Agric Res J **63(1)** (Special Issue): 137-148, February 2026
DOI: 10.5958/2395-146X.2026.00014.8

MACHINE LEARNING TECHNIQUES FOR CLASSIFICATION AND PREDICTION OF RICE LEAF FOLDER (*Cnaphalocrocis medinalis*) INCIDENCE USING WEATHER VARIABLES

P Durga Bhavani*, B Ravindra Reddy, K Kiran Prakash*, P Lavanya Kumari, D Ramesh, N Kamakshi and G Amrutha Lakshmi

Department of Statistics and Computer Applications,
Agricultural College, Bapatla-522 101, Andhra Pradesh, India

ABSTRACT

Rice (*Oryza sativa*) is a staple food crop in India and across Asia, but its production is threatened by the rice leaf folder, *Cnaphalocrocis medinalis* (Guenee) (Lepidoptera: Pyralidae), a major pest capable of causing yield losses of 63-80% under severe infestation. Climate change further exacerbates pest outbreaks, posing challenges to food security. This study applied machine learning classification algorithms including Multinomial Logistic Regression, C5.0 Decision Tree, Random Forest, and XGBoost to analyze rice leaf folder incidence based on five weather parameters. Secondary data were collected from the Agricultural Research Station, Bapatla, spanning 390 Standard Meteorological Weeks (SMWs) from January 2011 to December 2024 including weekly leaf folder catches and climatological averages: maximum temperature (T_{max}), minimum temperature (T_{min}), morning relative humidity (RHM), evening relative humidity (RHE), and total rainfall (RF). Multinomial Logistic Regression achieved the lowest accuracy (73.08%), while C5.0 Decision Tree accuracy improved to 83.08% with T_{max} as the most influential factor. XGBoost identified T_{min} as the key contributor achieving 93.59% accuracy. Random Forest outperformed all models yielding the highest accuracy (96.67%), a kappa value of 0.9155 and superior sensitivity, specificity, precision and F_1 scores. These results demonstrate that Random Forest is the most effective algorithm for predicting rice leaf folder incidence using weather data which offers a reliable tool for pest management and mitigation under changing climatic conditions.

Keywords: Classification, Decision tree, Machine learning, Multinomial Logistic Regression, Random Forest, Rice leaf folder, XGBOOST

www.IndianJournals.com
Not for Commercial Use
Downloaded From IP - 216.73.216.57 on dated 17-Nov-2026

Rice (*Oryza sativa*) is the principal staple food crop of India and much of Asia. India has the largest area under rice cultivation globally covering approximately 44 million hectares and producing about 112 million tonnes annually (Reddy *et al*., 2022). Rice production is constrained by numerous insect pests; more than 100 species have been reported, of which around 20 cause significant economic damage (Pathak and Dhaliwal, 1981). Among these, the rice leaf folder, *Cnaphalocrocis medinalis* (Guenee) (Lepidoptera: Pyralidae), is a major pest across Asian rice-growing regions capable of causing yield losses of up to 63-80% under severe infestation.

Historically regarded as a minor pest, *C. medinalis* increased in abundance during the late 1980s and has since attained major pest status in many parts of the world (Ahmed *et al*., 2010). Severe infestations, particularly when synchronized with susceptible crop growth stages, can result in substantial economic losses. Damage becomes critical when more than half of the flag leaf and the two youngest leaves per tiller are affected. While rice plants can often compensate for damage during the vegetative stage, infestations during the reproductive phase are economically more consequential as feeding on flag leaves directly reduces photosynthetic capacity and yield.

Climate is a key determinant of agricultural productivity, influencing both crop growth and pest dynamics. Climate change has significantly altered pest incidence patterns posing a growing threat to food security. Rising temperatures have been shown to intensify damage caused by pests such as the rice leaf folder (Jeong *et al*., 2014). Climatic variables also affect pest populations and their natural enemies with some insect species exhibiting increasing trends over time (Ahmed *et al*., 2013).

In recent years, machine learning-based classification algorithms, including C5.0, CART and Random Forest, have been increasingly applied to analyze pest incidence using weather parameters (Jinubala and Jeyakumar, 2021; Orounla *et al*., 2023). Among these, Extreme Gradient Boosting (XGBoost) has been reported to provide superior predictive accuracy and reliability in modeling complex relationships between climatic factors and pest occurrence (Hussain *et al*., 2022).

*Corresponding author: kavuri1611@gmail.com
Date of receipt: 11.03.2025, Date of acceptance: 30.11.2025

Therefore, the present study aims to evaluate and compare the performance of multiple classification algorithms in predicting rice leaf folder incidence based on key weather parameters. By identifying the most accurate and reliable model, this study seeks to contribute to the development of data-driven decision support systems for effective pest forecasting and sustainable rice production under changing climatic conditions.

MATERIALS AND METHODS

Data description

The secondary data used in the present study were collected from the Agricultural Research Station, Bapatla. A light trap was used to monitor and collect pest data in the plant ecosystem throughout the research trials at the research station. The pests collected in the trap were brought to the laboratory in the morning, where the number of individuals caught per day were manually counted. The daily counts were then summed and expressed as weekly cumulative catches. The dataset used in this study consists of 390 Standard Meteorological Week (SMW) observations including meteorological variables such as maximum temperature (T_{max}), minimum temperature (T_{min}), morning relative humidity (RHM), evening relative humidity (RHE) and rainfall (RF), which were collected from the weather stations installed at Bapatla. Standard meteorological week-wise cumulative catches of leaf folder population and weekly averages of climatological parameters during the period from January 2011 to December 2024 were considered for the study as per data availability.

In present study, rice leaf folder incidence serves as the target variable, while the meteorological parameters act as predictor variables. These variables were selected because climatic conditions significantly influence insect development, survival and reproduction (Bale *et al*., 2002; Sharma, 2010). Temperature plays a fundamental role in insect development through thermal accumulation processes, where each stage of insect growth requires a specific amount of heat units (Awmack and Leather, 2002). Maximum and minimum temperatures, therefore, act as key environmental indicators influencing pest growth rates and population fluctuations. Extreme temperatures may lead to reduced survival rates or disruption of reproductive cycles. Relative humidity also plays a significant role in the biological development of leaf folder populations. High humidity levels generally enhance egg survival, larval development and adult activity, whereas prolonged periods of low humidity may negatively affect pest survival due to increased desiccation stress (Bale *et al*., 2002). Rainfall exhibits a complex relationship with pest incidence. Moderate rainfall can increase humidity within the crop canopy creating favourable conditions for pest survival, whereas heavy rainfall events may reduce pest populations by physically dislodging larvae or washing away eggs from plant surfaces. Weather variables such as temperature, humidity and rainfall play crucial roles in pest dynamics (Khan and Ramamurthy, 2004).

