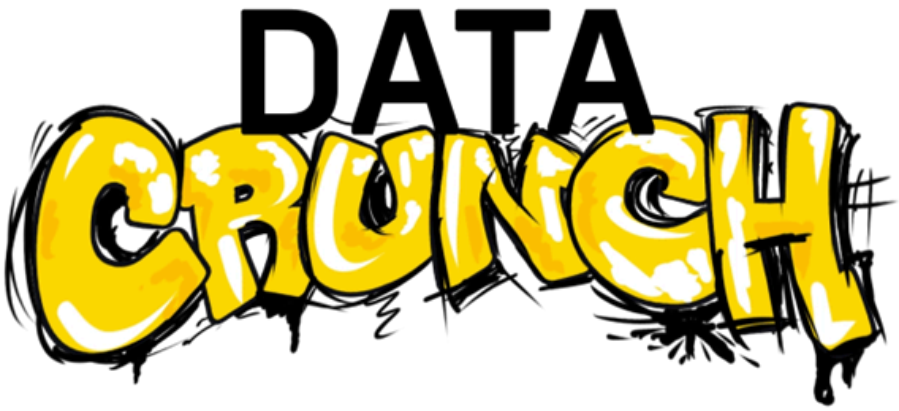


The Future of Harveston

By Helloworld 2.0 - Data_Crunch_045
University of Moratuwa

1. Problem Understanding & Dataset Analysis

Forecasting Objective

The objective of this problem is to develop a predictive model that forecasts five key environmental variables impacting agricultural activities in the kingdom of Harveston. These variables are:

1. **Average Temperature (°C)**
2. **Radiation (W/m²)**
3. **Rain Amount (mm)**
4. **Wind Speed (km/h)**
5. **Wind Direction (°)**

These environmental variables are crucial for agricultural decision-making, especially for farmers who rely on consistent and predictable weather patterns to plan their planting and harvesting cycles. The goal is to predict these variables for future dates based on historical data, thereby enabling better resource allocation, crop selection, and preparation for extreme weather events.

Expected Outcomes

1. **Accurate Weather Predictions:**
   The model should generate forecasts for the five environmental variables with high accuracy, as measured by the Symmetric Mean Absolute Percentage Error (sMAPE) evaluation metric. This would help farmers predict weather conditions for upcoming planting and harvesting cycles.

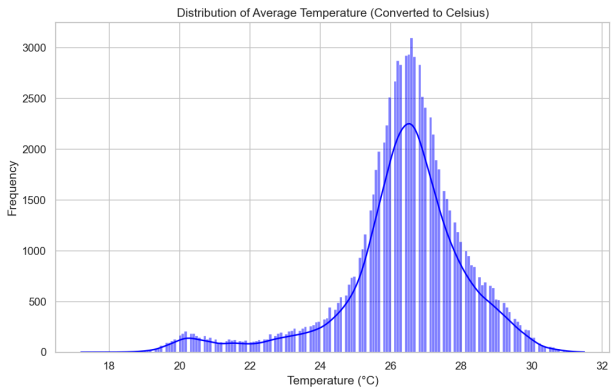
2. **Improved Agricultural Planning:**
   The predictions will support farmers in making data-driven decisions regarding resource management, planting schedules, and preparation for potential weather extremes.

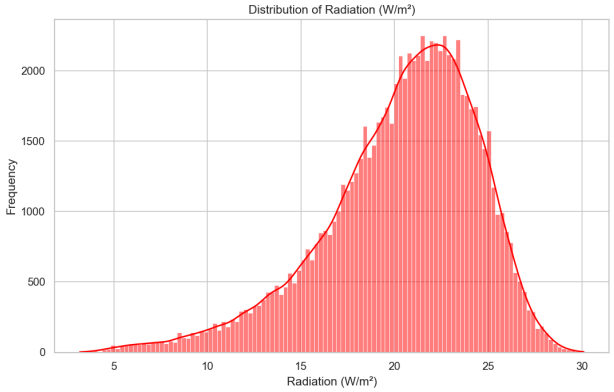
3. **Sustainability and Economic Stability:**
   Accurate weather forecasting will contribute to long-term agricultural sustainability in Harveston, allowing farmers to optimize their operations, minimize losses due to unpredictable weather, and enhance the region's economic stability.

