Comparative Analysis of Pest Prediction using Weather Parameters

Rahul Pandey
Dept of Electrical Engineering
IIT Bombay
23M1063@iitb.ac.in
23M1063

Geevar Jos
Dept of Electrical Engineering
IIT Bombay
23M1164@iitb.ac.in
23M1164

Aaditya Nagare
Dept of Electrical Engineering
IIT Bombay
21D070001@iitb.ac.in
21D070001

***Abstract*—Pest prediction is crucial in assessing the risk of pest outbreaks and disease spread, offering vital information for forecasting and monitoring pest and disease problems. In sustainable agriculture management, understanding the importance of different features within a farm is paramount. This study compares various machine learning models to evaluate their Accuracy, Mean Squared Error, R-squared score and much more. By leveraging predictive analytics, this research aims to empower agricultural stakeholders with actionable insights for making tactical decisions regarding pest control measures. The findings of this study contribute to enhancing pest management strategies, enabling proactive interventions to mitigate risks and promote sustainable farming practices.**

I. Introduction

The study aims to predict the extent of Pink Bollworm infestation in Coimbatore from August to September, which is a critical period for protecting cotton crops. Pink Bollworms tend to attack cotton in the later stages, especially in September, which can significantly impact the yield and quality of the crop. The study aims to forecast infestation levels and enable preemptive measures by analysing weather data and making comparative predictive analyses. Through a comprehensive assessment of various weather parameters, the study aims to gain a better understanding of infestation dynamics and improve agricultural management strategies, thus reducing the economic losses associated with Pink Bollworm infestation in the area.

II. Data Description

The Indian Council of Agricultural Research provides the dataset used in this project on their website ICAR CRIDA DataBase. It contains weather parameters and pest values for the weeks of years 2001 to 2009. The Data corpus parameters are described in Table I

TABLE I
Parameter Description for Data Corpus

Column	Description
Observation Year	The year in which the observation was recorded.
Standard Week	The standard week number.
Pest Value	The measured pest value.
MaxT(°C)	Maximum temperature in Celsius.
MinT(°C)	Minimum temperature in Celsius
RH1(RH2(RF(mm)	Rainfall in millimeters.
WS(kmph)	Wind speed in kilometers per hour.
SSH(hrs)	Sunshine hours.
EVP(mm)	Evaporation in millimeters.
Pest Value	Measured pest value **Target Label**

III. Data Pre-Processing

A. Data Imputation

The data has empty values in the Pest value field from week 1 to week 30. The empty values are filled with zeros as the cotton crop is yet to be sowed in the fields.

B. Data Transformation

To accommodate the broad spectrum of values within the ”Pest Value” target field, a logarithmic transformation is applied using Equation [1]. This approach preserves zero values while effectively scaling non-zero values to a more manageable range.

$$\ln{(x+1)} \tag{1}$$

IV. Exploratory Data Analysis

A heatmap is plotted to understand the correlation between the various training parameters available in the data in Figure 1.

The trends for the available years of the data for Pink Bollworm adult's infestation are understood through figure 2.

V. Analysis of Models

A. Training

The comparative analysis involves various models: Ordinary Least Squares (OLS), Linear Regression, Random Forest Regressor, Support Vector Machine (SVM) with Radial Basis Function Kernel (RBF), and Convolutional Neural Network (CNN). The CNN is trained using both the entire dataset's parameters and a rolling window approach, focusing solely on the 'Pest Value' column to forecast the pest value for the subsequent week. Standard scaling is applied during training to enhance model performance.