In the present study, rice leaf folder incidence was categorized into different infestation levels and machine learning classification models were used to identify the relationship between meteorological variables and pest occurrence. The statistical analysis and model development were carried out using R programming software.

Statistical analysis

Descriptive statistics

Descriptive statistical measures including mean, standard deviation, skewness, kurtosis, minimum, maximum, range, count and coefficient of variation (CV) were calculated to examine the central tendency, variability and distribution characteristics of rice leaf folder pest incidence and the selected meteorological variables (T_{max}, T_{min}, RHM, RHE and RF).

Multinomial Logistic Regression (MLR)

Multinomial Logistic Regression (MLR) is used when the dependent variable is categorical with three or more unordered classes. Unlike binary logistic regression, which models two outcomes, MLR models the probability of multiple class outcomes simultaneously. If the response variable has K classes, the model estimates K−1 logistic equations, where one category is treated as the reference class. The log-odds of class *j* relative to the reference class are modelled (Hosmer *et al*., 2013) as:

$$log\left(\frac{P(Y=j)}{P(Y=\ reference)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta n Xn \quad \ldots\ldots [1]$$

where,

P(Y = j) is the probability that the observation belongs to class j.

P(Y = reference) is the probability of the reference class.

X_1, X_2,....Xn are the predictor variables.

β_0, β_1, β_2 are the regression coefficients.

The parameters of the multinomial logistic regression model are estimated using the maximum likelihood estimation (MLE) method. To obtain class probabilities, the softmax function (Hastie *et al*., 2009) is used:

$$P(Y=j) = \frac{e^{\beta_j^T X}}{\sum_{k-1}^{K} e^{\beta_k^T X}} \quad \ldots\ldots\ldots\ldots [2]$$

www.IndianJournals.com
Not for Commercial Use
Downloaded From IP - 216.73.216.57 on dated 17-Nov-2026

where, *K* represents the number of classes and *X* represents the vector of predictor variables. The predicted class corresponds to the class with the highest probability. Model performance is evaluated using several classification metrics such as accuracy, confusion matrix, precision, recall, and F_1-score. The estimated coefficients are interpreted as the change in the log-odds of belonging to a particular class relative to the reference class for a one-unit change in the predictor variable.

DECISION TREE (C5.0 ALGORITHM)

The machine learning model considered for the analysis is the C5.0 Algorithm, which is Decision Tree based classifier algorithm. The decision tree is a technique to perform the classification. Decision trees have turned out as the most effective and notable techniques in machine learning and data mining. C5.0 algorithm is a development of the C4.5 algorithm, which is an expansion of the ID3 algorithm. It is a classification algorithm that is suitable for large datasets. C5.0 also uses a boosting method to generate and integrate various classifiers to deliver enhanced predictive accuracy. C5.0 works in three essential stages; initially, all samples are considered at the top of the tree called as the root node and forwarded them through to the second node called branch node. The branch node produces rules for a set of samples reliant on an entropy measure. At this point, the C5.0 creates a huge tree by considering all characteristic values and concludes the decision principle by pruning. It uses a heuristic technique for pruning depending on the statistical measure of splits. Hence fixing the better rule, the branch nodes send the last class value in the previous node known as the leaf node. Machine learning enables to predict an outcome using data about past events. C5.0 algorithm is used to develop a decision tree for classification. The C5.0 algorithm demonstrated high accuracy (88.99%) in classifying rice pest data using weather attributes (Jinubala and Jeyakumar, 2021). The decision tree is a predictive model that maps decision problems based on a set of conditional decisions. Decision tree takes the form of hierarchical structures similar to trees, with each node representing decisions based on certain features. Decision trees are used to form decision trees with very strong decisions. The decision tree method converts very large facts into decision trees that represent rules that can be easily understood.

The stages in the decision tree are

a. Entropy measures the impurity or randomness in the dataset. Calculate the entropy value of each attribute (proposed by Quinlan, 1993):

$$Entropy\ (S) = -\sum_{i=1}^{n} p_i * log_2 p_i \quad \ldots\ldots [3]$$

where p_i represents the probability of class i in dataset S.

b. Information gain: It measures the reduction in entropy after splitting the dataset based on an attribute.

$$Info\ Gain\ (S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\ (S) \ \ldots\ [4]$$

where represents the subset of dataset *S* after splitting based on attribute *A*.

c. Split information measures how broadly and uniformly the data is split.

$$Split\ info\ (A) = -\sum_{j=1}^{n} \frac{|D_j|}{|D|} * log_2 \frac{|D_j|}{|D|} \quad \ldots[5]$$

d. The gain ratio is used to select the best splitting attribute and is calculated as

$$Gain\ Ratio\ (A) = \frac{Info\ Grain\ (A)}{Split\ info\ (A)} \quad \ldots\ [6]$$

e. The attribute that has the highest profit ratio is selected to be the root (splitting attribute) and the attribute that has a profit ratio value lower than the root (root) is selected to be a branch.

f. Recalculate the value of the profit ratio of each attribute by excluding the attribute selected as root in the previous step.

g. Attributes that have the highest Gain Ratio will be made branches. Then iterations for steps 4 and 5 obtained a value of Gain of zero from all remaining attributes.

Random forest

Random forest is an ensemble learning classifier that consists of multiple decision trees, where each tree is trained on a random subset of the training data and a random subset of features. Each decision tree determines the class of a data object by using a series of attribute-based splits and the final classification is made based on the majority vote from all the trees. The internal nodes of a tree represent decision rules on attributes, while the leaf nodes represent the final predicted class. Random forest improves prediction accuracy and reduces overfitting compared to a single decision tree by averaging multiple models. The performance of a Random Forest model can be evaluated using metrics such as classification error and accuracy. To estimate this performance reliably, resampling methods like bootstrap (e.g. out-of-bag error estimation) and k-fold cross-validation are commonly used helping ensure that the model generalizes well to unseen data (Breiman, 2001).