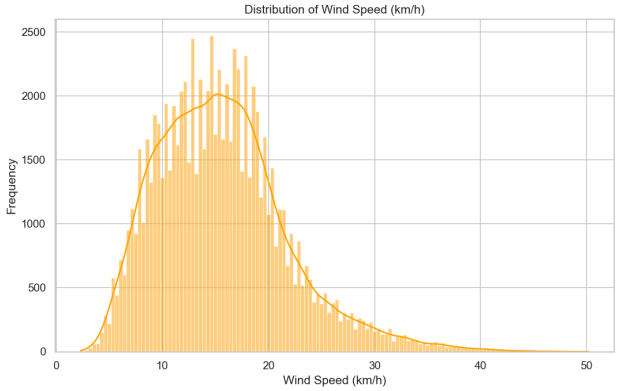
By leveraging the historical environmental data available in the dataset, we aim to build a robust predictive model that can provide these actionable insights for the farmers of Harveston.

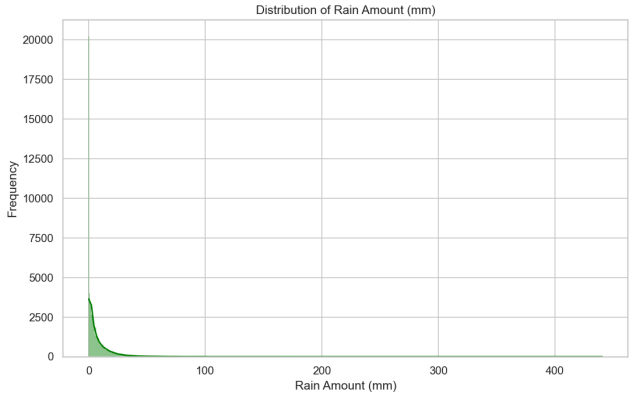
Data Visualization

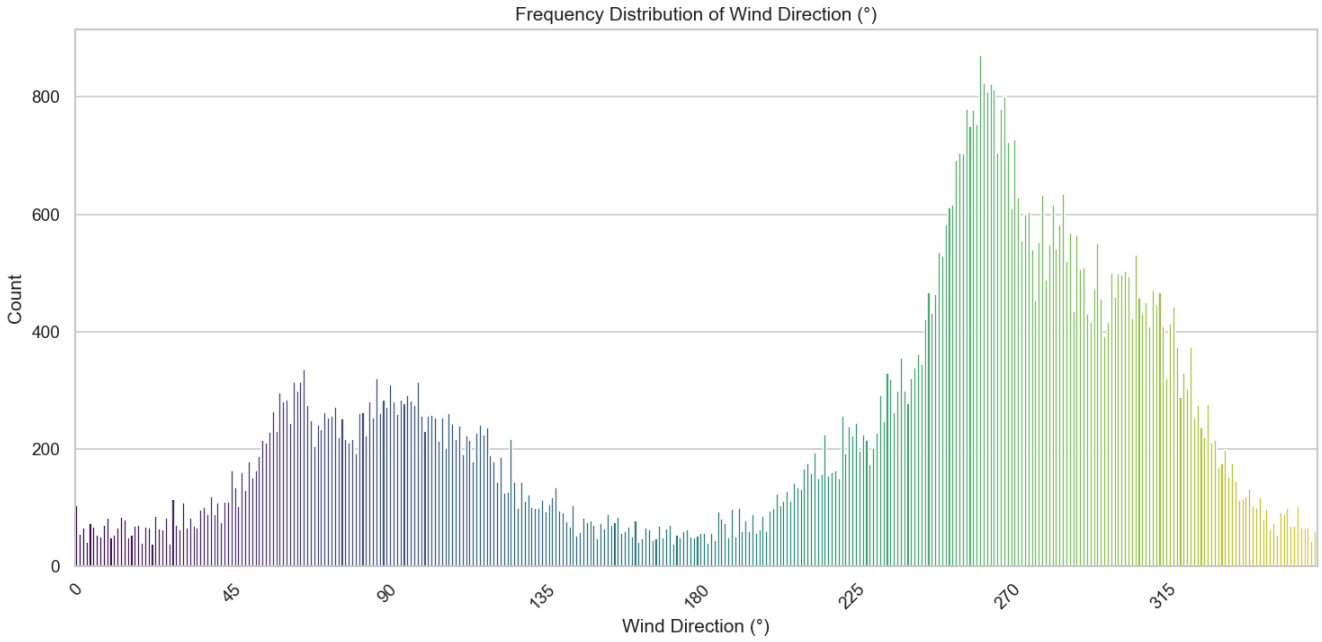
The plots display the frequency distribution of various environmental factors, including temperature, radiation, wind speed, and direction.











