the data was merged utilising the pandas .merge function.

2.2 Feature Extraction

Feature extraction was primarily done through initial research and data collection, this is because both websites offered several potential features, many of which would be benign to our model hence we concluded the most likely features to impact the actual crop yield based off other studies and when retrieving the data we simply selected the features through their convenient user interface provided. The final features used were; temperature, dew, humidity, precipitation, wind speed and cloud cover. It is important to note we initially tried using UV index however too many countries had no entries for the data and restricted the size of the model hence it was decided to drop the feature. To refine these features, when running our model we created a for loop to normalise each feature individually and remove values with a z-score greater than three. This is because out models can be largely impacted by outliers. As mentioned in the data preprocessing, the weather data was also condensed into yearly averages, as this was the only way to create equal dimensions between the two data sets.

2.3 Classification

Our data was strictly numerical and we wanted a numerical value for yield, not something such as will yield increase or decrease from the previous year. Hence classification is not within the scope of our project

2.4 Linear Regression

Linear regression is a fundamental supervised machine learning algorithm used for predicting a numerical outcome (a continuous target variable) based on one or more input features. It assumes a linear relationship between the input variables and the output variable. The goal of linear regression is to find the best-fitting linear equation that minimises the difference between the predicted values and the actual observed values. This minimization is typically achieved by adjusting the coefficients in the linear equation. It was our initial goal to develop a range of linear regression models to compare, analyse and ultimately predict the impact of each feature individually as well as the confounding effect. However as we progressed we came to the realisation that a traditional linear regression model was not suited for our problem due to the non-linear correlation of weather and crop growth at extreme values. Simply put, sunshine is good for most crops, however too much sunshine and the crops may die. A standard linear regression model can be represented in the equation:

$$Y = \beta 0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

Where Y is the predicted output, X_n are the input feature $\beta 0$ is the y-intercept and β_n are the coefficients associated with each input feature

After running a few initial linear regression models yielding bad results, we researched alternative non-linear regression models and decided to switch to an Extreme Gradient Boosting Regressor (XGBoost).

2.5 Extreme Gradient Boosting (XGBoost)

The XGBoost Regressor was used to create all our final models. XGBoost is a popular and powerful ensemble learning algorithm with the capability for both regression and classification tasks. It combines multiple decision trees to create a robust and accurate predictive model. The steps involved in the XGBoost model are:

- It starts with a single decision tree.
- It calculates the errors made by the tree on the training data.
- It adds a new decision tree to the ensemble to correct these errors.
- This process is repeated for a specified number of iterations or until a stopping criterion is met.
- The final prediction is a combination of the predictions from all the trees in the ensemble.

To elaborate, ensemble learning is a machine learning technique which combines the predictions of several models often referred to as base models or weak learners to ultimately create a stronger, more accurate model. The idea behind ensemble learning is to aggregate multiple predictions to reduce bias and variance of the final prediction, XGBoost does this through combining decision trees as the base models. Decision trees are used for both classification and regression, hence by extension explaining why XGBoost can perform either task. Decision trees are hierarchical structures that split data into subsets based on the value of input features, leading to a decision at each leaf node which is the terminating point of each individual tree. XGBoost employs 'gradient boosting' hence the name, which sequentially adds decision trees with each tree correcting the errors made by the previous. Finally continuous value predictions is a form of regression, where instead of trying to categorise a label, a continuous numerical value is predicted. As for the assumptions required for an XGBoost model, this is one of the fantastic parts, it is highly robust and versatile as it does not assume linearity, independence, homoscedasticity, or multicollinearity it only assumes the data is free from missing values and outliers due to its sensitivity to these entries. This once again highlights the important steps of preprocessing to ensure no NaN values or outliers were present. With this in mind, XGBoost appeared to fit our problem very nicely, allowing us to avoid issues such as the non-linearity of weather and the dependence you would find between various features such as higher temperatures often causing higher humidity. We then created several different iterations of the model which developed as we realised more steps to add for greater accuracy. Initially it was just the basic XGBoost regressor model using the xgboost library, we then realised the importance of altering hyperparameters and added a grid search algorithm to test a range of different hyperparameters to return the highest R^2 value. Finally we needed to filter outliers and implemented the normalisation component to eliminate values with a z-score greater than three. These progressive updates are displayed through *Figure 1 and 2.*

```python
# Iterate through each weather column
for column in weather_columns:
    # Define your features (X) which is the current weather column
    X = merged_data[[column]]

    # Impute missing values in X
    X = imputer.fit_transform(X)

    # Initialize a Linear Regression model
    model = LinearRegression()

    # Fit the model
    model.fit(X, y)

    # Make predictions
    y_pred = model.predict(X)

    # Evaluate the model
    mse = mean_squared_error(y, y_pred)
    r2 = r2_score(y, y_pred)

    # Store the model and results in the dictionary
    weather_models[column] = {
        'model': model,
        'mse': mse,
        'r2': r2
    }
```

```python
# Iterate through each weather column
for column in weather_columns:
    # Define your features (X) which is the current weather column
    X = merged_data[[column]]

    # Split the data into training and testing sets
    X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2, random_state=42)

    # Initialize the XGBoost Regressor model
    xgb_model = xgb.XGBRegressor()

    # Fit the best model on the training data
    xgb_model.fit(X_train, y_train)

    # Make predictions on the validation set
    y_pred = xgb_model.predict(X_val)

    # Evaluate the model
    mse = mean_squared_error(y_val, y_pred)
    r2 = r2_score(y_val, y_pred)

    # Store the model and results in the dictionary
    weather_models[column] = {
        'model': xgb_model,
        'mse': mse,
        'r2': r2
    }
```