Algorithm for Construction of Random Forest is:

Step 1: Let the number of training cases be 'n' and let

www.IndianJournals.com
Not for Commercial Use
Downloaded From IP - 216.73.216.57 on dated 17-Nov-2026

the number of variables included in the classifier be 'm'.

Step 2: Let the number of input variables used to make decision at the node of a tree be 'p'. We assume that p is always less than 'm'.

Step 3: Choose a training set for the decision tree by choosing k times with replacement from all 'n' available training cases by taking a bootstrap sample. Bootstrapping estimates the variability of a statistic by resampling the data and assessing how it deviates from its expected value. It is usually used for hypothesis tests. Simple block bootstrap can be used when the data can be divided into non-overlapping blocks. But, moving block bootstrap is used when we divide the data into overlapping blocks where the portion 'k' of overlap between first and second block is always equal to the 'k' overlap between second and third overlap and so on. We use the remaining cases to estimate the error of the tree. Bootstrapping is also used for estimating the properties of the given training data.

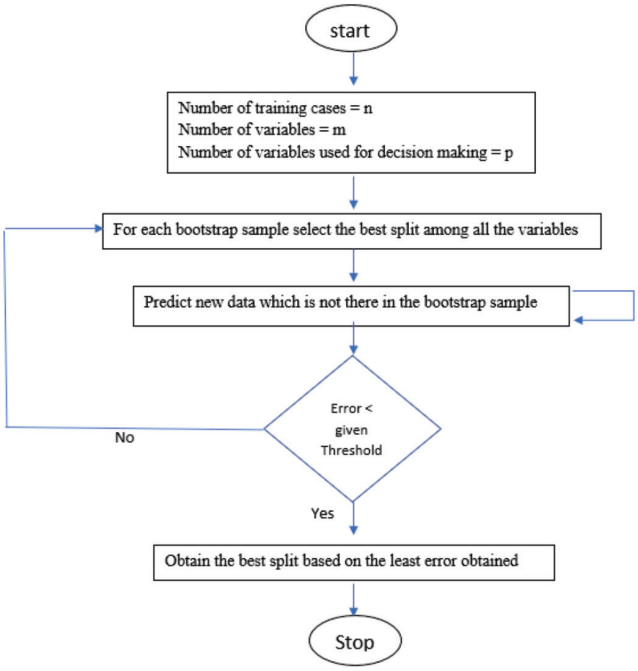


Fig. 1. Diagrammatic representation of Random Forest algorithm

Step 4: For each node of the tree, randomly choose variables on which to search for the best split. New data can be predicted by considering the majority votes in the tree. Predict data which is not in the bootstrap sample and compute the aggregate.

Step 5: Calculate the best split based on these chosen variables in the training set. Base the decision at that node using the best split.

Step 6: Each tree is fully grown and not pruned. Pruning is used to cut of the leaf nodes so that the tree can grow further. Here the tree is completely retained.

Step 7: The best split is one with the least error i.e the least deviation from the observed data set.

Extreme gradient boosting

XGBoost algorithm (Extreme Gradient Boosting) is a powerful and scalable machine learning algorithm used for classification and regression. In gradient tree boosting, models are not trained in isolation of each other, but rather builds trees in succession, where each model iteratively reduces errors made by previous models as described by Chen and Guestrin (2016). At each step, a new decision tree is added to minimize the current error.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad \text{..... [7]}$$

XGBoost uses the gradient (first derivative) and the Hessian (second derivative) of the loss function to approximate improvements efficiently. The algorithm approximates the loss function using Taylor expansion (2nd order):

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_i) \quad \text{..... [8]}$$

where,

$$g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)(gradient)$$

$$h_i = \partial^2{}_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)(Hessian)$$

XGBoost greedily builds a tree by choosing splits that maximize the Gain, defined using gradients and Hessians.

The gain from a split is:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \quad \text{..... [9]}$$

where,

G_L, H_L = sum of gradients and Hessaians for the left child

G_R, H_R = sum of gradients and Hessaians for the right child

Υ = regularization term for the number of leaves.

Regularization parameters Υ and λ control complexity where Υ represents minimum loss reduction to make a further partition, λ represents L2 regularization on weights. After each tree is built, its contribution is scaled down by a factor η (learning rate):

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \quad \text{..... [10]}$$

The learning objective combines two parts: 1. Loss Function (e.g., Mean Squared Error for regression, Log Loss for classification) 2. Regularization Term: Penalizes model complexity (to prevent overfitting)

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad \text{.....[11]}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \quad \text{..... [12]}$$

www.IndianJournals.com
Not for Commercial Use
Downloaded From IP - 216.73.216.57 on dated 17-Nov-2026

Where is the loss function to find the difference between the predicted value and the target value Ω represents the complexity of the model. It is a regularization term that has the function of smoothening the weight to avoid overfitting. Like in Random Forest, XGBoost can randomly sample features at each tree or split to prevent overfitting and improve speed.

Confusion matrix

Classification performance for each method was evaluated using a confusion matrix to calculate accuracy, precision and recall. The confusion matrix was employed to assess classification performance across all classes and the results are presented in Table 1.

Table 1. Confusion matrix of modelled predictions

	Actual values	
Predicted values	**Positive**	**Negative**
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True negative (TN)

Several matrices can be derived from confusion matrix, namely accuracy, sensitivity (recall), specificity, precision and F_1-score (Fawcett, 2006). The following formulae were used:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad \ldots.[13]$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP+FN} \quad \ldots.[14]$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad \ldots.[15]$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad \ldots.[16]$$

$$F_1\text{-Score} = 2^* \frac{Precision*Recall}{Precision+Recall} \quad \ldots[17]$$

Kappa criterion

The Kappa criterion (also called Cohen's Kappa) is a statistical measure that evaluates the agreement between predicted and actual classifications, adjusted for chance agreement (Cohen, 1960). It is especially useful when the data is imbalanced across classes. Accuracy can be misleading when classes are imbalanced. For example, if 90% of data belongs to one class, a model that predicts only that class will have 90% accuracy, but it's actually not useful. Kappa corrects for this by penalizing high accuracy that arises due to chance or imbalance.

Table 2. Strength of agreement based on Kappa statistic

Kappa value	Strength of agreement
< 0	Less than chance
0.01–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Pre-processing technique:

Min-Max Normalization (Min-Max scaling)

Min-max normalization is a data pre-processing technique used in machine learning and statistical analysis to rescale numerical variables to a fixed range, usually between 0 and 1 (Ahsan *et al.,* 2021). This transformation ensures that all variables contribute equally to the model and prevents variables with larger values from dominating those with smaller values.

