DOI: 10.15849/IJASCA.251130.05

Int. J. Advance Soft Compu. Appl, Vol. 17, No. 3, November 2025
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Climatic Intelligence for Coffee: Yield Forecasting in India Using Stochastic Machine Learning and Abiotic Factor Modelling

Santhosh C. S.[1], Umesh K. K.[2], and Narendra Khatri [3,*]

[1]Department of Computer Applications, JSS Science and Technology University,
Mysuru-570006, Karnataka, India
e-mail: sancs84@jssstuniv.in
[2]Department of Information Science & Engineering, JSS Science and Technology
University, Mysuru-570006, Karnataka, India
e-mail: umeshkatte@jssstuniv.in
[3]Department of Mechatronics, Manipal Institute of Technology, Manipal Academy of
Higher Education, Manipal-576104, Karnataka, India
e-mail: narendra.khatri@manipal.edu
*Corresponding Author

Abstract

Accurate forecasting of agricultural yields is vital for enhancing farm-level decision-making, securing food supply chains, and reducing environmental impacts. Coffee, being a globally significant commodity, demands robust predictive frameworks, especially in regions like India, a major producer of Arabica and Robusta varieties. Despite the importance, limited research has examined the role of abiotic and climatological variables in Indian coffee yield prediction, with most existing studies focusing on other coffee-producing countries, thereby restricting region-specific insights. To address this gap, the present study employs long-term datasets (2004–2022) obtained from the Central Coffee Research Institute (CCRI) and Coffee Research Station, Balehonnur, Karnataka. Using stochastic machine learning algorithms, key abiotic factors—including rainfall, temperature, sunshine, humidity, vapor pressure, and dew point—were analyzed through multivariate feature selection and correlation-based grouping. Predictive models such as Bayesian Ridge, Lasso Regression, Elastic Net, Extra Tree, Gradient Boosting, and Random Forest were evaluated. Results demonstrated that Group-3 predictors (Year, Relative Humidity, Rainfall, Temperature) offered the highest accuracy, with Bayesian Ridge and Lasso Regression models achieving R^2 values of 0.81 and 0.80, respectively, alongside low RMSE values. The findings emphasize precipitation as the most influential variable and highlight the potential of tailored machine learning approaches for reliable, region-specific coffee yield forecasting.

Keywords: *Coffee, Stochastic approach, machine learning (ML), yield prediction, abiotic variables.*

1 Introduction

Coffee is a globally beloved beverage that captivates and thrills countless individuals. In terms of international trade, coffee is second only to gasoline. The southern Indian states

Received 22 June 2025; Accepted 22 September 2025

of Karnataka, Tamil Nadu, and Kerala produce the majority of India's coffee, although even non-traditional locations like Orissa, Andhra Pradesh, West Bengal, Maharashtra, and the eastern and northern regions contribute significantly.

Lima et al., (2018) presented a detailed study of previous climate conditions for Brazil and coffee production, including seasonal rainfall, elevation, Robusta coffee growing area, and climatic water balance. Researchers spatialized mean monthly rainfall using spherical, linear, and exponential models employing geostatic-tics spatial interpolation by kriging. Geostatistical methods for spatial interpolation use R^2 and RMSE coefficients of determination. Results show Robusta coffee's vulnerability to low rainfall index, rainfall seasonability, and water shortage, which reduces coffee yield [1]. Monica et al., (2004) compared and analyzed the monoculture and agroforestry coffee types considering two years of measuring vegetative growth, nutritional status, coffee plant yield, and minimum and maximum temperatures. Statistical analysis Students' test at 5% probability level examined node count, branch length, leaf count, and area. Research shows that agroforestry methods boost tree growth and reduce maximum and diurnal temperatures and found that monoculture produced 2443 kg/hectare of coffee, whereas agroforestry produced 515 kg/hectare [2].

Wu et al. (2025) demonstrated that ML-assisted Vis-NIR hyperspectral imaging, enhanced by batch effect removal and few-shot learning, enables rapid, accurate, and non-invasive flavor quality prediction in green coffee beans [3]. Wang et al. (2015) analyzed 254 Ugandan coffee plots using GPS and boundary line analysis, revealing substantial regional yield gaps in Robusta and Arabica compared to attainable yields across five production zones [4]. Chengappa, et al., (2016) analyzed 33 years of climate data to assess variability. In Kodagu, researchers want to find climatic elements that have influenced coffee production over 33 years. They have collected 54 samples from Robusta and Arabica growth areas. Calculating each month's coefficient of variation (CV) assessed climate variable volatility. Coffee yields have been declining due to a combination of factors, including a slight decrease in overall rainfall, increased monthly variability, rising temperatures, and reduced relative humidity [5].

Della Peruta et al. (2025) employed the G-Dynacof model to project Arabica yield declines of 16–35% under climate change, highlighting uncertain adaptation potential from shade agroforestry [6]. Lorenzo et al. (2018) applied FAO’s AEZ to Tabasco, predicting that under RCP8.5, rising temperatures could reduce Robusta coffee potential yields approximately by 41% by 2050 despite stable suitable areas[7]. Rahn et al., (2018) estimated the daily rain-fed coffee output using a CAF2014 model and spatially variable soil and meteorological data. This work updates the CAF2014 model for spatially contextualized decision support system to analyze climate-diverse Mt. Elgon in Uganda and Mt. Kilimanjaro in Tanzania. Results shows that, given appropriate soil water storage capacity, 50% shadow cover at low altitudes enhances coffee production by 13.5% under present climate [8].

Gines et al. (2025) demonstrated that XRF-based elemental profiling with machine learning effectively differentiated Philippine Robusta coffee origins, achieving up to 84% classification accuracy [9]. Byrareddy et al. (2020) applied CROPWAT and hierarchical Bayesian modelling on 558 Vietnamese farms, integrating spatial–temporal variability to assess irrigation needs and predict coffee yield responses under challenging climatic conditions [10]. Kittichotsatsawat et al., (2022) developed artificial neural networks and multiple linear regression to predict annual coffee yields and match production to market demand by measuring areas, zones, rainfall, relative humidity, lowest and highest temperatures, and humidity levels. Multiple linear regression revealed promising prediction accuracy for cherry coffee output with $R^2=0.9524$ and RMSE = 0.0642 [11].