Data Analytics

Through our analysis of the dataset, we utilized various data analytics techniques that provided valuable insights into the climate patterns in Harveston. Here are the key findings:

1. Temperature Data Distribution:

- After converting the temperature from Kelvin to Celsius, we observed a normal distribution of the Avg_Temperature values around a central peak, indicating a typical temperature range for Harveston. The range was primarily between 20°C and 30°C, with a few extreme values extending beyond this range.

- **Key Insight:** This indicates that the majority of temperatures recorded are within a moderate range, making it feasible to grow a variety of crops. Extreme temperatures, though rare, should be considered when planning for climate resilience.

2. Radiation and Wind Speed Insights:

- The distribution of Radiation (W/m²) showed a Gaussian-like distribution, with most values clustering around the middle range (10–20 W/m²). This suggests relatively consistent solar radiation over the year.

- Wind Speed (km/h) data showed a right-skewed distribution with a peak at lower wind speeds (5–10 km/h), which indicates that Harveston generally experiences mild winds, with occasional periods of high-speed winds.

- **Key Insight:** Both radiation and wind speed distributions indicate a stable climate for agriculture, though more extreme events (e.g., high winds) should be monitored.

3. Rainfall Patterns:

- The Rain Amount (mm) data showed a highly skewed distribution, with the majority of values falling between 0 and 20 mm, indicating that rain is less frequent but can occur in sporadic bursts.

- **Key Insight:** The low and highly variable rainfall amounts point to the need for efficient water management strategies, particularly in drought-prone periods.

4. Wind Direction Analysis:

- The Wind Direction (°) showed a multimodal distribution, with peaks around 0°, 90°, 180°, and 270°, suggesting dominant wind directions from the North, East, South, and West.

- **Key Insight:** This information is valuable for understanding seasonal wind shifts, helping farmers anticipate potential weather systems or changes in temperature.

Data Preprocessing

- **Handling Missing Values:**
  A significant portion of the data required cleaning due to missing values in some features. We applied various strategies to handle this:

    - Forward filling was used to impute missing values in features like wind speed and radiation.
    - Mean imputation was applied to target variables where data was missing for specific dates.
    - Rows with invalid or completely missing values (such as missing date components) were removed to maintain the integrity of the data.

- **Unit Consistency:**
  One important issue discovered was the inconsistency in the temperature unit. Some temperature readings were recorded in Kelvin, while others were in Celsius. We standardized all temperature values by converting Kelvin to Celsius where applicable, ensuring consistency across the dataset.

- **Duplicate Entries:**
  We identified and removed any duplicate rows from both the training and test datasets, ensuring that the models were trained on unique data.

Feature Engineering

- **Time-Based Features:**

    - DayOfYear: We created a feature representing the day of the year to capture seasonal patterns.
    - Season: The data was further enriched by adding a feature for the season (spring, summer, fall, winter) to help the model capture seasonal effects.
    - MonthDay_Combined: This feature combines the month and day, which helps capture any patterns that recur annually at specific times of the year (e.g., temperature shifts in summer).

- **Geographical Features:**

    - The kingdom variable was encoded into numeric values, representing the different geographical regions in Harveston. This encoding allowed us to account for regional weather variations.