The transformation is performed using the following formula:

$$X_{scaled} = \frac{X-X_{min}}{X_{max}-X_{min}} \quad \ldots\ldots[19]$$

where,

X is original value.

X_{min} the minimum value of the feature.

X_{max} maximum value of the feature.

In this method, the minimum value of the variable becomes 0 and the maximum value becomes 1 while all other values are scaled proportionally between these limits. This scaling preserves the relative relationships among the original data values while ensuring they fall within a consistent range.

RESULTS AND DISCUSSION

Descriptive statistics

The descriptive statistics of rice leaf folder incidence along with the associated weather variables namely maximum temperature (T_{max}), minimum temperature (T_{min}), morning relative humidity (RHM), evening relative humidity (RHE), and rainfall (RF) based on 390 weekly observations is presented in table 3. The average leaf folder incidence was 70.0 with a high standard deviation of 142.0 indicating large variability in pest population over the study period. Among the meteorological variables, the average values were 30.2°C for T_{max}, 23.2°C for T_{min}, 83.0% for RHM, 72.3% for RHE and 22.4 mm for rainfall. The coefficient of variation was highest for leaf folder (202.8%) and rainfall (179.7%) suggesting highly fluctuating patterns while relative humidity variables exhibited comparatively lower variability (Kewal *et al.,* 2024). The positive skewness values for leaf folder (3.8) and rainfall (4.0) indicate right-skewed distributions with occasional extreme values, whereas T_{max} and RHM showed negative skewness. High kurtosis values for leaf folder (17.2) and rainfall (23.2) suggest the presence of

www.IndianJournals.com
Not for Commercial Use
Downloaded From IP - 216.73.216.57 on dated 17-Nov-2026

Table 3. Descriptive statistics of rice leaf folder incidence and weather variables

Descriptive statistic	Leaf folder	T_{max} (°C)	T_{min} (°C)	RHM (%)	RHE (%)	RF (mm)
Mean	70.0	30.2	23.2	83.0	72.3	22.4
Standard Deviation	142.0	4.2	4.4	6.4	8.7	40.2
Kurtosis	17.2	2.1	-0.1	2.8	0.0	23.2
Skewness	3.8	-1.4	0.6	-1.3	-0.2	4.0
Range	1084.0	24.0	21.0	43.4	44.9	363.8
Minimum	0.0	14.9	14.6	56.0	48.1	0.0
Maximum	1084.0	38.9	35.6	99.4	93.0	363.8
Count	390.0	390.0	390.0	390.0	390.0	390.0
CV %	202.8	13.8	19.0	7.8	12.0	179.7

extreme peaks and heavy tails in their distributions.

The seasonal population dynamics of rice leaf folder recorded through light trap catches from 2011 to 2024 across Standard Meteorological Weeks (SMW) (Fig. 2). The pest population remained very low during the early weeks of the year (SMW 1-30). A gradual increase was observed from around 35 SMW. The population reached its peak between SMW 40 and SMW 48 in most years which indicated the major infestation period. However, the magnitude of peak populations varied among years. After SMW 50, the population declined sharply (Ram *et al.,* 2014).

Multinomial logistic regression for rice leaf folder pest population

Multinomial logistic regression was used to classify rice leaf folder infestation levels into three categories (Low, Medium and High) based on weather parameters including maximum temperature (T_{max}), minimum temperature (T_{min}), morning relative humidity (RHM), evening relative humidity (RHE) and rainfall (RF) during the study period and low infestation was taken as the reference category (Agresti, 2013). Prior to model development, the predictor variables were standardized to remove scale differences. The multinomial logistic regression model was implemented in the R statistical environment using the 'nnet' package.

From the figure 3, the T_{max} effect plot shows that as the maximum temperature (T_{max}) increases, the probability of the Low LF category (blue line) steadily rises, while the probabilities of the Medium (orange line) categories decline. This indicates that higher maximum temperatures are associated with a greater likelihood of observing a lower leaf folder incidence (Saikia *et al*., 2019). From the table 4 it is evident that T_{max} is the strongest and most consistent predictor significantly reducing the likelihood of 'Medium' categories compared to 'Low'. As RHE increases, the probability of the Low LF category (blue line) decreases gradually, while the probability of the medium category (orange line) steadily increases. The probability of the High category (green line) remains low and nearly constant across the RHE range. From the table 4, RHE has some significance for the 'Medium' category. The other variables do not

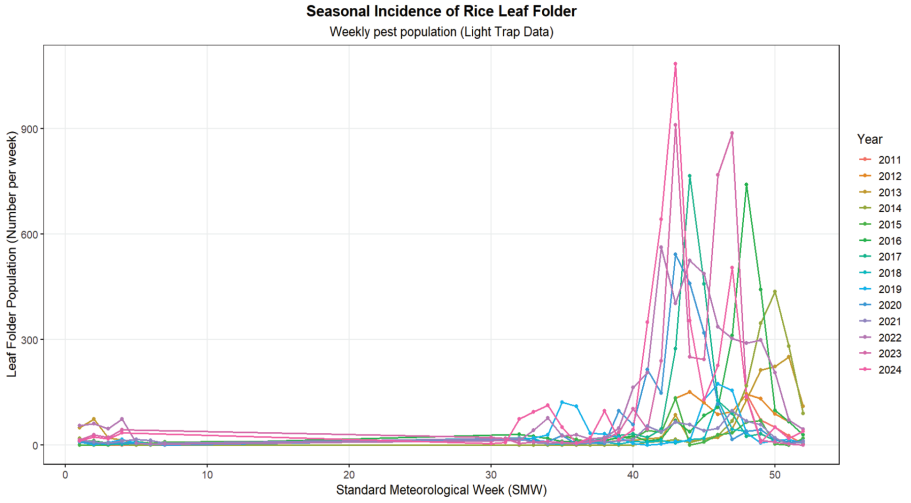


Fig. 2. Seasonal incidence of rice leaf folder based on weekly light trap data (2011-2024)

www.IndianJournals.com
Not for Commercial Use
Downloaded From IP - 216.73.216.57 on dated 17-Nov-2026

Table 4. Multinomial logistic regression results for comparing Medium and High leaf folder categories against the Low reference group