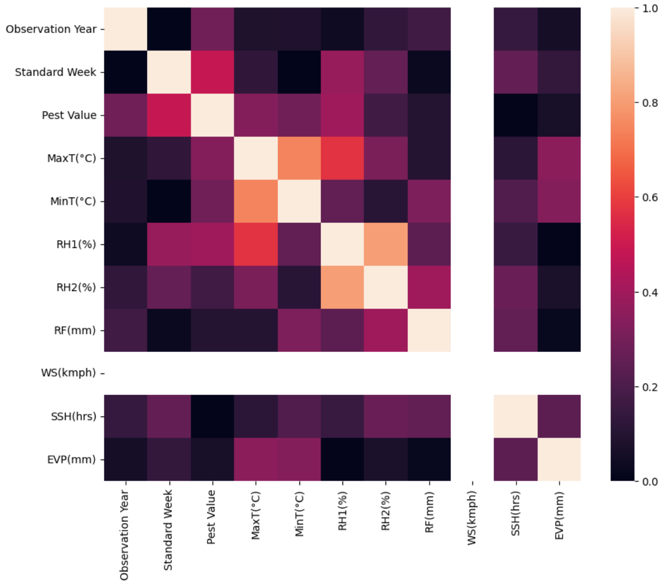

Fig. 1. Degree of Correlation between training Parameters

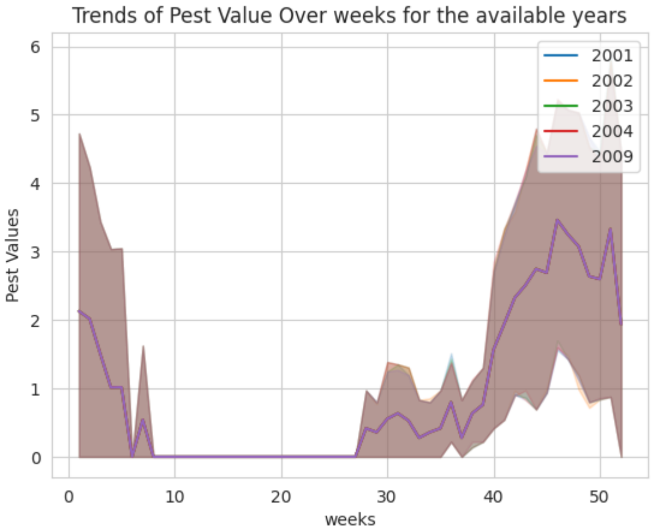

Fig. 2. Trends of Pink Bollworm adults degree of infestation across the years

B. Comparison

Following are the comparisons made through the comparative performance analysis

The Feature importance of of the parameters are estimated using the SHAP values of the SHAP values. SHAP (SHapley Additive exPlanations) values are a method used to explain the output of machine learning models. They provide insights into the contribution of each feature to the model's prediction for a particular instance. The core idea behind SHAP values is to assign each feature an importance score indicating how much that feature influenced the model's output.

TABLE II
MODEL PERFORMANCE COMPARISON

Model	MAE	RMSE	R-squared Score
OLS	0.9864	1.3883	0.498
Linear Regression	1.1292	1.4701	0.1476
SVM (RBF kernel)	1.0315	1.432	0.4670
Random Forest	0.5202	0.9141	0.7892
CNN	0.6542	1.1141	0.6869
CNN (Time Series)	0.0564	0.0901	0.8643

The Feature importance for the CNN model is plotted in Figure 3, whereas the Feature Importance for Random Forest Regressor is shown in 4.