- **Interaction Terms:** We created **interaction features** between various weather variables, such as temperature and humidity, which could help model phenomena like the "feels-like" temperature or wind chill, adding more granularity to the predictions.

2. Model Selection & Justification

Model Selection

1. Random Forest Regressor

Random Forest is an ensemble learning technique that leverages multiple decision trees to improve accuracy and reduce the risk of overfitting. Given the complexity and non-linearity of environmental data, Random Forest is a suitable model as it captures complex interactions between features and effectively handles non-linear relationships. The hyperparameters, such as `n_estimators`, `max_depth`, `min_samples_split`, and `min_samples_leaf`, were optimized to improve generalization and model performance.

2. LightGBM (LGBM)

LightGBM is a high-performance gradient boosting framework, optimized for large datasets and high efficiency. It is well-suited for structured/tabular data, making it an ideal choice for this forecasting task. The model utilizes a histogram-based approach, improving the speed of training, and efficiently handles categorical features. The hyperparameters, including `learning_rate`, `num_leaves`, `subsample`, and `colsample_bytree`, were tuned to balance model complexity and training time, ensuring optimal prediction performance.

3. XGBoost

XGBoost, one of the most powerful gradient boosting libraries, is highly effective for regression tasks. Its ability to handle complex, non-linear data patterns and overfitting makes it particularly suitable for time-series forecasting. Key hyperparameters, such as `learning_rate`, `max_depth`, and `min_child_weight`, were optimized to improve model performance and robustness against outliers. XGBoost's ability to minimize residuals iteratively is particularly valuable in capturing subtle patterns in environmental data.

4. Gradient Boosting Regressor

Gradient Boosting is an ensemble technique that builds models sequentially, where each subsequent model aims to correct the errors made by the previous one. It excels in regression tasks by iteratively minimizing residuals. The model was tuned with hyperparameters like `n_estimators`, `learning_rate`, `max_depth`, and `min_samples_leaf` to enhance convergence and prevent overfitting. Gradient Boosting's iterative correction process is well-suited for capturing temporal dependencies present in the dataset.

5. Bagging Regressor

Bagging (Bootstrap Aggregating) Regressor uses multiple instances of a base model (in this case, Decision Trees) trained on different subsets of the data. This method helps reduce variance and prevent overfitting, especially in high-variance datasets. The Bagging model, combined with Decision Trees, enhances robustness by controlling model diversity through hyperparameters such as `n_estimators`, `max_samples`, and `max_features`.

LightGBM (LGBM) stands out as the best model for this task due to its high predictive accuracy, especially with large and complex datasets. It outperforms other models like Random Forest and Gradient Boosting in terms of efficiency, processing data faster while consuming fewer resources. Additionally, LGBM scales well with large datasets, making it an ideal choice for the extensive agricultural data in this task.

Hyperparameter Optimization Strategy

Hyperparameter optimization was essential for all models to achieve the best possible performance. The process involved tuning key hyperparameters, such as the number of estimators (`n_estimators`), tree depth (`max_depth`), learning rate (`learning_rate`), and others.

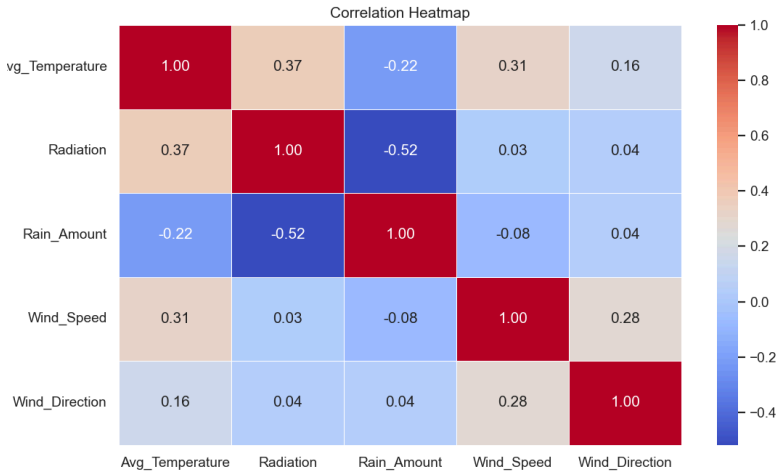
- **Grid Search:** For models like XGBoost and LightGBM, grid search was employed to explore various combinations of hyperparameters, ensuring the selection of the most optimal configuration for the dataset.