Freitas et al. (2025) calibrated an agrometeorological model for Arabica coffee yield estimation in Brazil, improving accuracy (RMSE = 8.65 ha^{-1} ; $R^2 = 0.65$) and highlighting climatic and irrigation influences [12]. Varshitha et al., (2022) forecasted soil fertility and agricultural production, researchers have analyzed pH, temperature, humidity, organic carbon, precipitation, NPK, and moisture content. Researchers have evaluated the performance of models such as SVM, Decision Trees, Naïve Bayes, and KNN, focusing on test results using R^2 of bagging technique values and comparing them to other approaches for forecasting soil fertility and production [13]. Byrareddy et al., (2019) examined fertilizer management on 798 Vietnamese and Indonesian Robusta coffee farms from 2008 to 2017. Nutrient management is improved by studying coffee-growing nations' fertilizer usage. Using yearly rainfall, solar radiation, and maximum temperature, fertilizer rates did not affect coffee output. With a 0.05 significance threshold, a one-way ANOVA is utilized to compare fertilizer use per province [14].

Pambudi et al. (2024) demonstrated that random forest ($R^2 > 0.9981$; RMSE < 1.346) accurately predicted pyrolysis behavior of oxidatively torrefied spent coffee grounds, advancing biomass-to-bioenergy modeling [15]. Muliasari et al., (2022) estimated Robusta coffee yield from July–December 2020 Bogor District. Using a geographic information system (GIS), the research examined the digital elevation model (DEM), agro-climate, soil physical and organic qualities, land use, and socioeconomic aspects such protected areas, lakes, highways, and rivers. Processing input data involves interpolation and categorization. In Bogor, 2% (5,227.78 ha^{-1}) is fairly suitable (S2), 33% (99,189.20 ha^{-1}) is marginal (S3), and 65% (194,808.40 ha^{-1}) is not suitable [16]. Veenadhari et al. (2014) built an easy-to-use web site named ‘Crop Advisor’ to estimate how climate would affect agricultural productivity. The C4.5 algorithm locates Madhya Pradesh districts' key climate factors affecting agricultural yield. Agriculture yield is affected by numerous climate parameters, as shown by the program. Samples include wheat, soybean, rice, and maize [17].

Pambudi et al. (2025) developed an explainable ML framework, where k-NN achieved $R^2 > 0.99$, effectively predicting thermogravimetric properties of oxidatively torrefied spent coffee grounds for bioenergy applications [18]. Natarajan et al., (2016) used fuzzy cognitive map (FCM) learning techniques, Data Driven Nonlinear Hebbian Learning (DDNHL), and Genetic Algorithm (GA), sugarcane production is categorized. The suggested study's FCM model predicts precision agriculture sugarcane production using soil and climate parameters. Factors affecting sugarcane production prediction are established by this solution [19]. Sirsat et al., (2019) developed a prediction model for grapevine phenology, predicting yield throughout growth phases and identifying key factors. The work involves the creation of a relational dataset that incorporates meteorological conditions, grapevine production metrics, phonological dates, fertilizer application details, soil analysis results, and maturity index data. Feature selection embedding approaches such as Random Forest, LASSO, Elastic net, and Spike slab address dataset dimensionality issues in generalized linear models. Evaluation of predictive models involves splitting the dataset into training and test sets, and computed RMSE and RRMSE values for the better performed models [20]. Kumar et al., (2019) forecasted the crop yield using meteorological and wheat yield data gathered for the years 1984 to 2015 from IARI, New Delhi. Stepwise and Lasso regression approaches have been used for variable selection and crop yield forecasting. R2, RMSE, and MAPE for stepwise regression are 0.81, 195.90, and 4.54%, while Lasso regression is 0.95, 99.27, and 2.7. Lasso is 1.89 and 1.64 percent, whereas stepwise is -8.5 and 10.14 percent. Lasso outperforms stepwise within a limited range [21]. Kaul et al., (2005) presented artificial neural network (ANN) models to multiple linear regression methods to predict Maryland corn and soybean yields under typical climatic conditions. The models utilized historic yield data from Maryland locations. ANN corn yield models in Maryland had R^2 and RMSEs of 0.77 and 1036, whereas linear regression had 0.42 and 1356. The Maryland ANN soybean yield model has a R^2 value of 0.81 and RMSE of 214, whereas the linear regression model has a R^2 of 0.46 and RMSE of 312 [22]. Astuti et al., (2024) used e-nose coupled with ANN for the classification of roasting profile of coffee [23].

Kim et al., (2014) examined machine learning agricultural pest prediction systems. Study presents Bayesian network, neural network, MLR and SVM algorithms, along with examples of their use. Studies focused on specific crops, predicted leaf wetness, pests, and diseases, and proposed techniques like generalized regression neural networks, Ridge regressions, lasso, MLR, Bayesian, elastic net, and Random Forest regressions and their outputs [24]. Shastry et al., (2016) ANN and MLR estimate wheat output using rainfall, transpiration, biomass, ESW, soil nitrogen (NO_3), soil evaporation, and historical wheat yield. Compared ANN and MLR outcomes using R2 and prediction inaccuracy. MLR, D-ANN, and C-ANN models achieved R^2 scores of 92.52, 95, and 97% on the test set MLR, D-ANN, and C-ANN models had average prediction errors of 4.19, 2.24, and 0.52% on the test set. In R^2 and percentage prediction error, C-ANN outperformed D-ANN and MLR. In the data set, C-ANN predicted wheat production better than MLR and D-ANN [25].

Shakoor et al., (2017) the research seeks to forecast intelligent agricultural information in Bangladesh. The six primary crops tested are rice varieties Aman, Boro, Aus, Potato, Jute, and Wheat. The forecast analyses static data like temperature, rainfall, and yield using Supervised Machine Learning. Decision Tree, ID3, predicted Aus, Aman, and Wheat better. KNN outperformed ID3 for Boro, Jute, and Potato predictions [26]. Mishra et al., (2016) article reassesses machine learning in agricultural production studies. The study uses novel data variables to find crucial linkages. Decision trees, Bayesian belief networks, information fuzzy networks, artificial neural networks, and regression analysis are methodologies. Support vector machines, k-means clustering, k-closest Neighbour, and time series analysis are also discussed in agriculture [27].