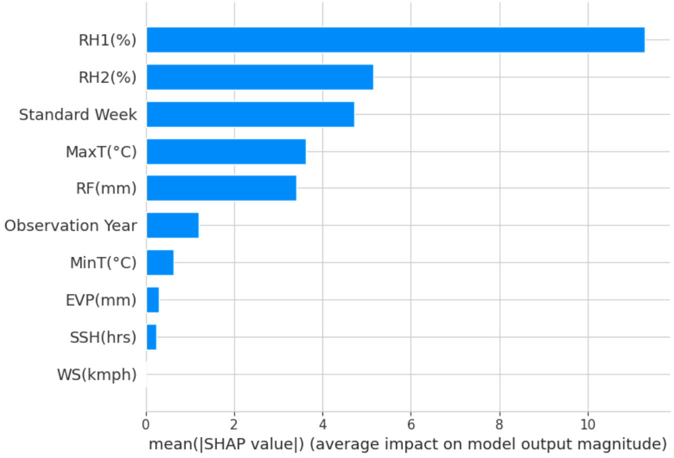

Fig. 3. Feature Importance in CNN Model

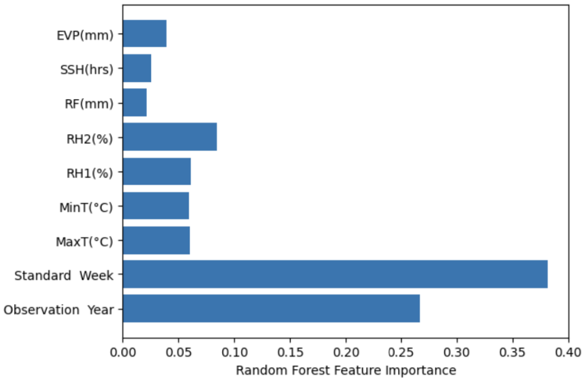

Fig. 4. Feature Importance in Random Forest Regressor

VI. Interpretation of Results

A. Linear Regression

The linear regression model has the highest MAE (0.9864) and RMSE (1.3883) compared to other models. This indicates that the linear regression model has the largest average difference between predicted and actual values. It also has the lowest

R-squared score (0.498) among all the models, signifying a weak correlation between the predicted and actual values.

B. Support Vector Machine (SVM)

The SVM model has a lower MAE (1.0315) and RMSE (1.432) compared to linear regression, but higher than the Random Forest and CNN models. It also has a lower R-squared score (0.4670) compared to Random Forest and CNN models, indicating a weaker correlation between predicted and actual values.

C. Random Forest

The Random Forest model has the lowest MAE (0.5202) and RMSE (0.9141) among all the models, indicating the smallest average difference between predicted and actual values. It also has a moderately high R-squared score (0.7892), signifying a good correlation between predicted and actual values.

D. CNN

The CNN model has a lower MAE (0.6542) and RMSE (1.1141) compared to linear regression and SVM models, but higher than the Random Forest model. It has a moderately high R-squared score (0.6869), signifying a good correlation between predicted and actual values.

E. Time Series CNN

The Time Series CNN model has the lowest MAE (0.0564) and RMSE (0.0901) among all the models, indicating the smallest average difference between predicted and actual values. It also has the highest R-squared score (0.8643), signifying the strongest correlation between predicted and actual values.

F. Overall Observations

The Random Forest and Time Series CNN models achieved the best performance based on the metrics provided in the table. They have the lowest MAE and RMSE, and the highest R-squared scores.

The CNN model trained on time series data proved to be the best among the models evaluated, successfully predicting time series patterns. Its exceptional performance, however, was solely attributed to its concentration on temporal patterns, ignoring the significance of weather parameters. Conversely, the CNN model trained with weather parameters displayed inferior performance compared to its time series counterpart. It's crucial to note that the relative humidity parameter was identified as a significant factor in the model.

The linear regression model performed the worst among the compared models based on the given metric.

VII. Conclusion

Based on the models, it appears that there is a strong link between high levels of humidity and the ideal growth conditions for the Pink Bollworm in the Coimbatore area. Therefore, it is crucial to take into account environmental factors when predicting pest infestations to improve accuracy and efficiency. Data transformation is vital for optimizing gradient and boundary optimization for models like CNN, SVM, and OLS and to standardize distance calculations in SVM and OLS. However, this step doesn't seem to play a significant role in training Random Forest-based models since they rely on thresholds rather than the distribution of parameter values.

It is also important to note that the choice of the best model may also depend on other factors that are not shown in the table, such as the specific task, the interpretability of the model, and the computational cost of training the model. For instance, while the Time Series CNN model has the best performance according to the given metrics, it might be a more complex model to train compared to a Random Forest model.

Acknowledgment

It is a great pleasure to express OUR deepest gratitude to Professor **Amit Sethi** for giving me this life- changing experience and prompt guidance throughout the **EE769 Introduction to Machine Learning Course**.

References

[1] Indian Council of Agricultural Research(ICAR), Data for Pink Bollworm for Coimbatore, http://icar-crida.res.in:8080/naip/AccessData.jsp. Accessed on Apr 2, 2024.

[2] Ronak Shah(203074001)), Pest Prediction from Weather Parameters Using Machine Learning Methods, Mtech Thesis for Electronic Systems, IIT Bombay, 2023.

[3] ChatGPT v3.5

[4] Gemini AI

[5] Colab code link