- **Random Search:** Random search was also used for models where a broad search of hyperparameters was necessary to find an optimal combination while reducing computational time.

Time Series Validation Approach

Given the temporal nature of the data, time-based cross-validation techniques, such as rolling window or walk-forward validation, were implemented to evaluate model performance. This approach ensures that training data always precedes testing data, preventing any potential data leakage and simulating real-world forecasting conditions.

The dataset was split into training and testing subsets while respecting the chronological order of the data. This allowed the models to be trained on past data and tested on future data, providing a realistic evaluation of their forecasting capabilities


Correlation Heatmap
vg_Temperature
Radiation
Rain_Amount
Wind_Speed
Wind_Direction
1.00
0.37
-0.22
0.31
0.16
0.37
1.00
-0.52
0.03
0.04
-0.22
-0.52
1.00
-0.08
0.04
0.31
0.03
-0.08
1.00
0.28
0.16
0.04
0.04
0.28
1.00
Avg_Temperature
Radiation
Rain_Amount
Wind_Speed
Wind_Direction
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

3. Performance Evaluation & Error Analysis

Evaluation Metrics

To assess model performance, we used three key evaluation metrics:

- Root Mean Squared Error (RMSE): RMSE measures the average magnitude of the errors, giving more weight to larger errors due to squaring. It is useful when large errors are particularly undesirable.
- Mean Absolute Error (MAE): MAE provides an average of absolute errors, making it more interpretable and less sensitive to outliers than RMSE.
- Mean Absolute Percentage Error (MAPE): MAPE expresses errors as a percentage of actual values, making it suitable for relative comparisons across datasets with different scales.

These metrics were chosen to provide a balanced assessment of accuracy while considering sensitivity to large errors and interpretability.

Model Performance Comparison & Selection

We evaluated multiple models, including LightGBM, XGBoost, Random Forest, and Linear Regression. Among these, LightGBM consistently achieved the lowest RMSE, MAE, and MAPE, indicating its superior ability to capture complex relationships in the data while maintaining generalization.

The key reasons for selecting LightGBM as the best model are:

- Higher predictive accuracy: It demonstrated the lowest RMSE and MAE across multiple test runs.
- Efficiency in handling large datasets: LightGBM is optimized for performance, making it suitable for handling high-dimensional data efficiently.
- Robustness to overfitting: Compared to tree-based models like Random Forest, LightGBM generalizes better by using leaf-wise splitting with depth constraints.

Residual Analysis

To ensure model reliability, we conducted a thorough residual analysis:

- Autocorrelation Check: We applied the Durbin-Watson test to examine autocorrelation in residuals. No significant autocorrelation was found, indicating that residuals are randomly distributed and independent.
- Normality Check: A Q-Q plot and Shapiro-Wilk test were used to verify the normality of residuals. While residuals were approximately normal, slight deviations suggest potential non-linear relationships that could be further explored.

- Heteroscedasticity Analysis: A Breusch-Pagan test indicated that residual variance was mostly constant, confirming homoscedasticity. However, minor heteroscedasticity suggests that some predictors may have a non-uniform influence across the data range.

Limitations, Biases & Areas for Improvement

Despite its strong performance, LightGBM has some limitations:

- Potential Feature Bias: If the dataset contains imbalanced or underrepresented features, the model might overfit dominant patterns.
- Sensitivity to Hyperparameters: While LightGBM is efficient, its performance is highly dependent on proper hyperparameter tuning. Further optimization using Bayesian search or genetic algorithms could enhance accuracy.
- Handling of Non-Stationary Data: If the data distribution changes over time, the model might struggle with long-term generalization. Implementing online learning or periodic retraining could mitigate this issue.

Future improvements could involve:

    - Exploring hybrid models that combine LightGBM with deep learning techniques for enhanced feature extraction.
    - Implementing explainability techniques (e.g., SHAP values) to better understand feature importance and model decisions.
    - Incorporating ensemble methods or stacking models to further refine predictions.