Chen et al., (2016) used support vector machines (SVMs) to examine how mean temperature, rainfall, relative humidity, sunshine hours, daily temperature range, and wet days affect paddy rice production variation in south-western China. The study compares SVM models to ANNs and MLR, analyzing performance accuracy using MAE, MRAE, RMSE, RRMSE, and R^2 metrics. Multivariate regression and artificial neural networks perform poorly compared to SVMs. Sunshine, daily temperature range, rainfall, relative humidity, mean temperature, and wet days affect rice production variation in the research region [28]. Gonzalez et al., (2014) research evaluates the accuracy of ML and linear regression approaches in predicting agricultural productivity across 10 datasets. Ranking multiple linear regression, M5-Prime regression trees, perceptron multilayer neural networks, support vector regression, and k-nearest Neighbour systems. Validating the models included four accuracy metrics: RMS, RRSE, MAE, and R. Real Mexican irrigation zone data is used to develop models for crops such as pepper, tomato, chickpea, maize, potato, and common bean, and their variations. Models are evaluated using two-year samples. The average root-mean-squared errors (RMSE) are lowest (5.14 and 4.91), RRSE errors are highest (79.46% and 79.78%), mean absolute error (MAE) is lowest (18.12% and 19.42%), and the average correlation factor is biggest (0.41 and 0.42) exhibited by the M5-prime regression trees and KNN approaches. M5-Prime is perfect for large-scale crop yield prediction in agricultural planning since it has the most crop yield models and the fewest errors [29].

In spite of the enormous number of studies to predict crop yields using machine learning methods, there is a glaring lack with regard to the utilization of abiotic factors in particular, particularly the utilization of coffee yield prediction. Though earlier studies have examined climatic factors affecting coffee production in different countries like Brazil, Uganda, Vietnam, and Mexico, the studies rarely incorporate the localized abiotic factors, i.e., the seasonal variation in rainfall, temperature, relative humidity, and solar radiation. Additionally, the literature makes use of data sets from a wide range of geographic locations, thus making cross-regional comparisons challenging due to differences in climate, soil, and cultivation practices. To the authors' knowledge, no studies have been performed based on the Indian scenario alone, especially coffee yield prediction using local abiotic data. Additionally, there are few studies with a detailed investigation of stochastic machine learning models like Bayesian Ridge, Lasso Regression, and Random Forest for coffee yield prediction.

This research effort attempts to bridge the available literature gap by solely utilizing datasets collected from Indian sources, i.e., the Central Coffee Research Institute (CCRI) and the Coffee Research Station at Balehonnur, Karnataka, for the period 2004-2022. The primary aim is to develop and validate stochastic machine learning models to forecast coffee yield using important abiotic factors like year, rainfall, temperature, relative humidity, sunshine, vapor pressure, and dew point. By employing advanced techniques like multivariate feature selection and correlation matrix analysis, this study will nominate the most important predictors of coffee yield. Secondly, the study includes comparative analysis of a set of machine learning models—i.e., Bayesian Ridge, Lasso Regression, Elastic Net, Extra Tree, Gradient Boosting, and Random Forest—to nominate the model having maximum accuracy in the prediction of coffee production. Thirdly, the study will investigate the role of precipitation as a determining factor of predictive accuracy and its consequential impact on model performance. The aim of this research is to provide action-oriented recommendations to the coffee growers and policymakers to maximize agricultural production and enhance crop yield prediction in the Indian coffee industry.

2 Materials and Methods

2.1 Area of Study and Dataset

The period covered by the data, 2004–2022, was collected from the Central Coffee Research Institute (CCRI) in the Balehonnur, Karnataka, and district of the Karnataka State of India. The dataset for the model employed in the investigation consists of abiotic components, totaling 10 input characteristics coffee growing seventeen blocks in coffee research station, Balehonnur. Table-1 depicts the datasets and provides a detailed description of ten abiotic and a yield parameter.

Table-1: Abiotic Parameters

Parameter Name	Range [mean]	Pearson Coefficient (R)
Year	2004-2022 [13.38]	0.540
Temperature – Minimum in Degrees	12.4 - 22.8 [15.54]	0.507
Temperature Maximum in Degrees	20.6 - 29.8 [27.78]	0.205
Sunshine – Minimum in Hours	1 – 6 [2.72]	0.067
Sunshine – Maxi mum in Hours	4.5 – 8 [6.91]	0.104
Rainfall in Centimetres	186.1cm - 329.5cm [268.87]	0.267
Relative Humidity – Minimum in Percentage (%)	32% - 74% [57.11]	0.333
Relative Humidity – Maximum in Percentage (%)	83% - 100% [94.11]	0.018
Vapour Percentage (VP) in Percentage (%)	16.7% - 25.63% [19.43]	0.266
Dew Point in Degrees	14.8 - 21.0 [16.85]	0.279

2.2 Proposed Methodology

Stochastic machine learning methods are used to develop and predict coffee production using abiotic factors. The preferred procedure is illustrated in Figure-1.

