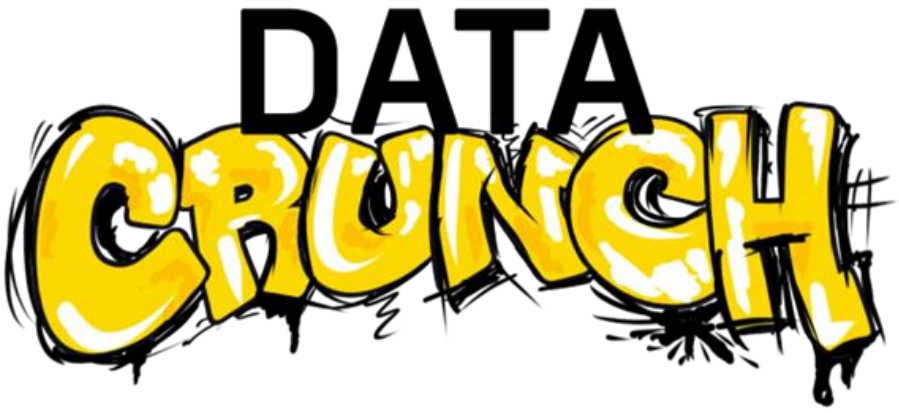


Preliminary Round Report

By Data diviners- Data_Crunch_152

University of Moratuwa

- # Problem Understanding & Dataset Analysis

  - ❖ **Forecasting Objective**

The problem requires developing time series forecasting models to predict five critical environmental variables in Harveston:

1. Average Temperature (°C)
2. Radiation (W/m²)
3. Rain Amount (mm)
4. Wind Speed (km/h)
5. Wind Direction (°), while only the year, month, day and kingdom are given.

These predictions will help farmers make informed decisions about planting cycles, resource allocation, and weather preparation.

- **Expected Outcomes**

1. Accurate multi-target time series forecasting model

2. Robust predictions that account for seasonal patterns and geographical variations

- **Key findings**

1. We identified that data spans **30** different kingdoms. So we need to consider the spatial concerns also during the feature engineering.

```
unique_kingdoms = train_data['kingdom'].nunique()
print(f"Number of unique kingdoms: {unique_kingdoms}")
```

```
Number of unique kingdoms: 30
```

2. During the dataset analysis, we identified date inconsistencies where certain non-leap years contained February 29, which is an invalid date. These errors were corrected by removing the affected entries. This preprocessing step ensured temporal consistency and improved model reliability.

```
# 1. Identify the rows to delete
february_29_rows_index = train_data[(train_data['Month'] == 2) & (train_data['Day'] == 29)].index
print(len(february_29_rows_index))
# 2. Delete the rows using the index
train_data = train_data.drop(february_29_rows_index)
```