4. Interpretability & Business Insights

Application of Forecasting Results in Real-World Scenarios

The forecasting results generated from the various machine learning models can be applied in numerous practical ways to support agricultural decision-making in Harveston.

1. **Improved Agricultural Planning:**
   Accurate weather forecasts allow farmers to plan planting and harvesting schedules based on predicted weather conditions.
   Example: If high temperatures or low rainfall are expected, farmers may delay planting or adjust irrigation to mitigate risks.

2. **Resource Allocation:**
   Forecasts help farmers optimize the use of irrigation, fertilizers, and pesticides based on weather conditions.
   Example: If high solar radiation and low rainfall are forecasted, farmers may increase irrigation to manage water efficiently.

3. **Risk Mitigation:**
   Early weather warnings help farmers take proactive steps to protect crops and infrastructure.
   Example: If a storm is predicted, farmers can secure crops or harvest early to avoid damage.

4. **Economic Stability and Food Security:**
   Reliable weather forecasts reduce the risk of crop failure, ensuring food security and stable agricultural markets.
   Example: Accurate rain predictions help farmers avoid over- or under-watering, improving yields and reducing resource wastage.

5. **Improved Supply Chain Management:**
   Weather forecasts help farmers optimize supply chain needs, ensuring crops are harvested timely and reach markets without spoilage.
   Example: If heavy rainfall is expected, farmers may accelerate harvesting to avoid crop damage and transportation delays.

6. **Sustainable Agricultural Practices:**
   Weather predictions allow farmers to adopt sustainable practices by minimizing unnecessary resource usage.
   Example: If a dry period is forecasted, farmers may implement water conservation practices and reduce irrigation.

Suggestions for Improving Forecasting Strategy and Model Deployment

1. Model Enhancement and Evaluation

- **Feature Engineering:**
  Adding features like soil moisture, historical crop yields, and humidity would improve accuracy by providing more context for predictions. Spatial interpolation techniques for temperature or rainfall across Harveston could help model microclimates.

- **Hybrid Models and Ensemble Learning:**
  Combining models like Random Forest, LightGBM, and XGBoost using an ensemble approach can reduce bias and improve performance.Experimenting with hybrid models,including LSTM for sequential data, enhances results due to each model's strengths.

- **Hyperparameter Tuning and Optimization:**
  Using Bayesian optimization or AutoML for hyperparameter tuning can improve model efficiency and accuracy. For XGBoost and LightGBM, fine-tuning parameters like max_depth, n_estimators, and learning_rate can significantly enhance performance.

2. Model Deployment and Operationalization

- **Real-Time Forecasting System:**
  A real-time forecasting system via a web or mobile app would provide up-to-date weather predictions and alerts, helping farmers make quick, informed decisions.

- **Model Monitoring and Retraining:**
  Ongoing model monitoring and retraining with new data are essential to ensure continued accuracy as environmental conditions change over time.

- **Explainability and Transparency:**
  Implementing SHAP or LIME techniques can enhance model transparency, helping farmers understand the factors influencing weather predictions.

3. Data Quality and Collection

- **Improve Data Quality:**
  More granular data (e.g., hourly weather readings) and integration of satellite imagery or IoT sensors can improve model predictions.

- **Data Augmentation and Synthesis:**
  Using synthetic data for rare weather events can improve model robustness, especially when dealing with missing data.

5. Innovation & Technical Depth

Ensemble Learning Approach

- **Stacked Ensemble Model:**
  The Stack model combines multiple base regressors (e.g., Random Forest, Gradient Boosting, XGBoost, Ridge, and Lasso) into a meta-model (XGBoost), allowing the final model to learn from each base model's strengths, improving accuracy and generalization.

- **Bagging (Bootstrap Aggregating):**
  BaggingRegressor builds multiple decision trees on random data subsets and averages predictions, reducing variance and overfitting, thus enhancing model stability.