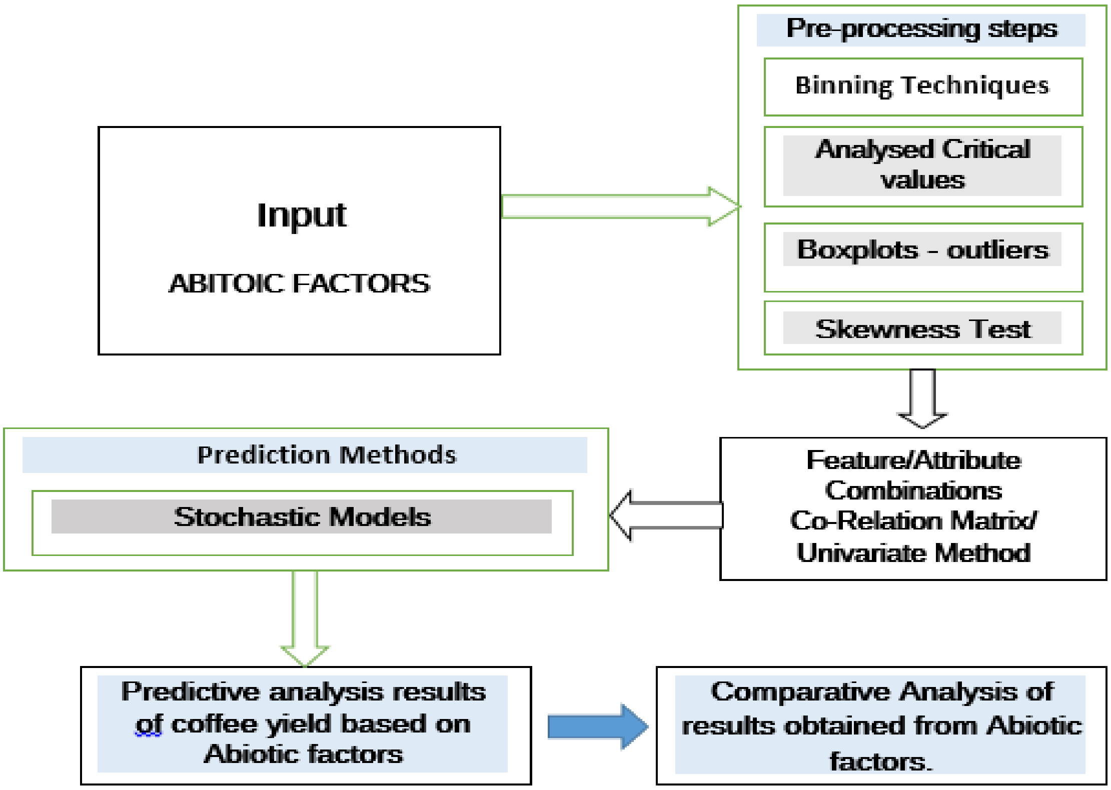

Input
ABITOIC FACTORS
Pre-processing steps
Binning Techniques
Analysed Critical values
Boxplots - outliers
Skewness Test
Feature/Attribute Combinations Co-Relation Matrix/ Univariate Method
Prediction Methods
Stochastic Models
Predictive analysis results of coffee yield based on Abiotic factors
Comparative Analysis of results obtained from Abiotic factors.

Figure 1. Block diagram of the proposed approach for predicting coffee yield based on abiotic factors

Applying Pearson's coefficient function and producing a correlation matrix with the Python library helps to check the positively correlated features towards yield. The most affecting features/predictable variables for yield were determined using a multivariate feature selection technique, which selects the best features based on statistical tests. For pre-processing an estimator, we have used selectkbest to eliminate all except the K highest scoring characteristics for prediction in univariate analysis. We have used box plots to express input features and removed outliers. A normal distribution, unaffected by skewness or outliers, is assured by the normality test. In continuation, a correlation matrix is generated to understand the correlations between input variables and the importance of features for coffee yield prediction.To pick the most prominent characteristics and increase yield, 1024 input feature combinations were produced and evaluated using univariate and multivariate feature selection techniques. After Python's feature significance function and correlation matrix output analysis, we chose three best feature groups, which were grouped into group-1, group-2, group-3, group-4 and group-5 input parameters and utilized to create the prediction model to assess yield output. The Correlation matrix depicts the degree and direction of relationships between numerous items as coefficients from -1 (strongly negative) to +1 (strongly positive), with zero indicating no link. After Feature Selection step, the grouped input parameters were given as inputs to proposed stochastic machine learning models to predict the actual yield based on historic data. Based on the performance metrics R^2, MAE, MSE and RMSE, models were evaluated. Comparative analysis is performed considering proposed stochastic machine learning models through scatter plot visualization based on R^2 and RMSE values.

2.3 Methods

Stochastic machine learning methods were discussed here.

2.3.1 Elastic Net (ENET) Regression: Elastic-net regression (ENET) was designed to solve complaints that the variable selection in Lasso regression was excessively data-dependent and unpredictable. Mixing Ridge and Lasso regression penalties yielded the best results. Ridge and Lasso algorithms provide a convex mixture of regressions in ENET [30]. It is seen here how the ENET uses C1 (lasso) and C2 (ridge regression) retributions:

Stage-1: C2_retribution = total of m beta_s2 from s=0 to m

Penalizing a model by its total absolute coefficient values is another common punishment. C1 retaliation mentions this. Low coefficient sizes and 0 coefficients remove the predictor from the model in C1 punishment.

Stage-2: Let's assume that C1_retribution = m-abs (alpha_s)*sum(s=0)*m.

The elastic net penalised linear regression model trains using C1 and C2 penalties. According to "The Elements of Statistical Learning," we use "beta" to set the weighted average of C1 and C2 retributions. C2 punishment weighting is -1 less than beta.

Stage-3: HereC2_retribution + (1 - beta * C1) = elastic_net_retribution (1 - beta * C1).

A loss function with an alpha of 0.5 would divide pay-outs evenly. When beta = 0, C2 vengeance is full, and when beta = 1, C1 retribution is full.

2.3.2 Lasso Regression: LASSO regression uses probabilistic classifiers. Lasso regularizes and selects variables. Absolute model parameters must be below Lasso's upper limit. The variable selection technique preserves non-zero coefficient variables after dimensionality reduction, thereby minimizing prediction errors [21].

In the variable selection technique, the tuning parameter χ determines how severe the penalty will be. As the parameter χ increases, more coefficients become zero, reducing variables. Increased tuning parameter χ causes bias and volatility, necessitating trade-offs. The Lasso eliminates response variable-unrelated features to reduce over fitting and increase model clarity.

Lasso solves this regularisation problem. Lasso solves the problem for a nonnegative χ value. Applying this formula:

$$\sum_{i=1}^{n}(E_i - \sum_j F_{ij}\beta_j)2 + \chi \sum_{j=1}^{p} \mid \beta_j \mid \ldots\ldots \text{(i)}$$

Where,

- The amount of shrinkage is determined by a nonnegative regularization parameter with a single value of χ.
- When χ equals zero, all characteristics are included; this is analogous to linear regression, which constructs a prediction model with just the residual sum squares.
- When χ is equal to zero, no feature is considered; as approaches infinity, more and more features are dropped.
- The level of bias increases as increases.
- As χ increases, so does prejudice.
- Higher variance is associated with lower χ values.
- The answer to issue (i) is E_i.
- N represents the total number of readings.
- The data (F_i) for observation i is a vector of q values.
- $The\ q - vector\ symbol\ is, \beta$.
- When χ increases, the percentage of non-zero components in β decreases.

2.3.3 Bayesian Ridge Regression: Bayesian regression uses Gaussian probability distribution, C2 regularisation, and posterior prediction optimization. The weighted coefficient's spherical Gaussian makes Bayesian and Bayesian Ridge regressions different. This research restricts and ranks inputs by prediction system value using Bayesian Ridge Regression. Predicts by adding coefficients to the weighted total. These coefficients picked the best features from each omits data set as feature significance ratings. Using the formula below:

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \varepsilon \ldots (\, ii \,)$$

Outliers or random sampling noise impact the model, resulting in x as a predictor, y as a response, c as a coefficient, and e as an error term.

2.3.4 Random Forest Regression: Ensemble-based data mining like Random Forest (RF) makes accurate predictions without over fitting. They use model aggregation-based learning [31]. Random forests combine binary decision trees, bootstrapped learning samples, and a random explanatory variable collection. In a third of samples, the RF method builds up to 2000 random trees using validation or "out-of-bag" predictions. Each tree is bootstrapped [31]. The Random Forest algorithm is carried out in a sequential manner as:

Step-1: Randomly choose and replace a selection of N training set data cases. The training process involves growing the original trees.