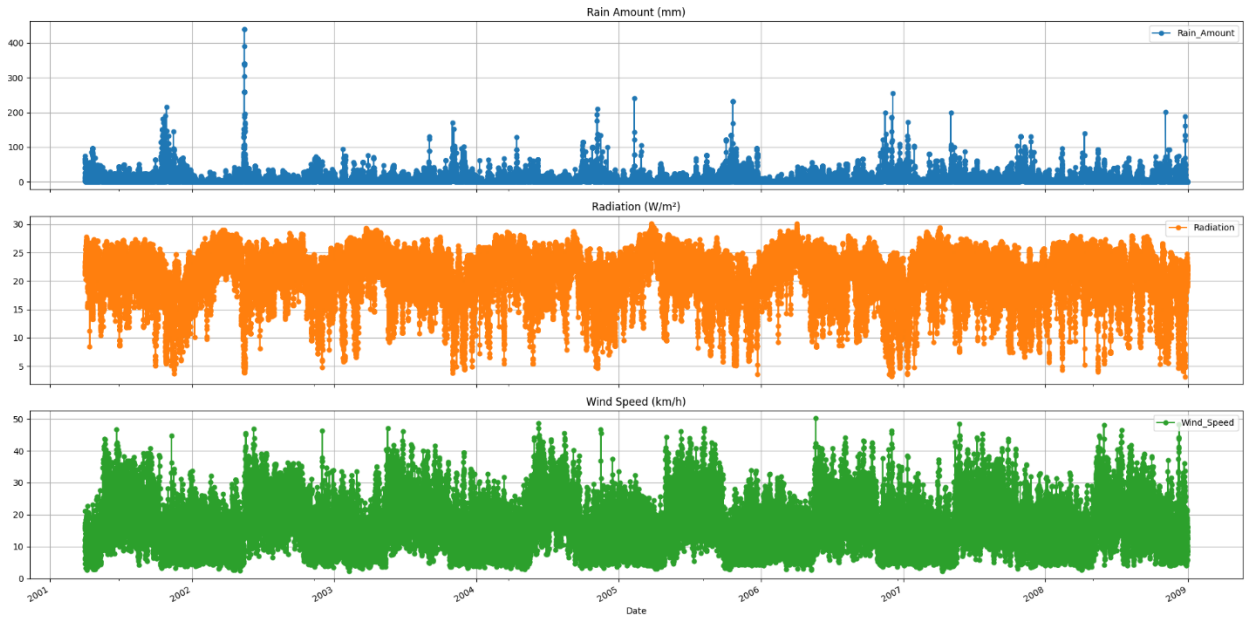
```
60
```

3. There are no missing values in the dataset.

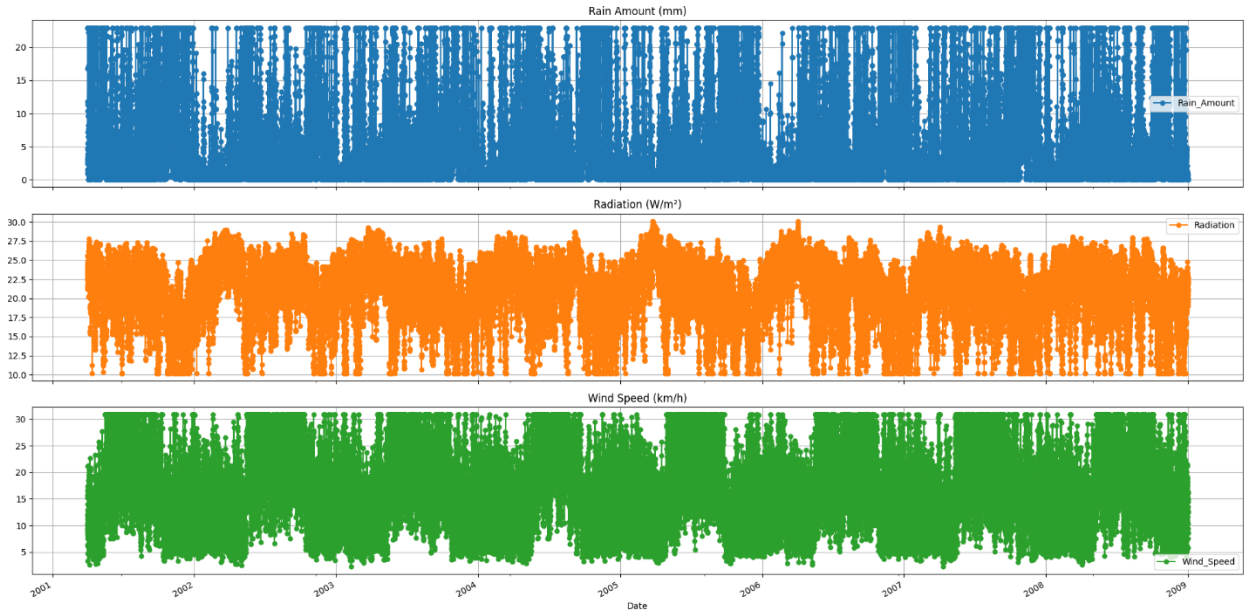
- **Preprocessing steps**

1. Temporal Handling:
    - Removed February 29 entries for consistency
    - Converted Year, Month, Day to proper Date format
    - Created cyclic encoding for month and day features (sin/cos transformations)
2. Geographical Features:
    - Added geo-clusters using K-Means (10 clusters) based on latitude/longitude
3. Outlier Treatment:
    - Identified and clipped outliers using IQR method for:
        - Rain_Amount (0.5% of data clipped)
        - Wind_Speed (0.3% of data clipped)
        - Radiation (0.4% of data clipped)

Before outlier removal:



After outlier removal:



- ## Feature Engineering & Data Preparation

  ❖ **Feature Creation**:

  - Rolling averages (3-day window) for temperature and wind speed
  - Lag features (1-day lag for temperature and rain)
  - Wind direction sin/cos encoding

  ❖ **Feature Selection**

  Dropped original categorical columns ('kingdom') and raw temporal features after creating derived features

  Retained engineered features that capture:

  - Temporal patterns (cyclic encodings)
  - Short-term dependencies (lags, rolling means)
  - Spatial patterns (geo-clusters)

❖ **Data Stationarity**

Applied MinMax scaling to target variables

Used Standard scaling for seasonal features

Considered differencing but found raw values with proper scaling worked better for this problem.