Figure 1. Showing our initial change from linear regression to XGBoost

```python
# Define customized parameter grids for each weather column
parameter_grids = {
    'temp': {
        'learning_rate': [0.1, 0.01, 0.001],
        'max_depth': [3, 5, 7],
        'n_estimators': [150, 200, 250]
    },
    'dew': {
        'learning_rate': [0.05, 0.02, 0.001],
        'max_depth': [3, 5, 7],
        'n_estimators': [150, 200, 250]
    },
    'humidity': {
        'learning_rate': [0.1, 0.01, 0.001],
        'max_depth': [3, 5, 7],
        'n_estimators': [500, 600, 700]
    },
    'precip': {
        'learning_rate': [0.05, 0.02, 0.001],
        'max_depth': [3, 5, 7],
        'n_estimators': [150, 200, 250]
    },
    'windspeed': {
        'learning_rate': [0.1, 0.01, 0.001],
        'max_depth': [3, 5, 7],
        'n_estimators': [150, 200, 250]
    },
    'cloudcover': {
        'learning_rate': [0.05, 0.02, 0.001],
        'max_depth': [8, 10, 12],
        'n_estimators': [150, 200, 250]
    }
}
```

```python
# Iterate through each weather column
for column in weather_columns:
    # Define your features (X) which is the current weather column
    X = merged_data[[column]]

    # Calculate Z-scores for the current feature
    z_scores = np.abs(stats.zscore(X))

    # Set a threshold for Z-scores (e.g., 3)
    threshold = 3

    # Filter out rows with Z-scores greater than the threshold
    filtered_indices = (z_scores <= threshold).all(axis=1)
    X = X[filtered_indices]
    y = y[filtered_indices]

    # Check if X and y have the same number of rows
    if X.shape[0] != y.shape[0]:
        print(f"Skipping {column} due to shape mismatch.")
        continue
```

Figure 2. Showing our hyperparameter grid and z-score filtering

3.0 Simulation Results

3.1 Key Findings

3.1.1 Impact of Weather on Onion Yield:

In this study, we aimed to determine the significant impact of weather conditions on onion yield. Onions, a vital agricultural crop, are known to be sensitive to environmental factors. By analysing historical data, we found that weather indeed plays a substantial role in influencing onion yield. The collective weather effect yielded an R-squared value of 0.63, signifying a strong influence on onion yield.