Step-2: Random Forest randomly selects m variables from M inputs (or predictors) at each node.

Step-3: The Random Forest approach grows each tree to its greatest size without trimming its structure.

Step-4: Pooling n trees' predictions provides a mean value for forecasting new data if there is a regression issue.

2.3.5 Extra Tree Regression: This new RF model uses "Extra Tree Regression". ETR creates unpruned judgments or regression trees. Bagging and bootstrapping are used in RF regression [32].

The instructions below show how to use the Extra Tree Regression Algorithm for numerical characteristics [32]:

Step-1: Dividing a node (A)

The learning subset A for the neighboring node that has to be divided is the input.

A node split $[x\ xy]$ or a zero split is the output.

Return 0 if Stop split (A) equals.

Anyhow, from all non-Consistent (in A) Applicant characteristics, choose

B attributes ($c_1 \ldots c_b$);

Describe the locations of the B divides $e_1 \ldots e_b . e_i$ = choose a random number

Splice $(A, x_i),\ v_j$ = 1, and then B;

To ensure that Count ($d *, A$) = max_i=1... B Count (e_i, A), return a split $e *$.

Step-2: Take a haphazard split. (A, x)

Inputs include an attribute x and x subclass A.

Results: x split

Let x_{Amax} and x_{Amin} denote the maximum and insignificant estimates of m in A, respectively.

Illustrate any boundary ac in accordance with $[x_{Amin}, x_{Amax}]$.

Send the split ($[x\ xy]$) back.

Step-3: **Reverse split (A)**

Enter: x subclass A binary output for x return TRUE if $|A| x_{min}$; Return TRUE if$A$'s properties are all consistent.

Return TRUE if the result is in accordance with A; FALSE otherwise.

The steps above describe the Extra-Trees splitting approach for numerical features.

2.3.6 Gradient Boosting Regression: In ensemble learning [33], gradient boosting regression tree approaches employ weak learner regression trees (decision trees) to create reliable forecasting models. Poorly trained models (regressors or classifiers) experience fewer errors (Singh et al., 2021) [34].

The gradient boosting tree, also known as the $f_n(x_t)$ algorithm, is the accumulation of n regression trees:

$$f_n\ (x_t)\ =\ \sum_{n\ i=1} f_i\ (x_t)\ \ ---------(iii)$$

Every$f_i(x_t)$ is a decision tree or regression. The equation estimates the new decision tree $f_{n+1}(x_t)$ to form the ensemble of trees:

$$\text{argmin} \sum\nolimits_{\text{t}} \text{L}(\text{y}_t\ .f_n(x_t)\ +\ f_{n+1}(x_t))\ \text{---}\ (\text{iv})$$

Where the loss function L (.) is differentiable. The steepest descent method solves this optimization. This study employed 0.2 learning rate and 100 estimator value. When learning rate is smaller, stopping before over fitting is easier.

2.4 Predictive Model Development:

All Six models—BRR, Lasso, ETR, BRR, Enet and RFR —were created on Windows 11 using Collab and Jupyter on an Intel core i7 laptop. To predict coffee yield (Y) in Central Coffee Research Institute (CCRI), coffee research station at Chikkamagaluru, the BRR model used patterns embedded in the K (=10) lots of Abiotic factors data matrix from Table-1 and its relationship with the objective variable, Y. Compared to all six models. This paper uses a cross correlation analysis between Xk (Input Parameters like Group-1, Group-2, Group-3, Group-4, and Group-5) and Y to examine the links between each component and coffee yield data.

For model creation, evaluation/selection, and testing, measured data was independently split into two sets: training () and testing (). Table 2 lays out the descriptive statistics of the Abiotic factors model's input parameters for the training and testing phases The most affecting features/predictable variables for yield were determined using feature important function using co-relation matrix and multivariate feature selection technique, which selects the best features based on multivariate statistical tests. For Pre-processing of an estimator we have used selectkbest to eliminate all except the K highest scoring characteristics for prediction in univariate analysis.

A total of 1024 distinct combinations of input features and total six models were employed in this study, with varying R² values for each group of feature combinations. A total of five ideal BRR models were created, taking into consideration the expected contributions of each of the abiotic characteristics to the estimated overall coffee production which are labelled as group-1, group-2 group-3, group-4 and group-5 taking into account the K feature values as subsets as 1, 2, 3, 4,..., 10. The below table-2 gives the detail information about selected input subsets for the prediction.

Table-2: Optimal Feature Subsets for Prediction Models

Models	Features Selected	Number of Features
1	Year + Rainfall + Temperature Minimum & Maximum, Sunshine Minimum & Maximum + Relative Humidity Minimum & Maximum + Vapor + Dew Point Vs. Yield,	10
2	Year + Rainfall + Temperature Minimum & Maximum, Sunshine Minimum & Maximum Vs. Yield	06
3	Year + Relative Humidity Minimum and Maximum + Rainfall + Minimum and Maximum Temperature Vs. Yield	06
4	Year + Rainfall + Temperature Minimum & Maximum + Vapor Vs. Yield	05
5	Year + Rainfall + Temperature Minimum & Maximum + Dew Point Vs. Yield	05

In order to construct its ensemble of decision trees, the ETR model used an n_estimaters (number of trees=1000) technique. To provide an unbiased comparison to the BRR model, it regressed the exploratory and response correlations between the data on coffee production and the abiotic factors measured in the testing phase. By iteratively trying out ensembles with numbers ranging from 100 to 600, in one-fold increments, we were able to maximize n_estimator, a critical parameter. Here, we found that 500 trees, with leaf=1 and fboot=1, produced the best ETR model.

To further compare, RF and GBR models were created for the same set of predictors (group-1, group-2, 3, 4 and group-5). Interestingly, the BRR model from group 3 performed better than the BRR models from groups 1 and 2. Additionally, when compared to other models built utilizing Groups 1, 2, and 3, the Best BRR model from Group 3 performed well.

Here, n is the total number of data points used in each prediction matrix for training and testing. Each trial's R² and RMSE were tracked according to the objective criteria, and the models were assessed using the Testing Datasets. The six model design parameters used in this investigation are shown in Table-3.