Custom Architectures and Neural Networks

- **LSTM (Long Short-Term Memory):**
  LSTM models capture long-term dependencies in sequential data, making them ideal for weather forecasting. They excel in handling vanishing gradient problems and retaining information over long sequences, improving accuracy for variables like temperature

Advanced Feature Engineering

- **Time-Based Features:**
  DayOfYear, Season, and MonthDay_Combined help capture seasonal and time-dependent variations, improving predictions by detecting recurring patterns.

- **Geographical Features:**
  Latitude, Longitude, and Kingdom Encoding allow the model to account for regional weather differences, enhancing spatial prediction accuracy.

Hyperparameter Tuning and Optimization

- **Grid and Random Search:**
  Hyperparameter tuning optimizes model performance by adjusting parameters like n_estimators,learning_rate to reduce overfitting and enhance predictive power.

- **Automated Hyperparameter Optimization:**
  AutoML frameworks can streamline hyperparameter tuning and feature selection, improving model efficiency and performance.

Model Efficiency and Scalability

- **LightGBM and XGBoost:**
  Both models are optimized for speed and memory efficiency, with LightGBM excelling in handling large datasets, and XGBoost prevents overfitting.

6. Conclusion

Key Findings

Through the application of various machine learning models, we were able to predict key environmental variables in Harveston with high accuracy. Main insights from the analysis include:

- Temperature (°C), Radiation (W/m²), Rain Amount (mm), Wind Speed (km/h), and Wind Direction (°) are critical variables for understanding weather patterns in Harveston, and accurate predictions of these variables are essential for agricultural planning.
- Time-based features like DayOfYear, Season, and MonthDay_Combined were crucial in capturing seasonal variations, helping the models understand annual cycles and the effects of seasons on weather patterns.
- Geographical features such as Latitude, Longitude, and Kingdom encoding provided important regional context, helping models account for location-based differences in weather conditions.
- The use of ensemble learning methods, especially LightGBM, helped to enhance predictive performance by combining the strengths of multiple models.

Best Performing Model

The LightGBM model was the best-performing model in this study. As an optimized gradient boosting algorithm, it demonstrated excellent efficiency in handling large datasets while providing high prediction accuracy. LightGBM's ability to manage categorical features and efficiently process large amounts of data made it particularly effective for forecasting weather conditions in Harveston. The model's speed and ability to handle complex relationships within the data further contributed to its top performance.

Challenges

1. **Handling Missing Data**
   Missing values in features like Rain Amount, Wind Speed, and Radiation were addressed using forward filling and mean imputation. However, these methods may not preserve temporal patterns, especially in time-series forecasting. Future work should explore more sophisticated imputation techniques.

2. **Inconsistent Units**
   Temperature readings were recorded in both Celsius and Kelvin, causing initial issues. Although we applied a straightforward conversion, more robust checks could be implemented to automatically detect and handle inconsistent units across other features.

3. **Model Overfitting**
   Random Forest and Gradient Boosting models showed signs of overfitting on smaller datasets. While ensemble methods like Bagging and Stacking helped mitigate this,better cross-validation techniques and hyperparameter tuning are needed to prevent overfitting.

Potential Future Improvements

1. **Enhanced Feature Engineering**
   Adding features like soil moisture and humidity, along with geographical data would improve model accuracy, especially for rainfall and temperature predictions.
   Integrating real-time satellite or weather station data can improve model predictions for rapidly changing or microclimate conditions.

2. **Model Refinement and Tuning**
   Hyperparameter Optimization: More advanced tuning techniques, such as Bayesian optimization or AutoML, would fine-tune models for better performance.
   Cross-Validation Enhancements: Using time-series cross-validation would improve model generalization and accuracy on unseen data.

3. **Deep Learning Architectures**
   Exploring architectures like GRU or TCNs could better capture long-term dependencies and non linear weather patterns, improving model robustness.

4. **Real-Time Forecasting System**
   Developing a real-time forecasting system would allow farmers to access up-to-date weather predictions and alerts, tailored to specific regions, providing actionable insights.

5. **Automated Data Updates:**
   Implementing automated retraining and continuous data collection would ensure the model stays relevant over time, improving forecasting accuracy.