Innovative Approaches to Agricultural Risk with Machine Learning

Sumi. M[1]*, S. Manju Priya[2]
Research Scholar, Department of Computer Science and Engineering,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India[1]
Professor, Department of Computer Science and Engineering,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India[2]

***Abstract*—Agriculture is fraught with uncertainties arising from factors like weather volatility, pest outbreaks, market fluctuations, and technological advancements, posing significant challenges to farmers. By gaining insights into these risks, farmers can enhance decision-making, adopt proactive measures, and optimize resource allocation to minimize negative impacts and maximize productivity. The research introduces an innovative approach to risk prediction, highlighting its pivotal role in improving agricultural practices. Through meticulous analysis and optimization of a farmer dataset, employing pre-processing techniques, the study ensures the reliability of predictive models built on high-quality data. Utilizing Variation Inflation Factor (VIF) for feature selection, the study identifies influential features critical for accurate risk classification. Employing techniques like KNN, Random Forest, logistic regression, SVM, Ridge classifier, Gradient Boosting and XGBoost, the study achieves promising results. Among them KNN, random forest, Gradient Boosting and XGBoost scored with high accuracy of 88.46%. This underscores the effectiveness of the proposed methodology in providing actionable insights into potential risks faced by farmers, enabling informed decision-making and risk mitigation strategies.**

Keywords*—Random forest; ridge classifier; logistic regression; gradient boosting; extreme gradient boost; Variation Inflation Factor; support vector machine; farmer risk prediction; agricultural risk

I. Introduction

Agriculture is a vital sector of any country; therefore, the growth and development of a country directly depend on agriculture. Agriculture is, not just only a means of subsistence or income, it's a way of living life for the human species [1]. Agriculture is the key source of food, forage, and energy and serves as the cornerstone of the economic growth of any country. Agriculture is the key source of food, forage, and energy and serves as the cornerstone of the economic growth of any country [2]. In the current Indian era, agriculture still plays a significant role in the lives of more than 80% of Indians who are directly or indirectly involved in farming activities. According to the census of India 2021, the agricultural sector of India employed 54.6 % of the total workers. The agriculture sector and allied sector provide 17.8 % of the nation's Gross Value Added [3].

Agriculture is one of the risky professions with uncertain outcomes and a variety of risks are faced by Indian farmers over the whole growing season. The World Bank defines "Agricultural risk as a combination of the possibility of a hazardous event or exposure and the severity of the losses that can be caused by the event or exposure" [4]. One of the most vital agricultural risks is the production or biological risk, which is mostly brought on by climate variability and is getting worse every day as a result of climate change [5]. However, many other factors such as financial, legal, marketing, technological, social, and human personal factors can contribute to agricultural risk in addition to this climate change effect and farmers have to deal with all the risk sources. For instance, events like insect pest attacks [6], bad quality of inputs, epidemics [7], volatile prices, and unavailability of inputs can also decrease the production as well as income of Indian farmers. Therefore, based on these risk components, agricultural risk can be broadly classified as economic, production, technological, institutional, and personal risk. Risk is classified into five categories viz., Economic risk, Production risk, Technological risk, Institutional risk, and Personal risk [8].

The main contributions of this study can be outlined as follows:

- Develop predictive models for farmers' risk prediction using machine learning (ML) techniques.
- Optimize feature selection through Variation Inflation Factor (VIF) analysis to enhance the accuracy of risk prediction.
- Evaluate the performance of various classifiers, including KNN, Random Forest, SVM, Ridge classifier, logistic regression, Gradient Boosting, and XGBoost, in predicting farmers' risk levels.

The rest of the paper is organized as follows: In Section II, a summary of literature is provided, highlighting areas that indicate a need for more investigation. In Section III, the methodology is explained in depth. Section IV goes into great detail about the results that the suggested strategy produced. A discussion is provided in Section V and finally, a summary of the findings is included in Section VI, which gives a conclusion to the paper.

II. Literature Review

Jinger et al. [9] introduced a fuzzy model designed for forecasting maize crop yields. They evaluated maize production by incorporating parameters such as temperature,

humidity, rainfall during different growth stages, and the sowing area. Upadhya et al. [10] proposed, fuzzy logic-based crop yield estimation, considering temperature, humidity, and soil moisture as input parameters. The parameters were subjected to fuzzy arithmetic, resulting in obtaining crisp values of yield. Trapezoidal membership functions were considered in the fuzzy modeling. Pandhe et al. [11] suggested a model, it was determined that if farmers were aware of the yield potential of the crops, they are planting beforehand, they would opt for crops with higher expected yields based on the climate of the region. With an accuracy of 87%, assessed through a 10-fold cross-validation technique, indicating a strong correlation between climate factors and crop yield.

Kalimuthu et al. [12] aid beginner farmers by providing guidance on suitable crop choices through the utilization of machine learning, advanced technology in crop prediction. The Naive Bayes algorithm, a supervised learning technique, was employed to achieve this objective. The approach involves the development of a supervised ML model using the naive Bayes Gaussian classifier with a boosting algorithm to predict crops with high accuracy. Consequently, the predicted crop seed serves as the output for the given input parameters. Mulla et al. [13] centered on exploring the prediction of crop yield and cost estimation. The methodology proposed employs tree algorithms to efficiently predict the outcomes. The study primarily encompasses several key implementation modules, including data acquisition, data exploration, prediction, and the development of a web application. Mohanty et al. [14] describe four functional components, which include predicting crop yield, predicting demand, determining supply and forecasting crop prices. The input datasets consist of a range of field values, demand, and remaining crop at year-end, encompassing yield, import and crop prices. Rani et al. [15] proposed a model for estimating commodity prices. By using techniques like Linear Regression, Random Forest, and Decision Trees. The model's successful application of decision trees, random forests, and linear regression suggests an appropriate estimation.

Chen et al. [16] investigated the complexities and challenges in agri-food supply chains (ASCs), highlighting the need for effective traceability and management. They designed a blockchain-based ASC framework to ensure decentralized security and traceability of agri-food products. Additionally, they proposed a Deep Reinforcement Learning-based Supply Chain Management (DR-SCM) method to optimize production and storage decisions for increased profits. Extensive simulations demonstrated the framework's reliability in maintaining secure, consistent, and unique tracing data. Moreover, the DR-SCM method consistently outperformed heuristic and Q-learning methods in various scenarios, achieving higher profits and exhibiting greater adaptability. The study concluded that integrating blockchain with DR-SCM significantly enhances traceability and profitability in ASCs, paving the way for further research on advanced algorithms in more complex environments. Rakhra et al. [17] aimed to address the myriad challenges encountered by farmers in accessing tool and equipment, as well as to ascertain their keen interest in equipment rental and sharing processes. Farmers were categorized into three groups—small, moderate, and large—based on the findings of the survey. To gain a deeper insight into the target variables, the dataset underwent training and testing splits. Standardization of the survey dataset was performed to ensure clarity and remove ambiguity.

Chelliah et al. [18] is grounded in satellite imagery and utilizes ML algorithms to achieve an accuracy enhancement. This paper introduces a target prediction algorithm aimed at guiding farmers regarding market target products and fostering improved relationships between farmers and bankers through centralized information about recent government plans. Additionally, a ML algorithm for crop prediction is proposed to augment agricultural revenue. The proposed model holds relevance for real-world research, facilitating the assessment of the acceptability of the financial forms detailed in this study.

Existing studies have explored various risk factors and modelling approaches, but there remains a lack of comprehensive frameworks that effectively integrate diverse data sources and advanced analytical techniques to provide actionable insights for farmers. Additionally, the majority of current research focuses on individual risk factors or employs simplistic modeling techniques, neglecting the multifaceted nature of agricultural risks and the potential interactions between different risk factors. Addressing this gap requires the development of sophisticated predictive models that leverage advanced ML algorithms, incorporate diverse data streams, and account for the dynamic and interconnected nature of agricultural systems. Such models have the potential to significantly enhance farmers' ability to anticipate, mitigate, and adapt to various risks, thereby improving agricultural sustainability, resilience, and productivity.