Table-3: Model Design Parameters for Bayesian Ridge, Lasso, Extra Tree, Gradient Boosting, Random Forest, and Elastic Net Regression based on Feature Groups

Model	Design Parameters	State of parameters in study
BRR	max_iter	int, default=300
	tol	float, default=1e-3
	alpha_1	float, default=1e-6
	alpha_2	float, default=1e-6
	alpha_init	float, default=None
	lambda_1	float, default=1e-6
	lambda_2	float, default=1e-6
	lambda_init	float, default=None
	compute_score	bool, default=False
	fit_intercept	bool, default=True
	verbose	bool, default=False
	coef_	array-like of shape (n_features,)
	n_features_in_	int
	feature_names_in	ndarray of shape (n_features_in_,)
	sigma_	array-like of shape (n_features, n_features)

	scores_	array-like of shape (n_iter_+1,)
ETR	n_estimaters (Number of Trees) Min_sample_split Min_sample_leaf Max_features Random_state	100,200,300,400,500,600,700,800 (Optimal value = 600) 5 , Int, Default = 2 Int, Default=1 Int, Group-1,2,3 subsets , Default =1.0 Int, Default = none.
GBR	n_estimaters (Number of Trees) Loss Learning Rate Sub Sample Criterion Min_sample_split Min_sample_ leaf Max_depth Random_State Max_Features	100,200,300,400,500,600,700,800 (Optimal value = 600) Default: Squared Error. Float , Default = 0.1 Float , Default = 1.0 Default : 'Friedman_mse' Int, Default = 2 Int, Default = 1 Int/None, Default = 3 Int, Default = none Int, Group-1, 2, 3 subsets, Default =None.
RFR	n_estimaters (Number of Trees) Max_leaf_node Criterion Max_depth Min_sample_Split Min_sample_leaf Max_features Max_leaf_node Bootstrap Surrogate	50,100,150,200,250,300,350,400 (Optimal Value = 200) Int, Default = 5 Default = 'Squarred Error' Int, Default = none Int, Default = 2 Int, Default =1 Int, Default = 1, Group-1, 2, 3 Subsets. Int, Default = none Bool, Default = True. On, Sample with replacement.
Lasso	alpha fit_intercept : normalize : copy_X : precompute : max_iter : tol : warm_start : positive intercept_ :	: float, optional Boolean boolean, optional, default False boolean, optional, default True True | False | 'auto' | array-like int, optional float, optional bool, optional : bool, optional float | array, shape = (n_targets,)
Enet	Alpha L1_ratio fit_intercept precompute max_iter tol warm_start positive random_state selection	float, default=1.0 float, default=0.5 bool, default=True bool or array-like of shape (n_features, n_features),default=False int, default=1000 float, default=1e-4 bool, default=False bool, default=False int,RandomState instance, default=None

		{‘cyclic’, ‘random’}, default=’cyclic’

2.5 Model Performance Evaluation Metrics

The study carried out compared measured yield data with anticipated yield data from the test phase to evaluate six stochastic models for coffee yield prediction. Analyzed R², mean absolute error, mean square error, root MSE. The core performance measures in Table-4 forecast production:

Table-4: Standard Performance Metrics

Sl. No	**Performance Metrics**	**Formula**
1	R-Squared (R^2)	$R^2\ Squared\ =\ 1\ -\ \frac{SSr}{SSm}$ SSr – Squared regression line sum error SSm - Squared mean line sum error.
2	Mean Absolute Error (MAE)	$\text{MAE} = \frac{1}{N}\sum_{i=1}^{i=N}(\lvert y_i - \hat{y}_i \rvert)$ There are N anticipated values. $T\text{he ith data}'\text{s real true value is represented by yi.}$ $\hat{\text{y}}_\text{i}$is the i-th data's anticipated value.
3	MSE, or Mean Squared Error	$MSE\ =\ \frac{1}{N}\sum(y_i - \hat{y}_i\,)^2$
4	RMSE, or Root Mean Square Error	$RMSE\ =\ \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\left(y_i - \hat{y}_i\right)^2}$

3 Results

During testing, scatterplots (fig. 3) were utilized to visually analyze the degree of agreement between actual yield and projected yield data to simplify all characteristics to a single scale without altering the stochastic models' range of values. In this study, we employed ordinary scalar fine tuning method and features were selected based on multivariate feature selection approach based on feature importance function along with co-relation matrix. Later , for prediction model developments five optimal models were chosen and categorized into five separate sets of parameters when analyzing and building the model: ***Group-1 parameters:*** *Year + Rainfall + Temperature Minimum & Maximum, Sunshine Minimum & Maximum + Relative Humidity Minimum & Maximum + Vapor + Dew Point Vs. Yield****, Group-2 parameters:*** *Year + Rainfall + Temperature Minimum & Maximum, Sunshine Minimum & Maximum Vs. Yield****, Group-3 parameters:*** *Year + Relative Humidity Minimum and Maximum + Rainfall + Minimum and Maximum Temperature Vs. Yield****, Group-4 parameters:*** *Year + Rainfall + Temperature Minimum & Maximum + Vapor Vs. Yield****, Group-5 parameters:*** *Year + Rainfall + Temperature Minimum & Maximum + Dew Point Vs. Yield****.*** With the use of scatterplots and Figure-2 displays performance measurements R-Square (R²) and RMS Error (RMSE), the actual and expected yield, and model error rate.

3.1 Bayesian Ridge Regression: To compare Bayesian Ridge regression to other models in the present study, 10,000-fold cross-validation was utilized. The outcome was more accurate and had less errors than the five models. Overall consistency testing for split sizes is in Table-5. Year, rainfall, lowest and highest temperatures, and shortest and longest sunlight days are most essential. Highest and lowest relative humidity Vapour and Dew Point, Five types of model findings exist.