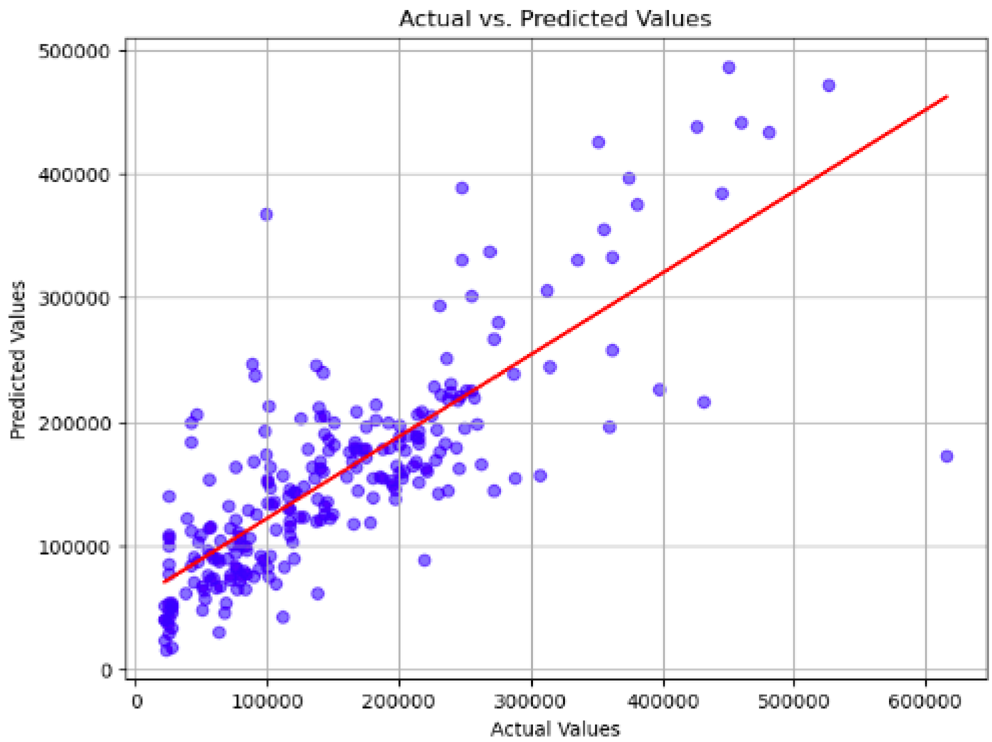

Figure 3. Actual vs Predicted Values For The Combined Weather Effect on Yield

3.1.2 Individual Weather Conditions:

To gain deeper insights, we proceeded to analyse the impact of individual weather conditions on onion yield. We evaluated temperature, humidity, precipitation, wind speed, and cloud cover separately. The results revealed that dew and temperature have the most significant impact on onion yield, with wind speed following closely. Humidity, precipitation, and cloud cover were found to be relatively less influential.

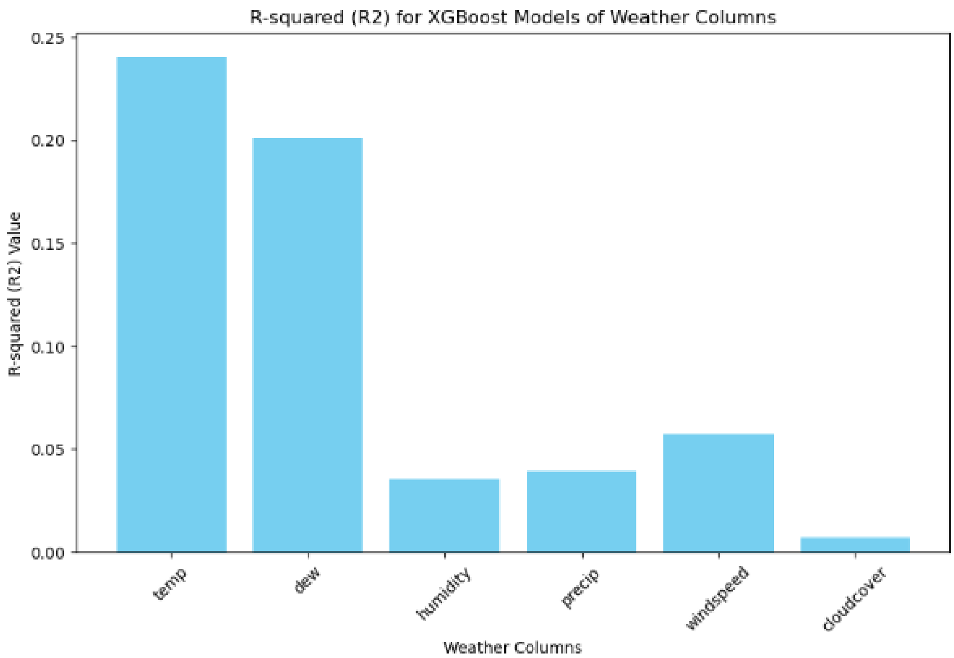

Figure 4. Bar Graph Depicting Individual Models of Weather Conditions Predicting Yield