	Class	Intercept	T_{max}	T_{min}	RHM	RHE	RF
Coefficient	Medium	-1.18	-0.29	0.002	0.218	0.3874	-0.126
	High	-3.62	-0.402	0.425	0.306	-0.1802	-0.629
Standard error	Medium	0.126	0.1429	0.144	0.139	0.14	0.135
	High	0.413	0.365	0.394	0.51	0.382	0.682
p-value	Medium	0	**0.03**	0.98	0.11	**0.01**	0.35
	High	0	0.27	0.28	0.35	0.54	0.63

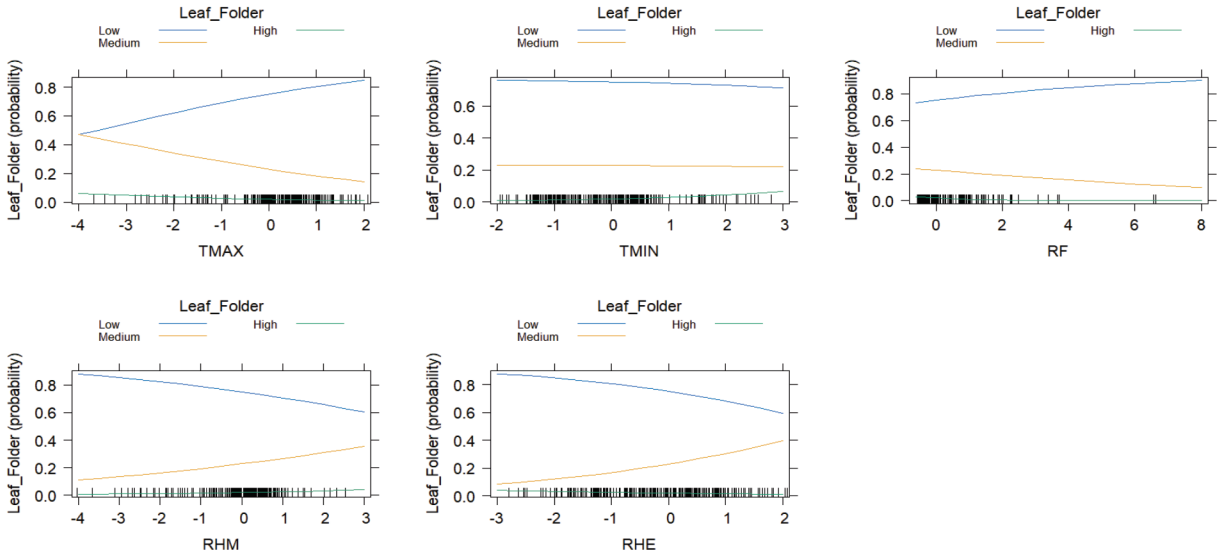


Fig. 3. The effect plots of predictors (T_{max}, T_{min}, RHM, RHE, RF) for predicting categorical leaf folder levels

www.IndianJournals.com
Not for Commercial Use
Downloaded From IP - 216.73.216.57 on dated 17-Nov-2026

show statistically significant effects which means their influence on LF category classification is uncertain in this model.

Table 5. Confusion Matrix for rice leaf folder using Multinomial Logistic Regression

	Actual		
Predicted	**Low**	**Medium**	**High**
Low	282	90	10
Medium	4	3	1
High	0	0	0

Table 6. Sensitivity and Specificity for different classes of rice leaf folder using Multinomial Logistic Regression

Metric	Low	Medium	High
Sensitivity	0.9860	0.0323	0.0000
Specificity	0.0385	0.9832	1.0000
Precision	0.7382	0.3750	0.0000
F_1 Score	0.8448	0.059	0.0000

The overall accuracy measured is 73.08 % and Kappa criterion is 0.0274, based on actual and predicted class values (Table 5). The sensitivity 98.6% of actual low values were correctly classified and specificity (3%) of many 'Medium' or 'High' class were wrongly classified as low (Table 6). Similarly, sensitivity (3%) of actual 'Medium' values were correctly predicted and specificity of 98.3% of many 'Low' or 'High' class were wrongly classified as Medium. The sensitivity is zero because no actual 'High' values were correctly predicted. However, the specificity of 100% indicates perfect performance in identifying non-'High' cases which means no other values were incorrectly classified as 'High'. For multiclass evaluation, the F_1 score provides a more balanced measure of performance. An F_1 score of 0 indicates that the model fails to correctly predict the 'High' class, whereas a score close to 1, such as 0.844, indicates strong predictive accuracy for the 'Low' class (Witten *et al.*, 2016).

Decision tree for rice leaf folder pest population

The C5.0 decision tree algorithm was employed to classify rice leaf folder infestation levels based on weather variables using the C5.0 package (Kuhn *et al.*, 2014). The infestation data were categorized into three

classes: Low, Medium and High. Five meteorological variables, T_{max}, T_{min}, RHM, RHE, and RF, were selected as predictor variables. As these variables were measured on different scales, Min-Max normalization was applied to rescale them to a range between 0 and 1. The normalized predictors were then used to construct the C5.0 decision tree model for classifying the severity of leaf folder infestation. The algorithm uses the information gain ratio to select the attribute that best splits the data. Based on the selected attribute, the dataset is partitioned into subsets and this process is recursively repeated for each subset. Tree growth continues until all observations within a node belong to the same class or no further information gain can be achieved at which point the stopping criterion is met.

Figure 4 presents the C5.0 decision tree constructed to predict leaf folder pest incidence across standard meteorological weeks during the period from 2011 to 2024 at ARS, Bapatla. The pest data were categorized into three classes: Low, Medium and High. The decision tree analysis reveals that temperature variables (T_{max} and T_{min}) play a dominant role in the initial splits, while relative humidity (RHE) and rainfall (RF) contribute to further refinement of the classification at deeper levels of the tree. Most terminal nodes are associated with the 'Low' category which indicates that the majority of observations fall within this class.

The overall classification accuracy of the model was 83.08%, with a Kappa coefficient of 0.5171, as calculated from the actual and predicted class values presented in table 7. According to table 8, the sensitivity for the 'Low' class was 96.5%, indicating that most actual 'Low' values were correctly classified. However, the specificity was 53.8%, which suggests that a considerable proportion of 'Medium' and 'High' cases were incorrectly classified as 'Low'.

For the 'Medium' class, the sensitivity was 51.6% which means that only about half of the actual 'Medium' cases were correctly predicted, while the specificity was 93.9%, which indicates that most 'Low' and 'High' cases were correctly identified as not belonging to the 'Medium' class. In contrast, the model failed to correctly predict any instances of the 'High' class resulting in a sensitivity of 0, although the specificity was 100%. This means that no other classes were misclassified as 'High'.