Table 5: Performance evaluation of the Bayesian Ridge Regression model during the testing phase, showing forecasted coffee yield for five groups with varying abiotic parameters. Performance metrics include R^2, MAE, MSE, and RMSE

<table>
<tr><td colspan="5">Group-1: All Parameters (Year + Rainfall + Temperature Min Max + Sunshine Min Max + Relative Humidity Min Max + Vapor + Dew Point) Vs. Yield</td></tr>
<tr><td>Quantity Shared</td><td>R^2</td><td>M A E
kg per ha</td><td>M S E
kg per ha</td><td>R M S E kg per ha</td></tr>
<tr><td>90:10</td><td>1.00</td><td>7.19</td><td>54.40</td><td>7.38</td></tr>
<tr><td>80:20</td><td>0.76</td><td>65.65</td><td>5371.34</td><td>73.29</td></tr>
<tr><td>70:30</td><td>0.64</td><td>78.06</td><td>7068.31</td><td>84.07</td></tr>
<tr><td>60:40</td><td>0.75</td><td>53.43</td><td>4018.23</td><td>63.39</td></tr>
<tr><td colspan="5">Group 2 of the Bayesian Ridge Model: Year + Rainfall + Minimum and Maximum Temperature and Sunshine vs. Yield</td></tr>
<tr><td>Quantity Shared</td><td>R^2</td><td>M A E
kg per ha</td><td>M S E
kg per ha</td><td>R M S E kg per ha</td></tr>
<tr><td>90:10</td><td>0.80</td><td>62.07</td><td>6136.41</td><td>78.34</td></tr>
<tr><td>80:20</td><td>0.61</td><td>91.43</td><td>8777.17</td><td>93.69</td></tr>
<tr><td>70:30</td><td>0.69</td><td>68.15</td><td>6041.25</td><td>77.73</td></tr>
<tr><td>60:40</td><td>0.75</td><td>53.37</td><td>4027.28</td><td>63.46</td></tr>
<tr><td colspan="5">Group 3 of the Bayesian ridge model: Year + Rainfall + Minimum Maximum Temperature + Relative Humidity Min Max Vs. Yield</td></tr>
<tr><td>Quantity Shared</td><td>R^2</td><td>M A E
kg per ha</td><td>M S E
kg per ha</td><td>R M S E kg per ha</td></tr>
<tr><td>90:10</td><td>0.89</td><td>39.98</td><td>3222.75</td><td>56.77</td></tr>
<tr><td>80:20</td><td>0.78</td><td>65.75</td><td>4924.62</td><td>70.18</td></tr>
<tr><td>70:30</td><td>0.81</td><td>54.81</td><td>3784.57</td><td>61.52</td></tr>
<tr><td>60:40</td><td>0.74</td><td>56.72</td><td>4123.13</td><td>64.21</td></tr>
<tr><td colspan="5">Group-4 of the Bayesian ridge model: Year + Rainfall + Minimum Maximum Temperature + Vapor vs. Yield</td></tr>
<tr><td>Quantity Shared</td><td>R^2</td><td>M A E
kg per ha</td><td>M S E
kg per ha</td><td>R M S E kg per ha</td></tr>
<tr><td>90:10</td><td>0.83</td><td>51.32</td><td>5347.65</td><td>73.13</td></tr>
<tr><td>80:20</td><td>0.72</td><td>71.80</td><td>6413.55</td><td>80.08</td></tr>
<tr><td>70:30</td><td>0.74</td><td>62.41</td><td>5062.73</td><td>71.15</td></tr>
<tr><td>60:40</td><td>0.75</td><td>54.92</td><td>3932.09</td><td>62.71</td></tr>
<tr><td colspan="5">Group-5 Variables (Year, Rainfall, Minimum, Maximum Temperature, and Dew Point) Used vs. Yield</td></tr>
<tr><td>Quantity Shared</td><td>R^2</td><td>M A E
kg per ha</td><td>M S E
kg per ha</td><td>R M S E kg per ha</td></tr>
<tr><td>90:10</td><td>0.81</td><td>56.26</td><td>5680.52</td><td>75.37</td></tr>
<tr><td>80:20</td><td>0.70</td><td>77.48</td><td>6810.57</td><td>82.53</td></tr>
<tr><td>70:30</td><td>0.74</td><td>64.82</td><td>5044.54</td><td>71.02</td></tr>
</table>

60:40	0.74	53.83	4094.82	63.99

From a side-by-side comparison of experimental and anticipated yield for the 70:30 split ratio, Bayesian ridge regression using group-3 parameters—year, rainfall, minimum and maximum temperatures, and relative humidity—had the strongest association (R-square = 0.81 and Root Mean Square Error = 61.52 kg per ha (figure 2)). The analysed performance indicators showed Bayesian Ridge regression model performed badly for parameters in groups 1, 2, 4, and 5. Unlike the five regression models, several inputs decreased coffee yield prediction accuracy, whereas fewer input parameters increased it. Each model split during testing has R2, MAE, MSE, and RMSE in Table-5.

3.2 Lasso Regression: The Lasso regression model was tested 10K times to see whether it enhanced consistency compared to other models. Table 6 demonstrates all parameter group split consistency. This model has five outcomes, and Group-3 variables—year + relative humidity minimum and maximum + rainfall + minimum and maximum temperature vs. yield—are most relevant.

Table 6: Performance Evaluation of the Lasso Regression Model for Different Split Ratios and Predicted Coffee Yield Across Five Groups Using Abiotic Factors, with R², RMSE, MAE, and MSE as Key Metrics

<table>
<tr><td colspan="5">Group-1Lasso Model: All Parameters (Year, Rainfall, Minimum and Maximum Temperature, Minimum and Maximum Amount of Sunshine, Minimum and Maximum Relative Humidity, Vapor Pressure, and Dew Point) vs. Yield</td></tr>
<tr><td>Quantity Shared</td><td>R²</td><td>MAE
kg per ha</td><td>MSE
kg per ha</td><td>RMSE kg per ha</td></tr>
<tr><td>90:10</td><td>1.00</td><td>5.03</td><td>25.55</td><td>5.06</td></tr>
<tr><td>80:20</td><td>0.77</td><td>65.23</td><td>5305.40</td><td>72.84</td></tr>
<tr><td>70:30</td><td>0.65</td><td>76.74</td><td>6907.54</td><td>83.11</td></tr>
<tr><td>60:40</td><td>0.41</td><td>102.34</td><td>13461.70</td><td>116.02</td></tr>
<tr><td colspan="5">Lasso Model Constructed Using Group-2 : Year + Rainfall + Temp Min Max + Sunshine Min Max Vs. Yield</td></tr>
<tr><td>Quantity Shared</td><td>R²</td><td>MAE
kg per ha</td><td>MSE
kg per ha</td><td>RMSE kg per ha</td></tr>
<tr><td>90:10</td><td>0.80</td><td>61.90</td><td>6003.11</td><td>77.48</td></tr>
<tr><td>80:20</td><td>0.61</td><td>91.52</td><td>8790.72</td><td>93.76</td></tr>
<tr><td>70:30</td><td>0.69</td><td>68.24</td><td>6059.05</td><td>77.84</td></tr>
<tr><td>60:40</td><td>0.75</td><td>53.43</td><td>4018.23</td><td>63.39</td></tr>
<tr><td colspan="5">Lasso Model Constructed Using Group-3 : Year + Rainfall + Temp Min Max + Relative Humidity Min Max Vs. Yield</td></tr>
<tr><td>Quantity Shared</td><td>R²</td><td>MAE
kg per ha</td><td>MSE
kg per ha</td><td>RMSE kg per ha</td></tr>
<tr><td>90:10</td><td>0.90</td><td>37.92</td><td>2992.74</td><td>54.71</td></tr>
<tr><td>80:20</td><td>0.78</td><td>66.20</td><td>4931.91</td><td>70.23</td></tr>
<tr><td>70:30</td><td>0.80</td><td>55.60</td><td>3869.30</td><td>62.20</td></tr>
<tr><td>60:40</td><td>0.74</td><td>57.65</td><td>4167.51</td><td>64.56</td></tr>
<tr><td colspan="5">Lasso Model Constructed Using Group-4 : Year + Rainfall + Temp Min Max + Vapor Vs. Yield</td></tr>
<tr><td>Quantity Shared</td><td>R²</td><td>MAE
kg per ha</td><td>MSE
kg per ha</td><td>RMSE kg per ha</td></tr>
<tr><td>90:10</td><td>0.84</td><td>49.29</td><td>5041.77</td><td>71.01</td></tr>
</table>