III. Materials and Methods

The study initiates the collection of a comprehensive farmer dataset, comprising diverse variables such as weather conditions, pest prevalence, disease outbreaks, input and product prices, technology adoption rates, and insurance coverage. Following dataset collection, a rigorous pre-processing phase, which includes tasks such as handling outliers, correlation finding, and encoding to ensure the dataset's quality and suitability for predictive modelling. Subsequently, the Variation Inflation Factor (VIF) technique [19] is employed to select the most influential features from the dataset, facilitating accurate risk classification. Various Machine Learning techniques, including K-Nearest Neighbor [20], Random Forest [21], logistic regressions [22], Support vector machines [23], Ridge classifier [24], Gradient Boosting [25], and XGBoost [26], are then trained on the selected features. Finally, the trained models are utilized for making predictions on new farming scenarios, providing valuable insights into potential risks faced by farmers and enabling informed decision-making and risk management strategies. Fig. 1 illustrates the block diagram depicting the architecture of the envisioned system.

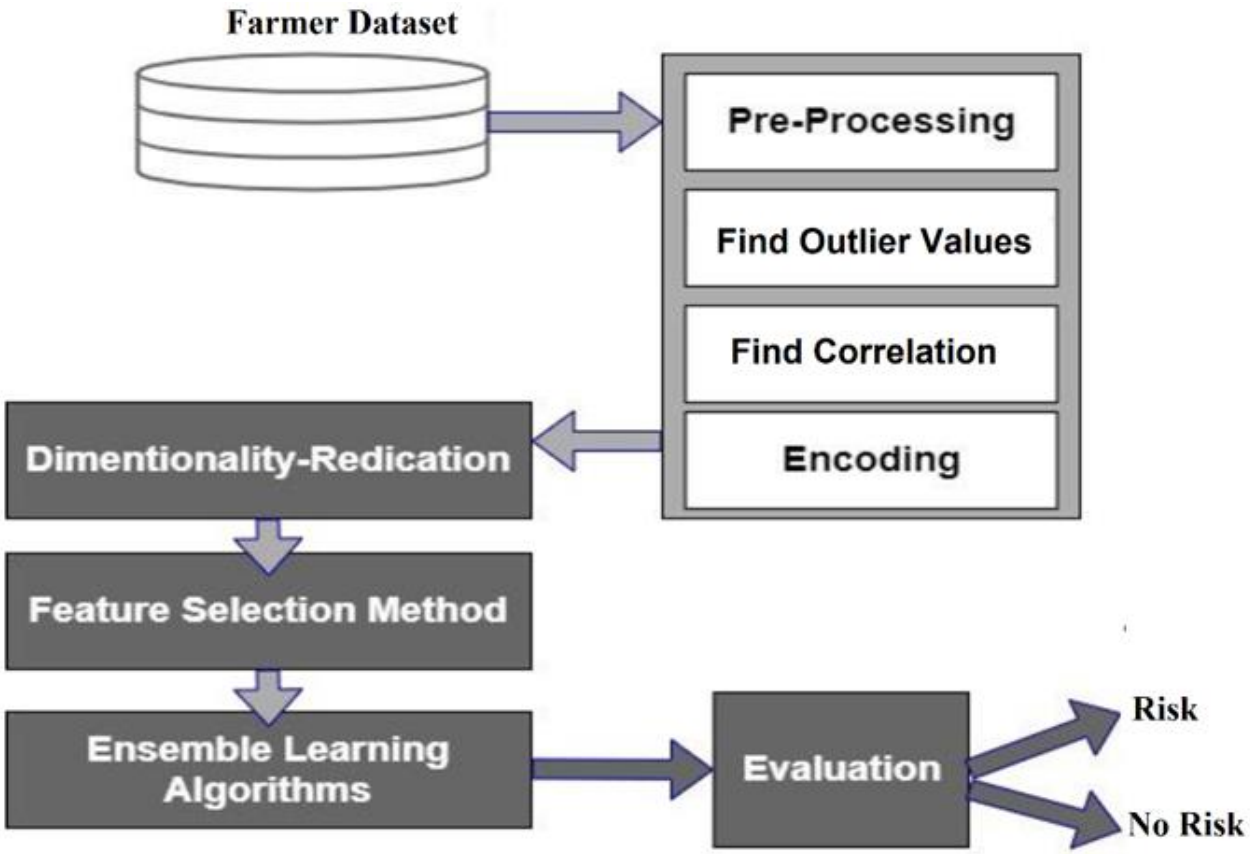


Fig. 1. Proposed farmers risk prediction system.

A. Dataset Description

The farmer dataset utilized in this study serves as a comprehensive repository of factors influencing farmers' risk. Collected through a nationwide survey, the dataset encapsulates the diverse perspectives and experiences of farmers across different regions of the country. Given the multifaceted nature of agricultural risk, the dataset captures a wide array of factors, ranging from climatic conditions and soil quality to crop varieties, farming techniques, and socioeconomic indicators. With a total of 12 features encompassing these diverse risk factors as tabulated in Table I, the dataset provides a rich foundation for developing ML model aimed at predicting and mitigating farmers' risk.

Each feature included in the dataset represents a distinct aspect of the agricultural ecosystem, reflecting the intricate interplay of environmental, socio-economic, and agronomic factors influencing farmers' risk levels. The sample dataset is depicted in Fig. 2. By synthesizing farmers' opinions and experiences, the dataset offers a holistic view of the challenges and opportunities faced by agricultural communities in terms of risk exposure. By analyzing such a dataset, predictive models can be trained to forecast risks effectively, helping farmers and stakeholders make informed decisions to mitigate potential adverse outcomes. The structured representation of these features allows for a comprehensive analysis, contributing to the development of robust risk prediction frameworks in agriculture.

TABLE I. FEATURES IN THE DATASET

Features	Range
Weather	Favorable- 0 Not Favorable -1
Pest	Absent-0 Present-1
Diseases	Absent -0 Moderate-1 Severe -2
Input Price	Non-Volatile-0 Volatile-1
Product Price	Increasing-0 Decrease-1
Product Type	Non-Perishable-0 Perishable-1
Duration	3 months to 6 months-1 Up to 3 months-0 More than 9 months-3 6 months to 9 months-2
Finance	Own Money-0 Bank Loan-1
Subsidies	Yes-0 No-1
Technology Adoption	Yes-0 No-1
Insurance	Yes-0 No-1
Eco Sensitive Zone	No-0 Yes-1
Target	No Risk-0 Risk-1

Weather	Pest	Finance	Diseases	Input Price	Product Price	Product Type	Subsidies	Technology Adoption	Insurance	Eco Sensitive Zone	Duration	Target
0	0	1	1	0	0	0	1	0	0	1	0	1
1	0	0	2	0	0	0	0	0	0	0	0	1
0	0	0	2	0	0	0	0	0	0	0	1	0
0	0	1	1	1	0	0	1	0	1	1	0	0
1	0	0	2	0	0	0	0	0	0	0	0	1
0	0	0	1	1	0	0	0	0	0	0	3	0
0	0	0	1	1	1	0	0	0	0	0	0	1
2	0	0	2	1	0	0	0	0	0	0	0	1
2	0	0	2	0	0	0	0	0	0	0	0	1
1	0	0	2	0	0	0	0	0	0	0	0	1

Fig. 2. Sample dataset visualization.

B. Data Preprocessing

Data preprocessing involves several essential steps aimed at preparing the data for analysis and modeling. This process typically involves cleaning, transforming, and organizing the data to ensure its quality, consistency, and relevance for predictive modeling purposes.