For multiclass evaluation, the F_1 score provides a more balanced measure of model performance. An F_1 score of 0 indicates that the model has failed to correctly predict the 'High' class, whereas a score close to 1, such as 0.904, which indicates strong predictive performance for the 'Low' class (Witten *et al.*, 2016).

Table 7. Confusion Matrix for rice leaf folder using C5.0 algorithm

	Actual Low	Actual Medium	Actual High
Predicted Low	276	45	3
Predicted Medium	10	48	8
Predicted High	0	0	0

Table 8. Sensitivity and Specificity for different classes of rice leaf folder using C5.0 algorithm

Metric	Low	Medium	High
Sensitivity	0.965	0.516	0
Specificity	0.538	0.939	1
Precision	0.851	0.727	0
F_1 Score	0.904	0.604	0

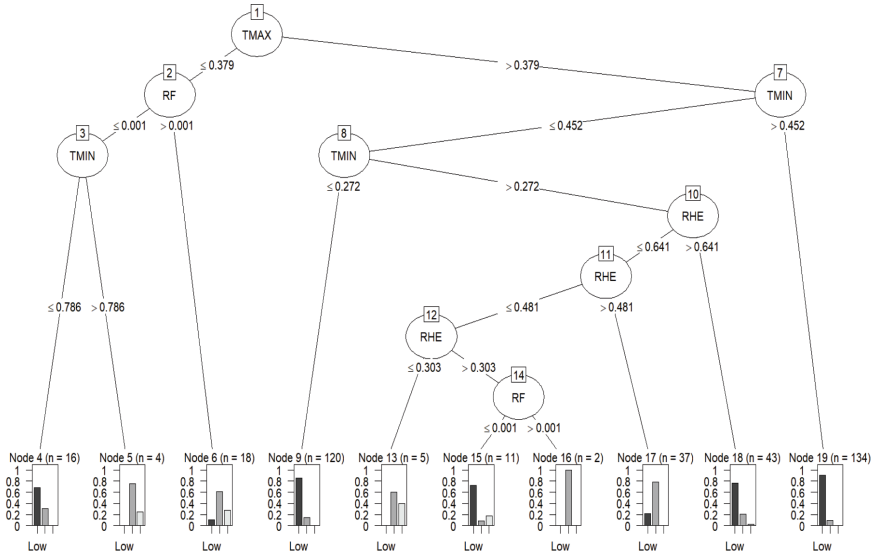


Fig. 4. Decision tree using C5.0 algorithm for leaf folder pest population at ARS, Bapatla

www.IndianJournals.com
Not for Commercial Use
Downloaded From IP - 216.73.216.57 on dated 17-Nov-2026

Random Forest for rice leaf folder population

As per methodology, Random Forest classification model is used for the prediction of Leaf folder pest count of standard meteorological weeks by using weather variables during the study period. Before the predictor variables were first normalized using Min-Max scaling to standardize their ranges between 0 and 1. The response variable representing leaf folder infestation was converted from a continuous measure into three categorical classes namely Low, Medium and High based on predefined threshold values. A Random Forest classification model (Breiman, 2001) was then developed and the model was fine-tuned by optimizing the number of trees (ntree = 500) and the number of variables considered at each split (mtry=2) to improve predictive accuracy and model performance. The Out-of-Bag (OOB) error estimate of the Random Forest model was 23.08%, indicating that approximately 76.92% of the observations were correctly classified which suggests a moderate to good predictive performance of the model (Liaw and Wiener, 2002). From the figure 5, the variable importance plot from the Random Forest model shows that minimum temperature (T_{min}) is the most influential predictor, followed by maximum temperature (T_{max}) and evening relative humidity (RHE). As per mean decrease in accuracy, rainfall has a moderate contribution to the model, whereas morning relative humidity (RHM) shows the least importance. Both Mean Decrease in Accuracy and Mean Decrease in Gini measures indicate that temperature variables play a dominant role in predicting the response variable.

The overall accuracy measured is 96.67 % and Kappa criterion is 0.9155 based on actual and predicted class values (Table 9). The sensitivity 99.6% of actual low values were correctly classified and high specificity of 93.2% rarely misclassifies other classes as low (Table 10). Similarly, the sensitivity for the 'Medium' class was 92.4% which indicates that most actual Medium values were correctly predicted, while the specificity of 97.9% shows that non-Medium values were accurately identified. According to Table 10, the sensitivity for the High class was 54.5% which means that just over half of the actual 'High' values were correctly classified. The specificity for this class was 100% which indicates that no other classes were incorrectly classified as 'High'. For multiclass evaluation, the F_1 score provides a reliable measure of model performance. F_1 scores close to 1, such as 0.986 and 0.928, indicate strong predictive accuracy for the 'Low' and 'Medium' classes, respectively (Witten *et al.*, 2016).

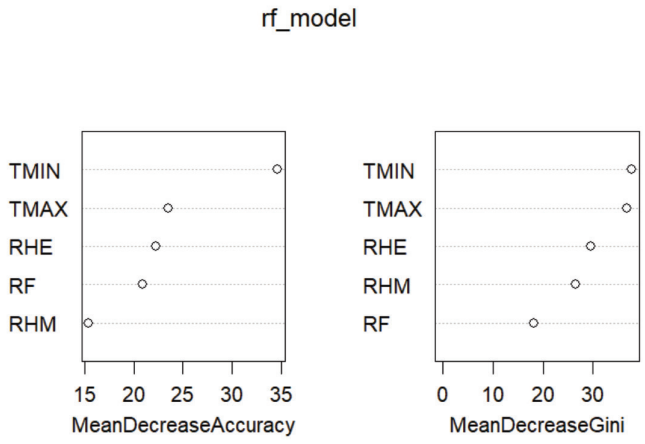


Fig. 5. Variable importance using Random Forest on leaf folder

Table 9. Confusion matrix for rice leaf folder using random forest

Predicted	Actual		
	Low	Medium	High
Low	285	7	0
Medium	1	86	5
High	0	0	6

Table 10. Sensitivity and Specificity for different classes of rice leaf folder using random forest

Metric	Low	Medium	High
Sensitivity	0.996	0.924	0.545
Specificity	0.932	0.979	1.000
Precision	0.976	0.934	1.000
F_1 Score	0.986	0.928	0.706