3.1.3 Nonlinearity in Crop Growth:

The low R-squared values for individual weather conditions could be indicative of the nonlinearity of onion crop growth in response to weather. We observed a point at which excessive heat or precipitation begins to have adverse effects on yield. This phenomenon suggests that onion growth may follow a nonlinear pattern in relation to specific weather conditions. To further explore this, we recommend employing non-linear regression models, such as XGBoost Regressor, which have shown promise in capturing complex relationships in the data.

3.1.4 Year of Harvest:

A separate analysis was conducted to examine the relationship between onion yield and the year of harvest. There is a positive gradient in the line of best fit which suggests a trend of increasing yield over the years and is supported by the R-squared values.
This observation aligns with the logical expectation that advancements in agricultural technology, increased knowledge, and the development of improved herbicides and fertilisers have contributed to gradual improvements in onion yield over time.

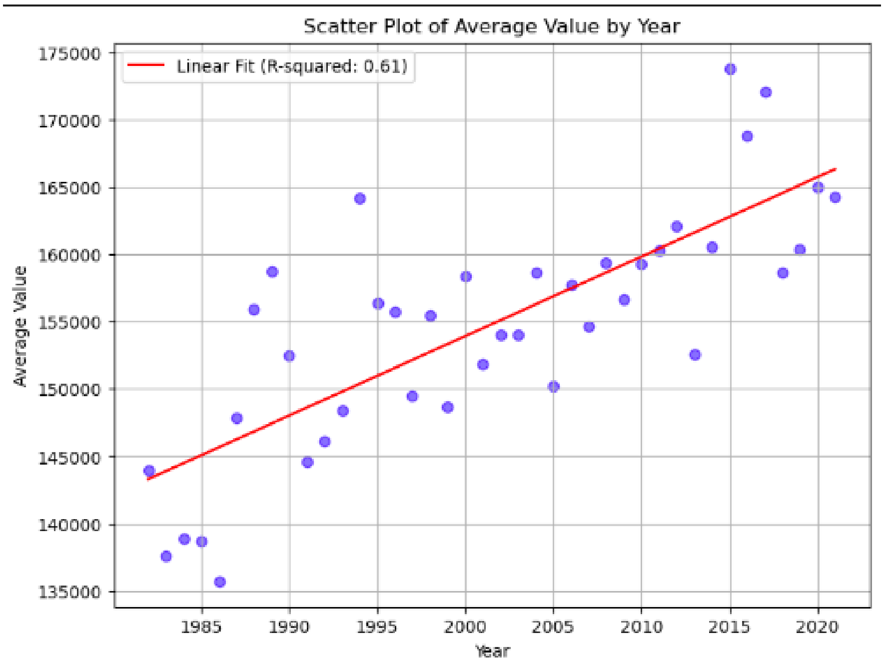

Figure 5. Scatter Plot of Average Yield vs Year to Determine Temporal Trend

In summary, our findings highlight the substantial influence of weather conditions, particularly temperature, dew, and to a small extent wind speed, on onion yield. The low R-squared values emphasise the complexity of this relationship, suggesting nonlinearity in crop growth patterns. Furthermore, the positive trend in yield over the years underscores the role of agricultural advancements in enhancing onion production. Future investigations should delve into nonlinear regression models to better capture the intricate interactions between weather and onion growth.

3.2 Issues Faced

Throughout the completion of this project we faced major challenges that forced us to re-evaluate our approach and although it slowed down the process severely for a large portion of the unit, it allowed us to generate creative solutions and produce a stronger final output than if the process had no setbacks.