1) Finding outlier: Outlier detection plays a crucial role in the preprocessing step, the dataset is partitioned into quartiles including Q1, Q2 (median), and Q3, from which the interquartile range (IQR) is calculated as the difference between the third and first quartiles. Subsequently, data points deviating below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are flagged as potential outliers. Outliers are exceptional conditions or extreme values indicating influential factors impacting risk assessments. Table II displays the dataset after the application of quartile ranges.

For instance, anomalies such as unusually high or low rainfall, atypical market fluctuations, or unexpected shifts in socio-economic indicators could signal outlier observations requiring closer examination. Detecting and addressing these outliers are essential as they could either represent genuine anomalies warranting further process or erroneous data entries capable of skewing risk prediction models. Therefore, integrating the quartile range method during preprocessing enables researchers to effectively identify and manage outliers, thereby ensuring the robustness and accuracy of subsequent analyses and predictive modeling efforts in farmers' risk assessment.

2) Finding correlation: Correlation analysis uncovering the relationship between different variable, pairwise correlation coefficient is computed to evaluate the strength and direction of the linear relationship between each pair of variables in the dataset. This entails figuring out Pearson correlation coefficients, which have a range of -1 to 1, with values near -1 denoting a strong negative correlation, values near 0 showing no linear link, and values closer to 1 indicating a significant positive correlation. Following the computation of correlation coefficients, a correlation matrix is constructed as shown in Fig. 3, offering a comprehensive overview of the relationships between all variables pertinent to farmers' risk prediction.

Significant findings include a strong positive correlation between 'Product Type' and 'Pest' (1.00), indicating that certain product types are more susceptible to pest infestations. 'Insurance' also shows a high positive correlation with 'Finance' (0.77), suggesting that better financial health is associated with higher insurance coverage. Conversely, 'Target' (representing risk) shows strong negative correlations with 'Finance' (-0.44), 'Diseases' (-0.35), and 'Subsidies' (-0.41), implying that better financial conditions, fewer diseases, and more subsidies are associated with reduced risk. Additionally, 'Technology Adoption' correlates positively with 'Product Type' (0.75) and 'Pest' (0.95), suggesting that technological advancements are more prevalent in certain product types and pest management.

TABLE II. DATASET AFTER APPLYING QUARTILE RANGES

Feature	Value
Weather	0
Pest	918
Finance	0
Diseases	0
Input Price	0
Product Price	1500
Product Type	918
Subsidies	0
Technology Adoption	1002
Insurance	1412
Eco Sensitive Zone	0
Duration	526
Target	0

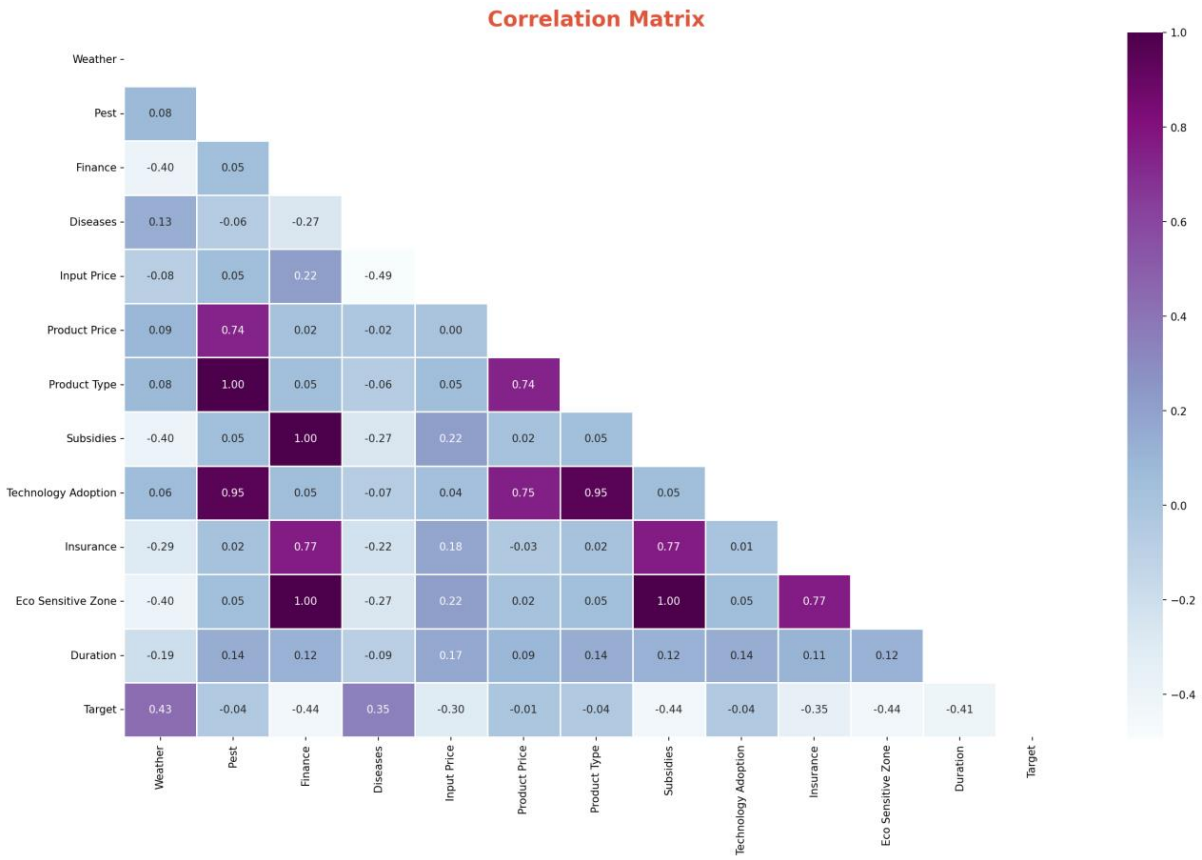


Fig. 3. Correlation matrix.

3) Encoding: Agricultural datasets contain categorical variables representing qualitative attributes such as crop types, farming practices, or geographical regions. However, most ML algorithms are designed to process numerical data, necessitating the conversion of categorical variables into a numerical format. During encoding, assigning unique numerical identifiers to each category within a categorical variable, enables computational models to effectively interpret and analyze the data.

C. Variation Inflation Factor

The paramount importance of mitigating multicollinearity risks ensures the reliability and accuracy of our models. To address this concern, we adopted a Variation Inflation Factor (VIF) approach, leveraging its iterative analysis to detect and manage multicollinearity effectively.

The VIF method facilitated the identification of correlated predictor variables, which might not exhibit significant effects when considered together but demonstrate their true significance when assessed independently. VIF computation involved conducting linear regressions for each predictor variable and obtaining the coefficient of determination (R^2). The VIF value was calculated using the Eq. (1).

$$VIF_i = \frac{1}{1-R_i^2} \quad (1)$$

VIF value of one indicates no correlation, increasing values signify stronger correlations with other variables. Models ignoring collinearity risks often exhibit high variance and instability, making it challenging to discern the relative importance of each variable and leading to inaccurate tests of significance. Features exhibiting VIF values exceeding 10,000 were deemed excessively collinear and consequently eliminated from the selection process. We adopted a practical interpretation guideline for VIF values: variables with VIF > 10 were removed outright, those with VIF > 5 were subject to scrutiny before elimination, and variables with VIF < 5 were deemed valuable and retained in the analysis, as shown in Fig. 4.

	Predictor	VIF
0	Finance	inf
1	Subsidies	inf
2	Eco Sensitive Zone	inf
3	Diseases	2.594175
4	Input Price	2.547087
5	Weather	2.050697
6	Duration	1.688319
7	Pest	NaN

Fig. 4. VIF Output for feature selection.

D. Proposed Classifier Models

Ensemble learning, a machine learning technique employed in our research, significantly bolsters accuracy and resilience in

forecasting by amalgamating predictions from multiple models. By harnessing the collective intelligence of the ensemble, this approach aims to mitigate errors or biases inherent in individual models. The methods utilized in the proposed study encompass a diverse range, including K Nearest Neighbor (KNN), Random Forest, Gradient Boosting, XGBoost, Support Vector Classifier (SVC), Logistic Regression, and Ridge Classifier. By leveraging the strengths of these various algorithms, our ensemble learning framework endeavors to provide robust and reliable predictions for farmer risk prediction tasks.