XGBoost algorithm for rice leaf folder population

In present study, the Extreme Gradient Boosting (XGBoost) algorithm was used to classify rice leaf folder pest incidence based on weather variables. The response variable was then converted into factor and numeric format which is required for the XGBoost algorithm. The predictor variables used in the model were maximum temperature (T_{max}), minimum temperature (T_{min}), morning relative humidity (RHM), evening relative humidity (RHE) and rainfall (RF). These variables were normalized using Min-Max normalization to scale the values between 0 and 1 and improve model performance. The XGBoost model was trained using the multi-class softmax objective function with parameters such as maximum tree depth = 4, learning rate (eta) = 0.1 and 100 boosting rounds (Chen *et al.*, 2015). The figure 6, displays variable importance using this algorithm on leaf folder population. It tells us how important each feature was in the structure of the boosted decision trees within the model. It shows that T_{min} has a higher relative importance followed by T_{max} (Sumido *et al.*, 2023).

The overall accuracy measured is 93.59 % and

www.IndianJournals.com
Not for Commercial Use
Downloaded From IP - 216.73.216.57 on dated 17-Nov-2026

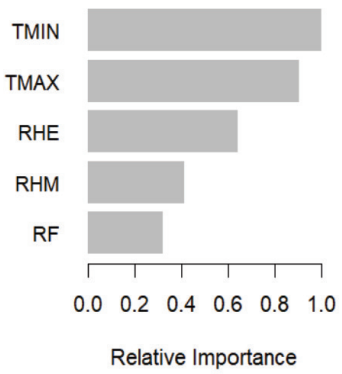


Fig. 6. Variable importance using XGBOOST for leaf folder

Kappa criterion is 0.832, based on actual and predicted class values (Table 11). The sensitivity 99.3% of actual low values were correctly classified and specificity high of 82.6 % rarely misclassifies other classes as low (Table 12). Similarly, the sensitivity for the 'Medium' class was 80.6% which indicates that a majority of the actual Medium values were correctly predicted, while the specificity of 97.64% shows that non-Medium values were accurately identified. The sensitivity for the 'High' class was 54.54% (Table 12). It means that just over half of the actual High values were correctly classified. The specificity of 100% indicates near-perfect performance, with no other classes misclassified as 'High'. For multiclass evaluation, the F_1 score provides a balanced measure of performance. An F_1 score close to 1, such as 0.966, indicates strong predictive accuracy for the 'Low' class (Witten *et al.*, 2016).

Table 11. Confusion matrix for rice leaf folder using XGBOOST

Predicted	Actual		
	Low	Medium	High
Low	284	18	0
Medium	2	75	5
High	0	0	6

Table 12. Sensitivity and Specificity for different classes of rice leaf folder using XGBOOST

Metric	Low	Medium	High
Sensitivity	0.993	0.8065	0.5454
Specificity	0.826	0.9764	1.0
Precision	0.9404	0.9146	1.0
F1 Score	0.966	0.857	0.706

Performance analysis

The performance of different machine learning algorithms for predicting rice leaf folder pest incidence is presented in Table 13. Among the models, the Random Forest algorithm showed the best performance (Orounla *et al.*, 2023) with the highest F_1 scores for Low (0.986) and Medium (0.928) classes, along with the highest overall accuracy (0.9667) and Kappa coefficient (0.9155). The XGBoost model also demonstrated strong performance with an accuracy of 0.9359 and Kappa value of 0.832. In contrast, the C5.0 algorithm showed moderate performance, while the Multinomial Logistic Regression model recorded the lowest accuracy (0.7308) and Kappa coefficient (0.027) indicating poor classification performance particularly for the Medium and High incidence classes. Overall, the Random Forest algorithm outperformed the other models in predicting rice leaf folder pest incidence.

Table 13. Performance comparison of Machine Learning Algorithms using F_1 Score, Accuracy and Kappa Metrics for rice leaf folder pest incidence

F_1 Score	Low	Medium	High	Accuracy	Kappa criterion
Random Forest algorithm	0.986	0.928	0.706	0.9667	0.9155
XGBoost	0.966	0.857	0.706	0.9359	0.8320
C5.0 algorithm	0.904	0.604	0.000	0.8308	0.5171
Multinomial Logistic Regression	0.844	0.059	0.0000	0.7308	0.0274

CONCLUSION

The Machine Learning Algorithms, namely Multinomial Logistic Regression, C5.0, Random Forest and XGBoost were used to analyse pest incidence based on weather data of standard meteorological weeks during the period from Jan 2011 to Dec 2024 at ARS, Bapatla. The study revealed that rice leaf folder incidence showed considerable variability across the cropping season with the highest infestation occurring during SMW 40-48. Weather variables played an important role in influencing pest incidence, particularly temperature and relative humidity. Among the models evaluated, the multinomial logistic regression model achieved moderate accuracy (73.08%) but showed poor prediction for Medium and High infestation classes. The C5.0 decision tree improved classification performance with an accuracy of 83.08%, though it still struggled to identify High infestation levels. The XGBoost model provided higher predictive accuracy (93.59%) with strong agreement between predicted and observed classes. Among all models, the Random Forest model performed the best, achieving the highest accuracy (96.67%) and Kappa value (0.9155), indicating excellent predictive capability. Therefore, Random Forest was found to be the most suitable model for predicting rice leaf folder infestation using weather variables. The results of this study can be used

www.IndianJournals.com
Not for Commercial Use
Downloaded From IP - 216.73.216.57 on dated 17-Nov-2026

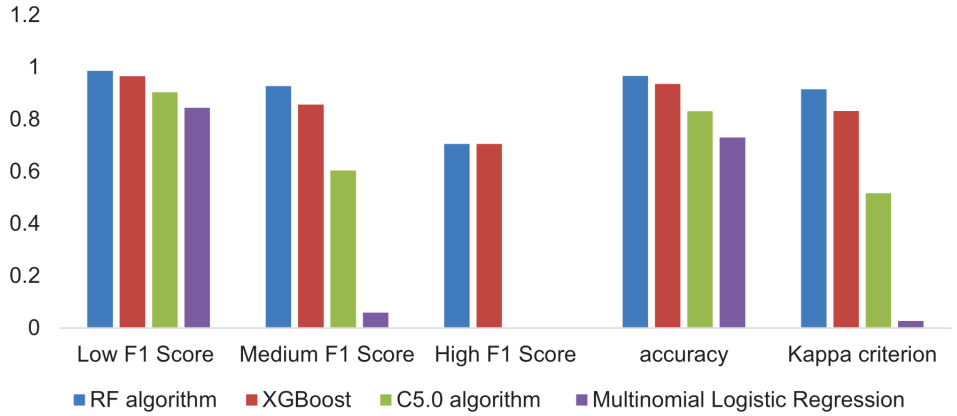


Fig. 7. Performance comparison of Machine Learning Algorithms using F_1 Score, Accuracy and Kappa Metrics for rice leaf folder pest incidence

www.IndianJournals.com
Not for Commercial Use
Downloaded From IP - 216.73.216.57 on dated 17-Nov-2026

by farmers, agriculturists and developers as a basis for developing Decision Support Systems and Integrated Pest Management Systems.