3.2.1 Problem One: Dataset Difficulties

One of the primary challenges we encountered during this project was related to the datasets we used. The data we relied on for our analysis came from two distinct sources: onion crop yield data and historical weather data. These datasets posed several difficulties:

The onion crop yield data was collected from the Food and Agricultural Organization of the United Nations (FAOSTAT), and the historical weather data was obtained from VisualCrossing, however we were very delayed in getting both data sets as we were unable to find a data bank that suited the needs of our project, both being global, easily accessible, having a historical database and being free. The bottleneck we encountered was that many weather data banks required a subscription to use their service so we resolved this by splitting $30 between all group members allowing us to have the data we needed and proceed with our project

3.2.3 Problem Two: Availability and Completeness:

We discovered that not all countries had complete datasets spanning the entire time frame of our analysis (1982-2021). This posed a significant issue, as missing data points prevented the completion of our model. We had to filter out countries with incomplete data, which impacted the scope of our analysis.

3.2.4 Problem Three: Matching the Two Data Frames:

The two datasets had different structures and formats, making it challenging to merge them effectively. While the weather data included daily reports for each country, the yield data was available on a yearly basis. To align the data correctly, we had to aggregate the daily weather data into yearly averages. This step was crucial for meaningful analysis but added complexity to the data preprocessing process.

These dataset difficulties led to additional data cleaning, filtering, and restructuring, which prolonged the initial phases of our project.

3.2.5 Problem Four: Linear Regression Shortcomings and XGBoost Transition:

We quickly found out that whilst linear regression is easily interpretable, quick and non-computationally expensive the results it yielded were lacklustre and severely impacted any significance to our report. This meant as deadlines were approaching we had to find an alternative solution or else we would have very limited and meaningless findings. Upon some research into alternative continuous target variable machine learning options, XGBoost appeared to suit our needs greatly. However, whilst the code for the combined effect of weather on yield was relatively easy to adapt, we were unable to change the individual feature models to XGBoost due to dimension and scalar errors that we were unable to resolve easily. Eventually by reassessing how the data was originally formatted we found a more effective way to align them and after fixing this the XGBoost models all were able to run

4.0 Discussion and Recommendations

4.1 Significance of Findings

The primary focus of this study was the application of the XGBoost regressor model to predict onion crop yield using historical weather data. The results and analysis revealed several key insights into the relationship between weather factors and crop yield, as well as the potential for future research and practical applications.

4.2 Influence of Individual Weather Factors:

Our analysis underscored the varying impacts of different weather factors on onion crop yield. Notably, temperature and dew emerged as the most influential variables, exhibiting a positive correlation with yield. On the other hand, variables such as precipitation, humidity, windspeed, and cloud cover appeared to have negligible effects on crop yield. The R-squared values for these individual weather factors peaked at 0.24, indicating that no single weather phenomenon could singularly determine onion crop yield. This observation is in line with the complex and multifaceted nature of agricultural systems, where yield is influenced by an interplay of numerous variables.

4.3 The Combined Effect of Weather Factors:

A crucial insight gained from this study is that while individual weather factors may have limited predictive power, their combined impact is substantial. When considering the combined influence of all weather features, the model's performance significantly improved, as reflected by an R-squared value of 0.63. This suggests that it is the cumulative effect of various weather factors that plays a pivotal role in determining onion crop yield. These findings reinforce the idea that a holistic approach to modelling agricultural yield, considering multiple contributing factors, is essential for accurate predictions.

4.4 Temporal Trends in Onion Yield:

In addition to the weather-related findings, we observed a noteworthy temporal trend in onion yield. Over time, there has been a consistent increase in onion crop yield. This upward trajectory can be attributed to a range of factors, including improved agricultural knowledge, technology, and practices. The scatter plot analysis demonstrated this trend with an R-squared value of 0.61, indicating a reasonably strong correlation. This historical context underscores the importance of considering long-term trends in yield predictions.

4.5 Recommendations for Future Work:

Based on the insights gained from this study, several recommendations for future research and applications can be made:

Feature Selection: To improve the model's accuracy and reduce noise, it is advisable to remove the weather features found to be negligible (humidity, windspeed, cloud cover) in subsequent research. Additionally given the time, implementing new features that may have been overlooked such as soil moisture may improve the model.