1) K Nearest Neighbour: The k Nearest Neighbors (kNN) algorithm operates by assigning a class label to a test point based on the majority class of its k nearest neighbors [27]. In the 1-NN approach, the class of the closest neighbor is directly assigned to the test point, which can lead to errors if the nearest neighbor is an outlier.

However, by considering a larger k value, such as in the kNN approach with k = 7, the influence of outliers is mitigated as the class assignment is determined by the majority class among the k nearest neighbors. This approach improves the reliability of class assignments, where the majority class among the k = 7 nearest neighbors yields a more accurate classification compared to the 1-NN approach. The choice of distance and similarity measures plays a crucial role in various pattern recognition tasks.

Let's denote our training dataset as $D = \{(x_i, y_i)\}\, for\ n\ terms$ where x_i represents the feature vector for i^{th} sample and y_i represents the corresponding risk level. Euclidean distance measures the similarity between feature vectors. For two feature vectors x_i and x_j the Euclidean distance is given by Eq. (2).

$$D\ (x_i, x_j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \tag{2}$$

When presented with a new data point, x, to predict the risk level, the k nearest neighbors to x are identified based on the calculated distances. In regression tasks such as predicting risk levels, the average of the risk levels of the k nearest neighbors is utilized as the prediction as illustrate by Eq. (3).

$$\hat{y}_{new} = \frac{1}{k}\sum_{i=1}^{k} x_i \tag{3}$$

Where, $\hat{y}_{new}$ is the predicted risk level for the new data point x_i,y_i are the risk level of k nearest neighbors.

2) Random forest: Predictions of multiple decision trees are combined in Random Forest to produce a robust and accurate final prediction [28]. By introducing randomness during both the training and prediction phases, Random Forest mitigates overfitting and increases diversity among the constituent trees. The decision tree construction process would involve selecting the most informative features at each node to effectively partition the data based on factors such as weather conditions, pest infestation, disease prevalence, market prices, crop types, financial factors, technological adoption, insurance coverage, and environmental considerations.

Given a dataset with N data points and M features, each decision tree T_i is built by recursively partitioning the feature space based on selected features. At each node j, a split is made by selecting the feature f that maximizes information gain or minimizes impurity. The decision tree construction process can be represented mathematically as in Eq. (4).

$$f_j = arg\ arg\ max\ _f Gain(D_j, f) \tag{4}$$

Where D_j represents the dataset at node j and $Gain(D_j, f)$ denotes the information gain achieved by splitting on feature f. Bootstrap sampling allows us to create diverse training datasets that capture various combinations of weather patterns, pest and disease occurrences, market conditions, financial situations, technological adoption rates, and other relevant factors affecting farmers' risk. The bootstrap sampling process can be expressed as in Eq. (5).

$$D_i = Bootstrap\ sampling(D) \tag{5}$$

where D is the original dataset and D_i represents the bootstrap sample for tree T_i. In the prediction phase, the Random Forest algorithm aggregates predictions from all decision trees. For regression tasks like the proposed method, the final prediction $\hat{y}_{RF}$ is calculated as the average prediction across all trees as depicted in Eq. (6).

$$\hat{y}_{RF} = \frac{1}{T}\sum_{i=1}^{T} \hat{y}_i \tag{6}$$

where T is the total number of trees in the forest and $\hat{y}_{RF}$ is the prediction from tree T_i.The aggregated prediction considers the combined insights from all decision trees trained on diverse subsets of features, enabling a comprehensive assessment of the potential risks faced by farmers based on factors.

3) Gradient boosting: Gradient Boosting sequentially constructs a series of weak learners with each subsequent learner focusing on the residuals or errors of its predecessor. By iteratively refining predictions based on the gradient of a predefined loss function, Gradient Boosting enhances predictive accuracy and resilience by placing emphasis on previously mis-predicted data points [29]. This approach is particularly advantageous in agricultural risk prediction scenarios, where nonlinear and complex relationships between predictors and outcomes prevail due to the multitude of interacting factors at play. Gradient boosting trees usually have deeper trees, such as ones with 8 to 32 terminal nodes.

Given a training dataset comprising features X and corresponding risk labels y, the algorithm iteratively fits a series of weak learners $h_i(x)$ to the residuals or negative gradients of the loss function. At each iteration t, the model updates its prediction $\hat{y}_t$ by incorporating a weighted contribution from the new weak learner $h_i(x)$. The final prediction $\hat{y}$ is obtained as the sum of all individual weak learner predictions, represented mathematically as in Eq. (7).

$$\hat{y}\,(x) = \sum_{t=1}^{T} \gamma_t h_t(x) \tag{7}$$

where γ_t denotes the learning rate or shrinkage parameter, regulating the influence of each weak learner, and T signifies the total number of iterations. The primary goal is to minimize the loss function, commonly expressed as the mean squared error for regression tasks or cross-entropy loss for classification tasks, by iteratively adjusting the parameters of the weak learners. Through this iterative refinement process, Gradient Boosting optimizes the model's capacity to capture intricate relationships inherent in agricultural data, furnishing farmers with precise risk assessments tailored to their specific contexts, thereby facilitating informed decision-making and effective risk management strategies.

4) XGBoost: XGBoost enhances predictive capabilities through its advanced ensemble learning techniques. It is an implementation of gradient-boosted decision trees designed for speed and performance. XGBoost operates by constructing an ensemble of decision trees in a sequential manner, where each new tree attempts to correct errors made by the previous ones. It incorporates several advanced features such as regularization to prevent overfitting, parallel processing for faster computation, and a sparsity-aware algorithm to handle missing data effectively.

During the training phase, given a dataset comprising features X and corresponding risk labels y, XGBoost iteratively builds decision trees to minimize a predefined objective function. Each decision tree $h_t(x)$ is trained to predict the residuals or negative gradients of the loss function. Prediction $\hat{y}$ is obtained as the sum of predictions from all decision trees, with parameters such as the learning rate γ_t controlling each tree's contribution. XGBoost optimizes a regularized objective function, consisting of a loss term measuring prediction error and a regularization term penalizing model complexity. The objective function is expressed as in Eq. (8).

$$Obj = \sum_{i=1}^{N} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \omega(f_k) \quad (8)$$

where L $(y_i, \hat{y}_i)$ represents the loss function, N is the number of data points, K is the number of trees, and $\omega(f_k)$ is the regularization term for the k[th] tree. XGBoost employs L1 and L2 regularization techniques to control model complexity and prevent over fitting, ensuring stability and enhancing model robustness.

5) Support Vector Machine: SVM employs a dataset comprising features *X* and corresponding risk labels y, where represents a matrix of m data points and n features, and y denotes a vector of risk labels for each data point. The SVM algorithm endeavors to delineate a hyperplane, represented as in Eq. (9).

$$W^T x + b = 0 \quad (9)$$

which effectively segregates the data points into various risk classes while maximizing the margin between these classes. SVM formulates an optimization objective aimed at finding the optimal hyperplane by simultaneously minimizing the classification error and maximizing the margin. This objective function is expressed as in Eq. (10).

$$min_w, b\ \frac{1}{2}||w||^2 + X\sum_{1=1}^{m} \varepsilon_i$$

$$\Sigma\upsilon\beta\varphi\varepsilon\chi\tau\ \tau o\ y^i w^T x^i + b \geq 1 - \varepsilon_i \varepsilon_i \geq 0 \quad (10)$$

Where, C is the regularization parameter control ling the balance between maximizing the margin and minimizing the classification error, while ε_i represents slack variables indicative of the classification error for each data point. SVM can adeptly handle nonlinear decision boundaries by employing kernel functions K (x, x') to map input features into higher-dimensional spaces. The decision function of the SVM model is then expressed as in Eq. (11).

$$f(x) = sign(\sum_{i=1}^{m} \alpha_i y^i \mathrm{K}(\mathrm{x}, y^i) + \mathrm{b} \quad (11)$$

Through training on the provided dataset, SVM determines an optimal hyper plane that effectively separates different risk levels based on input features.