After data preprocessing and feature engineering and selection

```
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   ID                            84900 non-null  int64
 1   latitude                      84900 non-null  float64
 2   longitude                     84900 non-null  float64
 3   Avg_Temperature               84900 non-null  float64
 4   Avg_Feels_Like_Temperature    84900 non-null  float64
 5   Temperature_Range             84900 non-null  float64
 6   Feels_Like_Temperature_Range  84900 non-null  float64
 7   Radiation                     84900 non-null  float64
 8   Rain_Amount                   84900 non-null  float64
 9   Rain_Duration                 84900 non-null  int64
 10  Wind_Speed                    84900 non-null  float64
 11  Wind_Direction                84900 non-null  int64
 12  Evapotranspiration            84900 non-null  float64
 13  geo_cluster                   84900 non-null  int32
 14  month_sin                     84900 non-null  float64
 15  month_cos                     84900 non-null  float64
 16  day_sin                       84900 non-null  float64
 17  day_cos                       84900 non-null  float64
 18  temp_rolling_mean             84900 non-null  float64
 19  wind_rolling_mean             84900 non-null  float64
 20  temp_lag1                     84899 non-null  float64
 21  rain_lag1                     84899 non-null  float64
 22  wind_direction_sin            84900 non-null  float64
 23  wind_direction_cos            84900 non-null  float64
dtypes: float64(20), int32(1), int64(3)
memory usage: 15.9 MB
```

- ## Model Selection & Justification

LSTM-based Sequence-to-Sequence Model:

- Chosen for its ability to:
  - Capture temporal dependencies in time series data
  - Handle multiple input and output features
  - Learn complex non-linear patterns
- **Model Details**:
  - 2 LSTM layers (128 and 64 units) with dropout (0.3)
  - Dense layers for final prediction
  - Adam optimizer with MSE loss
  - Early stopping with patience=5
- **justification**
  - Compared with simpler models (ARIMA, Random Forest) in initial exploration
  - LSTM outperformed for multi-step, multi-variate forecasting
  - Handles the multiple seasonalities in the data (daily, yearly patterns)
  - Can effectively use the created lag features and rolling statistics
- **Hyperparameter Optimization**
  - Lookback window: Tested 30, 60, 90 days - 90 days performed best
  - Layer sizes: Experimented with different architectures
  - Dropout: 0.3 provided good regularization
  - Batch size: 64 worked well with our data size
- **Validation Approach**
  - Time-based validation split (last 20% of sequence for validation)
  - Walk-forward validation during testing

- ## Performance Evaluation & Error Analysis

Evaluation Metrics

Primary metric: sMAPE (competition requirement)

Additional metrics tracked during development:

- MAE (Mean Absolute Error)
- RMSE (Root Mean Squared Error)
- R^2 Score

Model Performance

- Validation loss plateaued after 10 epochs
- Final training MAE:

```
=== Validation sMAPE Scores ===
Avg_Temperature: 5.62%
Radiation: 6.93%
Rain_Amount: 61.42%
Wind_Speed: 14.56%
Wind_Direction: 17.02%
```

Error Analysis

- Highest errors occurred for Rain amount, after clipping the outliers also, the error doesn't come low.
- Wind direction showed some variability in errors
- Temperature predictions were most stable

Limitations

- Model may struggle with unprecedented extreme events
- Geographic variations could be better captured
- Longer training might improve performance but requires more resources

- ## Interpretability & Business Insights

Real-world Applications

1. **Planting Decisions**: Temperature and rain predictions help determine optimal planting times
2. **Resource Allocation**: Radiation forecasts assist in planning irrigation needs
3. **Risk Management**: Wind predictions help prepare for potential crop damage

Deployment Recommendations

1. **Continuous Learning**: Implement model retraining with new data
2. **Uncertainty Estimation**: Add prediction intervals for risk assessment
3. **Regional Specialization**: Train kingdom-specific models for better accuracy

- ## Innovation & Technical Depth

Innovative Approaches

1. **Multi-target LSTM**: Simultaneous prediction of all five targets
2. **Hybrid Feature Engineering**: Combining temporal, spatial, and weather features
3. **Custom Scaling**: Different scaling strategies for targets vs seasonal features
4. **Cyclic Encoding**: Proper handling of temporal periodicities

Technical Depth

1. Implemented sophisticated time series preprocessing
2. Developed custom sequence generation and forecasting functions
3. Incorporated geographical clustering into temporal model
4. Robust pipeline from raw data to predictions

- ## Conclusion

  The developed LSTM-based forecasting model provides accurate predictions for Harveston's critical environmental variables. The solution addresses the competition requirements through careful feature engineering, appropriate model selection, and thorough evaluation. The approach balances technical sophistication with practical applicability, offering real value to Harveston's agricultural planning.

  Recommendations for Improvement

1. Incorporate external weather data sources for improved accuracy
2. Develop ensemble models to combine strengths of different approaches
3. Implement more sophisticated attention mechanisms in the LSTM architecture
4. Add explicit modeling of inter-variable dependencies