Generalisation to Multiple Crops: Although we determined that Onions would be a viable representative crop to demonstrate the impacts of weather on crop yield, extending the model's application to predict the yield of various crops beyond onions holds promise. Investigating how different crops respond to weather conditions can provide valuable insights for diversified agriculture.

Real-World Deployment: To assess the practical utility of the model, future work should involve deploying it on current weather data. This would allow us to evaluate its ability to predict crop yield in real-time, potentially offering valuable decision-support tools for the agricultural industry.

4.6 Commercialisation and Translation Opportunities:

Our target audience was initially farmers, with a goal to educate them on the best crop growth strategies. However, the project's findings set up a strong foundation for us to commercialise by developing a prototype that can be presented to potential investors as well as for translation in the form of education to train future agronomists. Valuable opportunities may include:

- Agricultural Technology Startups such as BrightFarms or Blue River Technology Inc. that may have greater reach and ability to create user-friendly platforms for farmers to access and apply the model's recommendations.
- Government initiatives whereby they can utilise the predictive model in implementing agricultural policies and programs aimed at further improving the efficiency of irrigation systems and thus resource allocation.
- The education sector in which the knowledge gained from this project may be incorporated into the agricultural curriculum to train the next generation of farmers in effective and efficient farming practices.

5.0 Conclusion

The aspiration for the project is to mitigate world hunger by maximising yield of production of onions around the globe, and hence this model would support farmers to choose a location for the crop and adjust the amount of irrigation water to save resources. Throughout this project, the linear regression model and XGBoost Regressor model were explored through machine learning to identify the ideal weather conditions and global location for onion production. Our final results correlate to the proposed prediction of temperature and dew being the most important factor. However it is important to note that no single factor determines the yield as is the combined contribution of them which ultimately determines the yield. We believe our model has value in the agricultural world and could be expanded to predict many other types of crops, hence increasing its global utilisability

6.0 References

[1] "ISSUE 3 OF 2021," Mar. 2021. Accessed: Oct. 25, 2023. [Online]. Available: https://www.nab.com.na/wp-content/uploads/2022/04/Market-Intelligence-Report-Onion.pdf

[2] "World Hunger Facts To Know In 2023 | Concern Worldwide," Concern Worldwide US, Oct. 12, 2023. https://concernusa.org/news/world-hunger-facts/

[3] E. Soken-Huberty, "10 Facts About Water Scarcity," Human Rights Careers, Jun. 09, 2020. https://www.humanrightscareers.com/issues/facts-water-scarcity/

[4] OECD, "Water and Agriculture - OECD," Oecd.org, 2016. https://www.oecd.org/agriculture/topics/water-and-agriculture/

[5] J. Kubala, "9 Impressive Health Benefits of Onions," Healthline, Dec. 18, 2018. https://www.healthline.com/nutrition/onion-benefits (accessed Oct. 30, 2023).

[6] M. Hickey, "Onion growing," Jan. 2005. Available: https://www.dpi.nsw.gov.au/__data/assets/pdf_file/0017/126305/Onion-growing-Agfact-H8.1.22.pdf

[7] W. Cho, M. H. Na, Y. Park, D. H. Kim, and Y. Cho, "Prediction of Weights during Growth Stages of Onion Using Agricultural Data Analysis Method," Applied Sciences, vol. 10, no. 6, p. 2094, Mar. 2020, doi: https://doi.org/10.3390/app10062094.

[8] Md. N. Islam, A. Wang, J. S. Pedersen, J. N. Sørensen, O. Körner, and M. Edelenbos, "Online measurement of temperature and relative humidity as marker tools for quality changes in onion bulbs during storage," PLOS ONE, vol. 14, no. 1, p. e0210577, Jan. 2019, doi: https://doi.org/10.1371/journal.pone.0210577.

[9] P. A. J. van Oort, B. G. H. Timmermans, R. L. M. Schils, and N. van Eekeren, "Recent weather extremes and their impact on crop yields of the Netherlands," European Journal of Agronomy, vol. 142, no. 126662, Jan. 2023, doi: https://doi.org/10.1016/j.eja.2022.126662.