6) Logistic regression: Logistic regression is a statistical model and supervised machine learning algorithm that uses data analysis to predict the probability of an event or observation. The most common logistic regression models a binary outcome, which can take two values like true/false or yes/no. Dataset containing features such as weather conditions, pest prevalence, diseases outbreak, input and product prices, product type, duration of farming activities, financial factors, subsidies availability, technology adoption, insurance coverage, and the presence of eco-sensitive zones. These features collectively form the input matrix X, where a farming scenario is represented by each row and each column corresponds to a specific feature. The model aims to predict the likelihood of a particular risk, represented as the target variable y, given the feature vector. The probability p(y=1|x) of the occurrence of the risk as a function of the input features. The logistic regression model applies the logistic function to transform the linear combination of features into a probability between zero and one. The function is defined as in Eq. (12).

$$\pi(\psi=1|\xi) = \frac{1}{1+e^{-z}} \quad (12)$$

Where β_0, $\beta_1 x_1, \beta_2 x_2, \beta_n x_n$ is the linear combination of features and coefficients, where $\beta_0, \beta_1, \beta_n$ are the coefficients or weights assigned to each feature and x_0, x_1, x_n are the values of the corresponding features for a given farming scenario. The coefficients $\beta_0, \beta_1, \beta_n$ are estimated during the training phase using optimization techniques such as maximum likelihood estimation or gradient descent. Once the coefficients are determined, the logistic regression model can predict the probability of occurrence of the risk for new farming scenarios based on their feature values. By setting a threshold probability, we can classify farming scenarios into different risk categories, providing valuable insights for farmers to make informed decisions and mitigate potential risks effectively.

7) Ridge classifier: The Ridge Classifier serves as a potent tool for classification tasks, effectively modeling the probability of various risks based on pertinent features. The Ridge Classifier aims to predict the probability of a specific risk occurrence, denoted as the target variable(y), given the feature vector(X). Fig. 5 shows the basic architecture of Ridge classifier.

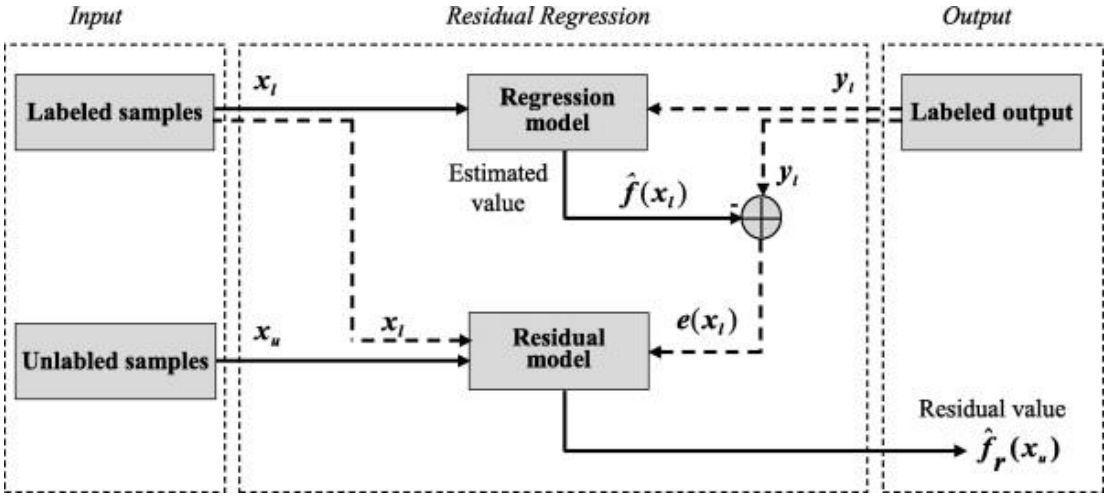


Fig. 5. Basic architecture of ridge classifier.

The dataset comprising features such as weather conditions, pest prevalence, disease outbreaks, input and product prices, product types, duration of farming activities, financial factors, subsidies availability, technology adoption, insurance coverage, and the presence of eco-sensitive zones. These features collectively constitute the input matrix (X). Ridge Classifier extends the logistic regression model by incorporating regularization to mitigate over fitting and improve model generalization. The objective function for Ridge Classifier can be formulated as in Eq. (13).

$$min_w||X_w - y||^2 + \alpha||w||^2 \quad (13)$$

where w represents the weight vector containing the coefficients for each feature, X is the feature matrix, y is the target variable, and α is the regularization parameter controlling the strength of regularization. The first term $||X_w - y||^2$ represents the residual sum of squares, measuring the difference between the predicted and actual target values. The second term $\alpha||w||^2$ is the L2 regularization term, penalizing large coefficients to prevent overfitting.

The Ridge Classifier optimizes the objective function to find the optimal weight vector w that minimizes the loss function while balancing the trade-off between fitting the training data and regularization. By incorporating the Ridge regularization term, the model is more robust to noisy data and less sensitive to multi-collinearity among features. Thus, the Ridge Classifier effectively predicts farmers' risk levels based on a comprehensive set of features, offering valuable insights for informed decision-making and risk management in agriculture.

E. Hardware and Software Setup

The proposed study utilized the Google Collaboratory platform in conjunction with the Microsoft Windows 10 operating system to establish a robust computational environment. The modeling process involved the application of the Python programming language, leveraging the Keras package and Tensorflow backend for training. The conceptualized models specifically configured to accept preprocessed and augmented datasets, ensuring precise decision-making capabilities. To assess the efficacy of the proposed model evaluating the predictions of the model on the test dataset.

IV. Experimental Results

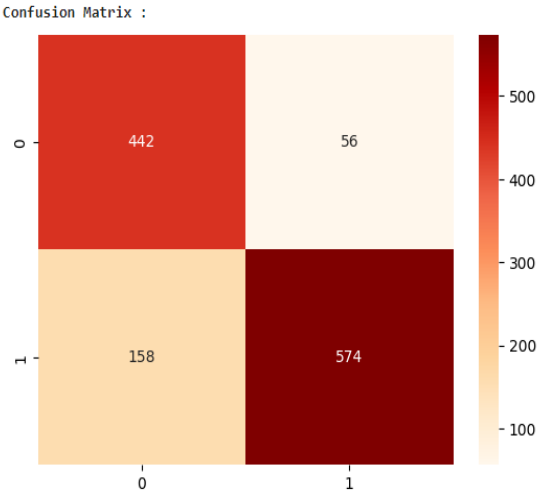
Performance parameter, accuracy is used to evaluate the effectiveness of classification model. Accuracy provides a general measure of model performance, it may not be sufficient when dealing with imbalanced datasets, where one class dominates the others.

The performance evaluation of prediction models involved the assessment of various classifiers, including K-Nearest Neighbors, Random Forest, Gradient Boosting, XGBoost, Support Vector Classifier, Logistic Regression, and Ridge Classifier. The success of these models in predicting farmers' risk levels can be attributed to their ability to capture complex relationships between various features such as weather conditions, pest prevalence, financial factors, and technological adoption. By leveraging the collective knowledge from multiple features, these classifiers were able to effectively differentiate between different risk levels faced by farmers. Additionally, the ensemble nature of Random Forest, Gradient Boosting, and XGBoost allows them to handle nonlinear relationships and interactions between features, contributing to their superior performance. Among these classifiers, KNN, Random Forest, Gradient Boosting, and XGBoost emerged as the top performers, achieving an impressive accuracy score of 88.46%. The indication in Table III shows that farmers' risk levels were correctly predicted in 88.46% of cases.