LITERATURE CITED

Agresti A 2013. *Categorical data analysis*. 3rd edn., Hoboken: John Wiley & Sons.

Ahmed H, R B Khan, Sharma D, Jamwal V V S and Gupta S 2010. Seasonal incidence, infestation and trap catches of *Cnaphalocrocis medinalis* (Guenee) in rice. *Annals Pl Prot Sci* **18**(2): 382-83.

Ahmed N, Holmquist B, Nasrin S, Ali M P, Bari M N, Begum M A, Afsana N and Rabbi M F 2013. Impact of climate change on rice insect pests and their natural enemies. **In**: *International Conference on Climate Change Impact and Adaption.* pp. 15-17.

Ahsan M M, Mahmud M P, Saha P K, Gupta K.D and Siddique Z 2021. Effect of data scaling methods on machine learning algorithms and model performance. *Technologies* **9**(3): 52.

Ahsan M M, Parvez Mahmud M A, Saha P K, Gupta K D and Siddique Z 2021. Effect of data scaling methods on machine learning algorithms and model performance. *Technol* **9**(3): 52.

Awmack C S and Leather S R 2002. Host plant quality and fecundity in herbivorous insects. *Annu Rev Entomol* **47**: 817-44.

Bale J S, Masters G J, Hodkinson I D, Awmack C *et al.* 2002. Climate change and the dynamics of crop pests. *Global Change Biol* **8**(1): 1-16.

Breiman L 2001. Random forests. *Machine Learning* **45**(1): 5-32.

Chen T and Guestrin C 2016. XGBoost: A scalable tree boosting system. **In**: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785-94.

Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I and Zhou T 2015. Xgboost: extreme gradient boosting. R package version 0.4-2. **1**(4): pp.1-4.

Cohen J 1960. A coefficient of agreement for nominal scales. *Educ Psychol Meas* **20**(1): 37-46.

Fawcett T 2006. An introduction to ROC analysis. *Pattern Recognit Lett* **27**(8): 861-74.

Hastie T, Tibshirani R and Friedman J 2009. The elements of statistical learning: Data mining, inference and prediction, Springer, New York.

Hosmer D W, Lemeshow S and Sturdivant R X 2013. Applied logistic regression. 3rd edn., Wiley, New York.

Hussain S, Memon F, Nighat A, Memon F A and Hussain M 2022. Weather based localized crop prediction using machine learning. *Quest Res J* **20**(2): 106-15.

Jeong H K, Kim C G and Moon D H 2014. An analysis of impacts of climate change on rice damage occurrence by insect pests and disease. *Korean J Environ Agric* **33**(1): 52-56.

Jinubala V and Jeyakumar P 2021. Classification of level of rice pest incidence based on weather information. *Int J Environ Clim Chang* **11**(2): 32-43.

Kewal R, Agrawal N, Yadav V K, Kripashankar and Dixit R 2024. Seasonal incidence of rice leaf folder *Cnaphalocrocis medinalis* (Guenee) in rice at Kanpur, Uttar Pradesh. *Int J Adv Biochem Res* **8**(10S): 225-31.

Khan Z H and Ramamurthy V V 2004. Influence of weather factors on the activity of rice leaf folder, *Cnaphalocrocis medinalis* (Guenee). *Ann Plant Prot Sci* **12**(2): 263-66.

Kuhn M, Weston S, Coulter N and Quinlan R 2014. C50: C5. 0 decision trees and rule-based models. R package version 0.1. 0-21, URL http://CRAN. R-project. org/package C, 50.

Liaw A and Wiener M 2002. Classification and regression by randomForest. *R news* **2**(3): 18-22.

Orounla B R, Sode A I, Salako K V and Kakai R G 2023. Empirical performance of CART, C5. 0 and Random Forest classification algorithms for Decision Trees. *African J Appl Stat* **10**(1): 1399-1418.

Pathak M D and Dhaliwal G S 1981. Trends and strategies for rice insect problems in Tropical Asia, International Rice Research Institute, Laguna, Philippines. https://infosat.icrisat.org/8592/

Quinlan J R 1993. C4.5: Programs for machine learning. Morgan Kaufmann Publishers, San Mateo, California.

Ram M, Sachan S K and Singh G 2014. Study on population build-up of rice leaf folder *Cnaphalocrocis medinalis* (Guenee) in relation to weather factors. *Int J Adv Res* **2**(10): 75-77.

Reddy B N K, Rathod S, Kallakuri S, Sridhar Y, Admala M, Malathi S, Pandit P and Jyostna B 2022. Modelling the relationship between weather variables and rice yellow stem borer population: A count data modelling approach. *Int J Environ Clim Chang* **12**(11): 3623-32.

Saikia R, Saikia D K, Islam A N and Devee A 2019. Seasonal population build-up of leaf folder (*Cnaphalocrocis medinalis* Guenee) in paddy in response to meteorological parameters. *J Agrometeorol* **21**(2): 230-32.

Sharma H C 2010. Effect of climate change on insect pests of crops and their management. *J Crop Improv* **24**(4): 304–22.

Sumido E C, Feliscuzo L S and Aliac C J G 2023. Pest classification and prediction: Analyzing the impact of weather to pest occurrence through machine learning. *J Eng Sci Technol* **18**: 124-38.

Witten I H, Frank E, Hall M A and Pal C J 2016. Data mining: Practical machine learning tools and techniques. 4th edn., Morgan Kaufmann, Burlington.

www.IndianJournals.com
Not for Commercial Use
Downloaded From IP - 216.73.216.57 on dated 17-Nov-2026