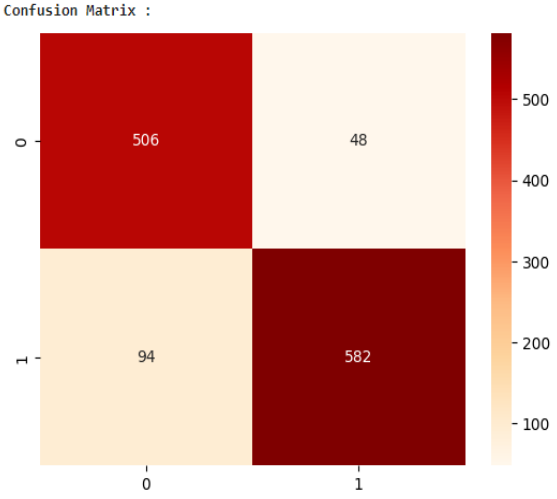
TABLE III. Model Comparison

Model	Accuracy
KNN	88.46
Random Forest	88.46
Gradient Boosting	88.46
XG Booster	88.46
SVC	88.05
Logistic Regression	82.60
Ridge Classifier	82.03

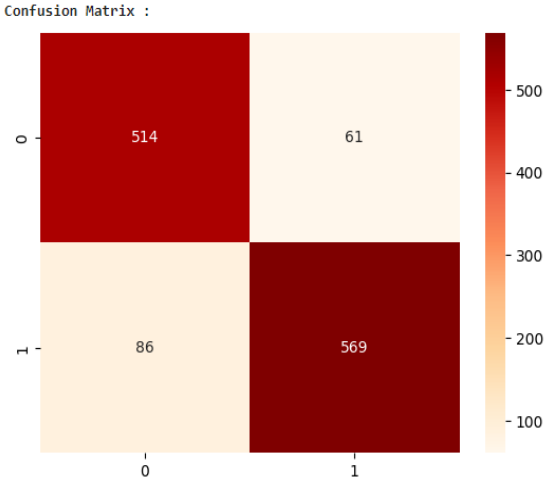
A confusion matrix is a tabular representation used to evaluate the performance of a classification model by summarizing the counts of true positive, true negative, false positive, and false negative predictions. It consists of rows and columns corresponding to actual and predicted classes, respectively, where each cell represents the count of instances. The main diagonal of the confusion matrix contains the counts of correct predictions, while off-diagonal elements indicate misclassifications. This matrix provides valuable insights into the model's ability to accurately classify instances and helps identify common types of errors such as false positives and false negatives. By analyzing the confusion matrix, stakeholders can assess the strengths and weaknesses of the classification model and make informed decisions regarding model improvement and optimization strategies. Fig. 6 shows the confusion matrix of the proposed models.



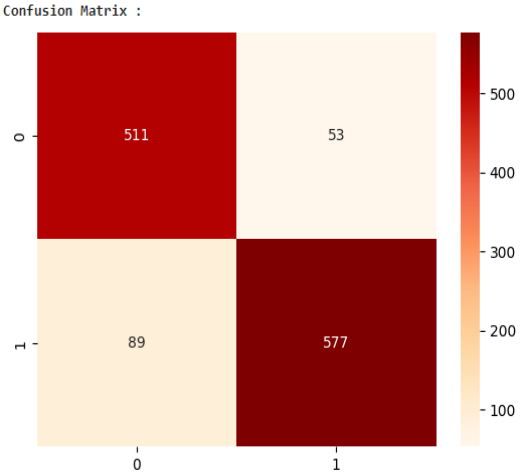
(a) Logistic Regression



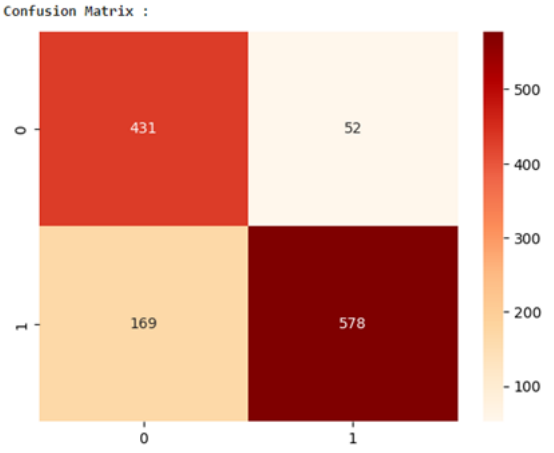
(b) KNN



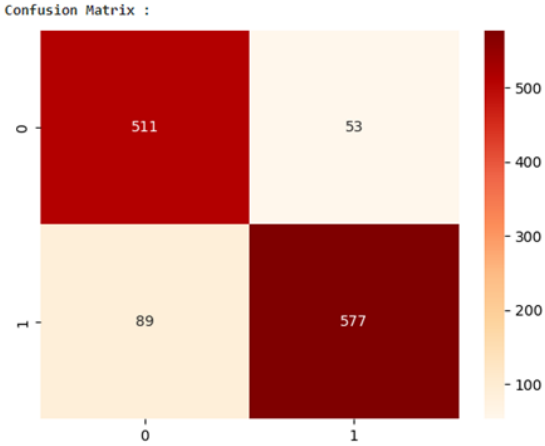
(c) SVM



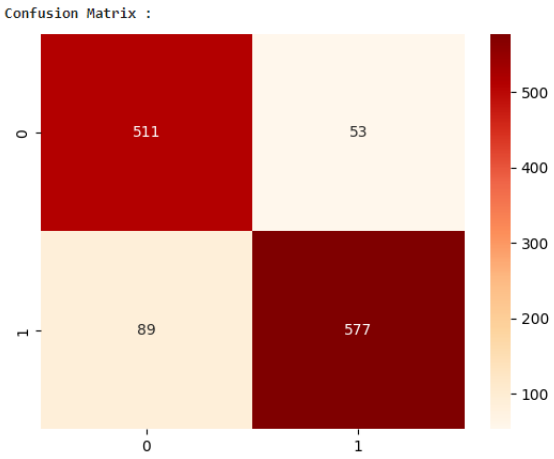
(d) Random Forest



(e) Ridge Classifier



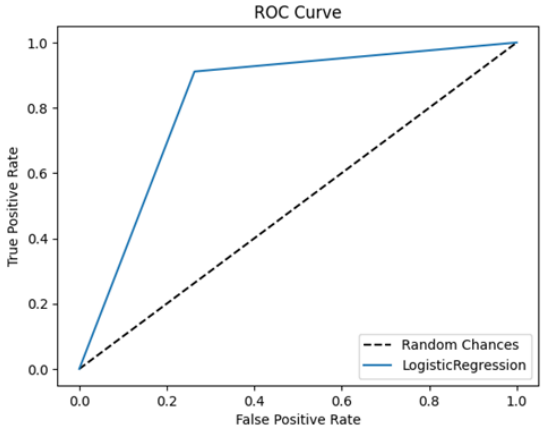
(f) Gradient Boosting



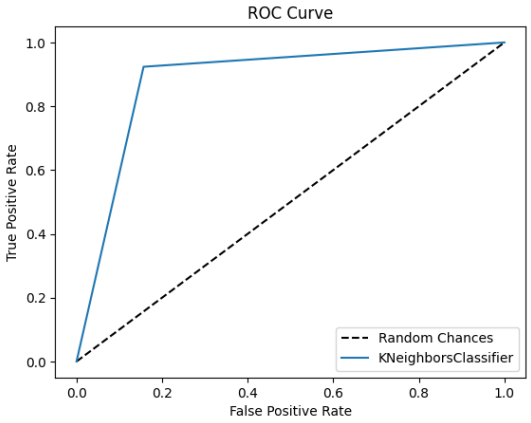
(g) XG Boost

Fig. 6. Confusion matrix.

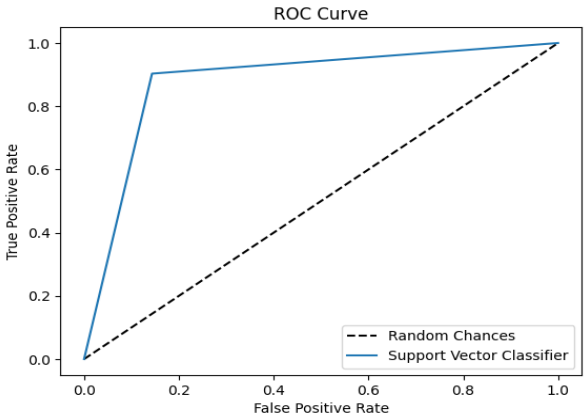
The Receiver Operating Characteristic curve is used to assess the performance of binary classification models by plotting the true positive rate against the false positive rate across various threshold values. The True Positive Rate, also known as sensitivity or recall, is the ratio of correctly predicted positive observations to the total actual positives. The False Positive Rate, on the other hand, is the ratio of incorrectly predicted positive observations to the total actual negatives. The ROC curve provides a comprehensive visualization of a classifier's ability to distinguish between the positive and negative classes, with a steeper curve indicating higher discriminative power. The area under the ROC curve quantifies the overall performance of the classifier, with a value closer to 1 indicating better performance. ROC curves are particularly useful for evaluating classifiers in imbalanced datasets and for selecting an optimal threshold value that balances sensitivity and specificity based on the specific requirements of the application. Fig. 7 shows the ROC curves of the proposed models.



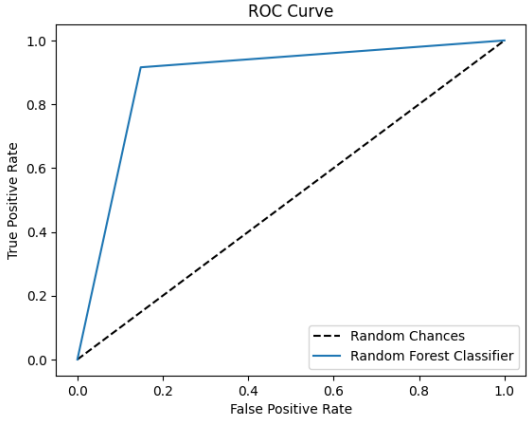
(a) Logistic Regression



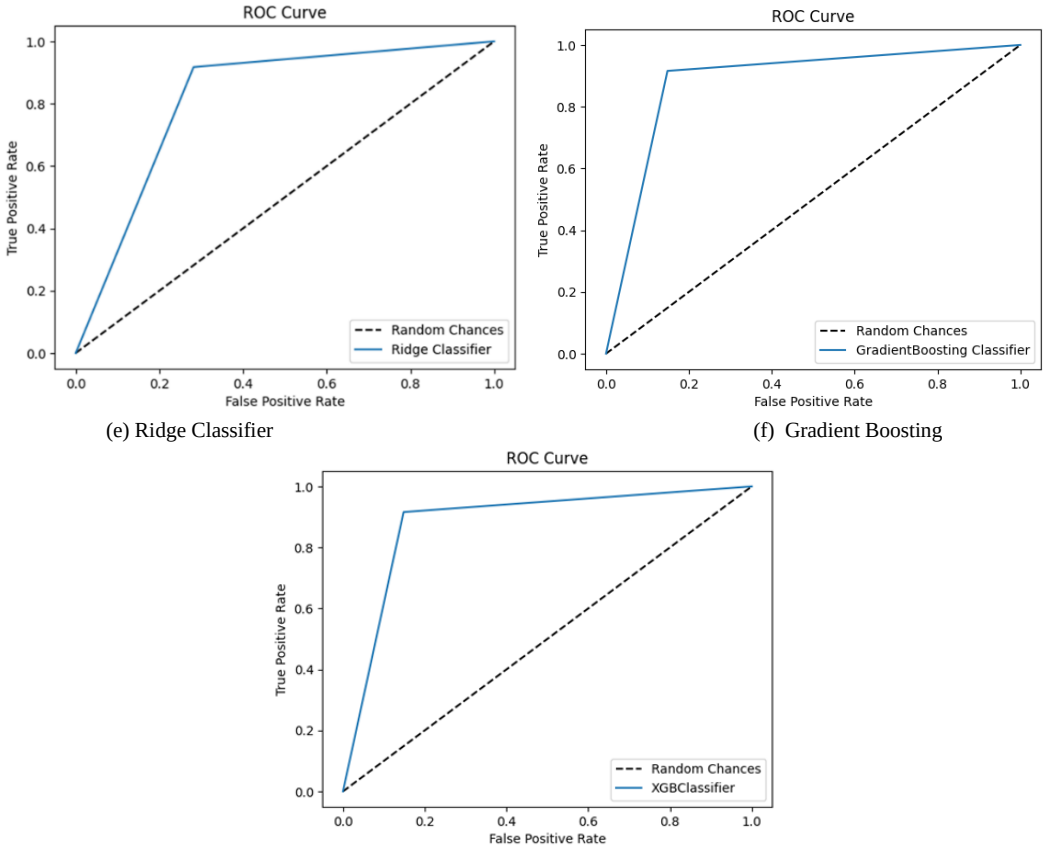
(b) KNN



(c) SVM



(d) Random Forest



(e) Ridge Classifier

(f) Gradient Boosting

(g) Extreme Gradient Boosting

Fig. 7. ROC Curve.

V. Discussion

Fig. 8 provides a visual representation comparing the performance of various classifiers, including Logistic Regression, KNN, SVM, Random Forest, Ridge Classifier, Gradient Boosting, and XGBoost. The results indicate that KNN, Random Forest, Gradient Boosting, and XGBoost all achieved the highest accuracy rate of 88.46%. These methods are closely followed by the SVM algorithm, which demonstrated a slightly lower accuracy of 88.05%. The superior performance of these algorithms can be attributed to their ability to handle complex patterns and interactions within the data effectively. Notably, ensemble methods such as Random Forest, Gradient Boosting, and XGBoost tend to provide robust predictions by combining the strengths of multiple base learners, which might explain their high accuracy in this context.

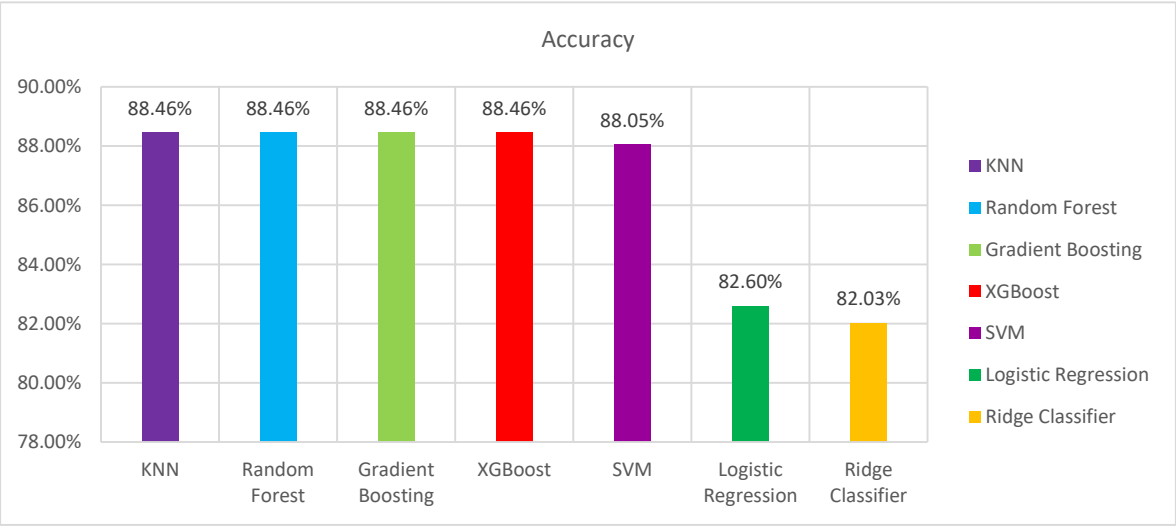


Fig. 8. Visualization of performance comparison of proposed models.

Logistic Regression and Ridge Classifier, however, exhibited lower accuracies, with 82.60% and 82.03% respectively. These methods, being more simplistic linear models, might not capture the nonlinear relationships in the data as effectively as the other more complex algorithms. Logistic Regression is a fundamental classification technique that is easy to implement and interpret but may fall short in performance compared to advanced models like ensemble methods and SVM. Similarly, Ridge Classifier, while being effective in regularizing the model to prevent over fitting, might not perform optimally in scenarios requiring sophisticated decision boundaries.

The slight edge in accuracy for the ensemble methods and SVM over logistic and ridge regression models emphasizes the importance of algorithm selection in predictive analytics. Ensemble methods, which combine multiple models to improve prediction accuracy, and SVM, known for its high-performance margin maximization, prove to be more adept in this case of farmer risk prediction.

Overall, the comparison underscores the effectiveness of advanced machine learning techniques, particularly ensemble methods and SVM, in achieving high prediction accuracy. These results suggest that employing such algorithms can significantly enhance the predictive performance in farmer risk prediction models, thereby supporting better decision-making and risk management strategies in agricultural practices. Future work could explore the integration of these models with more comprehensive feature sets and hyper parameter tuning to further optimize prediction outcomes.

VI. Conclusion

Agriculture, which makes up the majority of India's economy, is the primary backbone of our rural economy. Risk in agriculture is the result of a hazardous event, which is expressed as a combination of the likelihood and magnitudes of the risk. By analyzing the given farmer dataset and optimizing it through pre-processing techniques, the study ensures that the predictive models are built on high-quality data, thereby enhancing the reliability of the risk predictions. Through the utilization of Variation Inflation Factor (VIF) for feature selection, the study identifies the most influential features for accurate risk classification, demonstrating a meticulous approach towards model optimization and performance improvement. Utilizing a diverse array of techniques including KNN, Random Forest, Logistic Regression, SVM, Ridge Classifier, Gradient Boosting, and XGBoost, the study demonstrates significant progress. Notably, KNN, Random Forest, Gradient Boosting, and XGBoost exhibit exceptional performance, achieving a notable accuracy rate of 88.46%. The proposed farmers' risk prediction study represents a significant contribution to agricultural decision-making and risk management strategies. The study also acknowledges the potential for further improvement through the integration of Deep Learning Models, suggesting avenues for future research and development in agricultural risk prediction.

References

[1] Antle, J. M., & Ray, S. (2020). Sustainable agricultural development. Palgrave Studies in Agricultural Economics and Food Policy. 1st ed. Palgrave Macmillan Cham, 10, 978-3.

[2] Ferreira, H., Pinto, E., & Vasconcelos, M. W. (2021). Legumes as a cornerstone of the transition toward more sustainable agri-food systems and diets in Europe. Frontiers in Sustainable Food Systems, 5, 694121.

[3] Khan, M. A. (2021). Impact of agriculture sector on sustainable development of indian economy: An analysis. Ama, Agricultural Mechanization in Asia, Africa & Latin America, 52(02), 10.

[4] Huet, E. K., Adam, M., Giller, K. E., & Descheemaeker, K. (2020). Diversity in perception and management of farming risks in southern Mali. Agricultural Systems, 184, 102905.

[5] Shahzad, A., Ullah, S., Dar, A. A., Sardar, M. F., Mehmood, T., Tufail, M. A., ... & Haris, M. (2021). Nexus on climate change: Agriculture and possible solution to cope future climate change stresses. Environmental Science and Pollution Research, 28, 14211-14232.

[6] Peace, N. (2020). Impact of climate change on insects, pest, diseases and animal biodiversity. International Journal of Environmental Sciences & Natural Resources, 23(5), 151-153.

[7] Ceballos, F., Kannan, S., & Kramer, B. (2020). Impacts of a national lockdown on smallholder farmers' income and food security: Empirical evidence from two states in India. World Development, 136, 105069.

[8] Rajpoot, K., Singh, A., & Sunil, J. (2023). Ranking of major agricultural risks using Garrett's ranking technique in Jabalpur district of India.

[9] Jinger, Jyoti, and Shiv Kumar. "Maize Yield Prediction Considering Growth Stages using Fuzzy Logic Modelling." International Journal of Engineering Research & Technology (IJERT) 9.4 (2021).

[10] Upadhya, S. M., & Mathew, S. (2020). Implementation of fuzzy logic in estimating yield of a vegetable crop. In Journal of Physics: Conference Series (Vol. 1427, No. 1, p. 012013). IOP Publishing.

[11] Pandhe, A., Nikam, P., Pagare, V., Palle, P., & Dalgade, D. (2019). Crop yield prediction based on climatic parameters. International Journal of Research in Engineering and Technology (IJRET), 6(03).

[12] Kalimuthu, M., Vaishnavi, P., & Kishore, M. (2020, August). Crop prediction using machine learning. In 2020 third international conference on smart systems and inventive technology (ICSSIT) (pp. 926-932). IEEE.

[13] Mulla, S. A., & Quadri, S. A. (2020). Crop-yield and price forecasting using machine learning. International journal of analytical and experimental modal analysis, 12(8), 1731-1737.

[14] Mohanty, M. K., Thakurta, P. K. G., & Kar, S. (2023). Agricultural commodity price prediction model: a machine learning framework. Neural Computing and Applications, 35(20), 15109-15128.

[15] Rani, S., Kumar, S., Jain, A., & Swathi, A. (2022, October). Commodities Price Prediction using Various ML Techniques. In 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS) (pp. 277-282). IEEE.

[16] Chen, H., Chen, Z., Lin, F., & Zhuang, P. (2021). Effective management for blockchain-based agri-food supply chains using deep reinforcement learning. IEeE Access, 9, 36008-36018.

[17] Rakhra, M., Sanober, S., Quadri, N. N., Verma, N., Ray, S., & Asenso, E. (2022). Implementing Machine Learning for Smart Farming to Forecast Farmers' Interest in Hiring Equipment. Journal of Food Quality.

[18] Chelliah, B. J., Latchoumi, T. P., & Senthilselvi, A. (2024). Analysis of demand forecasting of agriculture using machine learning algorithm. Environment, Development and Sustainability, 26(1), 1731-1747.

[19] Cheng, J., Sun, J., Yao, K., Xu, M., & Cao, Y. (2022). A variable selection method based on mutual information and variance inflation factor. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 268, 120652.

[21] Mucherino, A., Papajorgji, P. J., Pardalos, P. M., Mucherino, A., Papajorgji, P. J., & Pardalos, P. M. (2009). K-nearest neighbor classification. Data mining in agriculture, 83-106.

[22] Rigatti, S. J. (2017). Random forest. Journal of Insurance Medicine, 47(1), 31-39.

[23] Speelman, D. (2014). Logistic regression. Corpus methods for semantics: Quantitative studies in polysemy and synonymy, 43, 487-533.

[24] Suthaharan, S., & Suthaharan, S. (2016). Support vector machine. Machine learning models and algorithms for big data classification: thinking with examples for effective learning, 207-235.

[25] Hazarika, B. B., & Gupta, D. (2023). Affinity based fuzzy kernel ridge regression classifier for binary class imbalance learning. Engineering Applications of Artificial Intelligence, 117, 105544.

[26] Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.

[27] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[28] Kataria, A., & Singh, M. D. (2013). A review of data classification using k-nearest neighbour algorithm. International Journal of Emerging Technology and Advanced Engineering, 3(6), 354-360.

[29] Pal, M. (2005). Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1), 217-222.

[30] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21.