80:20	0.72	71.64	6398.22	79.99
70:30	0.74	62.43	5083.55	71.30
60:40	0.75	55.07	3924.01	62.64
Lasso Model Constructed Using Group-5 : Year + Rainfall + Temp Min Max + Dew Point Vs. Yield				
Quantity Shared	R^2	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.82	55.91	5528.22	74.35
80:20	0.70	77.58	6762.33	82.23
70:30	0.75	65.05	5041.00	71.00
60:40	0.74	54.14	4084.65	63.91

Group-3-parameter lasso regression model (Year, Relative Humidity, Rainfall, and Temperature) yields the greatest determination constant (R-square = 0.80 and Root Mean Square Error = 62.20 kg per ha) (fig. 2). Compare the observed yield to the projected yield for a 70:30 splitting fraction. Even if lasso regression failed, groups 1, 2, 4, and 5 parameters. These models fail to extract because the Bayesian Ridge regression model's coffee yield forecast only little reduced with the same inputs and barely changed with group-3 parameters, unlike the Lasso regression model. Table-6 displays the testing results of divides models using R^2, MAE, MSE, and RMSE.

3.3 Elastic Net Regression: We tested elastic net regression's consistency compared to other probabilistic models with a 10Kfold increase. Table-7 shows overall split consistency, with notable variances among parameter groups.

Table 7: Performance of Elastic Net Regression in forecasting coffee yield using abiotic factors across five groups, with testing phase results for different split ratios and performance metrics (MAE, MSE, RMSE)

Group-1 Elastic Net Model: Year + Rainfall + Temp Min Max + Sunshine Min Max + Relative Humidity Min Max + Vapor + Dew Point vs. Yield				
Quantity Shared	R^2	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.71	91.91	8454.73	91.95
80:20	0.78	67.68	4989.60	70.64
70:30	0.72	71.09	5858.16	76.54
60:40	0.48	84.36	8365.08	91.46
Elastic Net Model Constructed Using Group-2 : Year + Rainfall + Temp Min Max + Sunshine Min Max Vs. Yield				
Quantity Shared	R^2	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.65	104.07	10852.35	104.17
80:20	0.73	68.89	6050.81	77.79
70:30	0.64	73.47	7038.19	83.89
60:40	0.44	87.21	8919.88	94.45
Elastic Net Model Constructed Using Group-3 : Year + Rainfall + Temp Min Max + Relative Humidity Min Max Vs. Yield				
Quantity Shared	R^2	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.73	91.59	8396.96	91.63
80:20	0.78	67.26	4914.64	70.10

70:30	0.69	70.67	5772.07	75.97
60:40	0.48	83.97	8287.57	91.04
Elastic Net Model Constructed Using Group-4 : Year + Rainfall + Temp Min Max + Vapor Vs. Yield				
Quantity Shared	R^2	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.64	104.80	11005.62	104.91
80:20	0.73	69.85	6228.52	78.92
70:30	0.64	74.06	7161.53	84.63
60:40	0.43	87.72	9030.26	95.03
Elastic Net Model Constructed Using Group-5 : Year + Rainfall + Temp Min Max + Dew Point Vs. Yield				
Quantity Shared	R^2	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.65	104.07	10852.35	104.17
80:20	0.73	68.89	6050.81	77.79
70:30	0.65	73.41	7017.06	83.77
60:40	0.44	87.19	8907.59	94.38

Visually comparing testing procedures. Group 1's elastic net regression model (All parameters vs. Yield) had the highest constant of determination for the 70:30 splitting proportion, using Year, Rainfall, Temperature, Sunshine, Relative Humidity, Vapour, and Dew point as predictor variables (R-square = 0.72 and Root Mean Square Error = 76.54 kg per ha) (fig 2 Performance measurements showed that the Elastic Net Regression model failed for group-2, 3, 4, and 5 parameters. Other inputs helped the Elastic Net regression model forecast coffee yield, indicating that parameter data can't predict. Table-7 compares testing efficiency for several splitting models using R^2, MAE, MSE, and RMSE.

3.4 Random Forest Regression: We tested the random forest model with tens of thousands of samples to determine whether it improved reliability over regression models. Table-8 indicates significant parameter group differences in splitting.

Table 8: Performance evaluation of the Random Forest Regression model for coffee yield prediction using abiotic factors across five groups, with testing phase results including R^2, MAE, MSE, and RMSE

Group-1 Random Forest Model: Year + Rainfall + Temp Min Max + Sunshine Min Max + Relative Humidity Min Max + Vapor + Dew Point vs. Yield				
Quantity Shared	R^2	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.78	60.98	6743.17	82.12
80:20	0.53	91.37	10612.55	103.02
70:30	0.53	81.41	9287.23	96.37
60:40	0.22	94.91	12445.00	111.56
Random Forest Model Constructed Using Group-2 : Year + Rainfall + Temp Min Max + Sunshine Min Max Vs. Yield				
Quantity Shared	R^2	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.78	61.30	6660.84	81.61
80:20	0.53	91.91	10738.22	103.63
70:30	0.52	84.15	9474.17	97.34

60:40	0.25	92.75	12030.19	109.68
Random Forest Model Constructed Using Group-3 : Year + Rainfall + Temp Min Max + Relative Humidity Min Max vs. Yield				
Quantity Shared	R^2 kg per ha	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.78	60.89	6708.19	81.90
80:20	0.53	91.77	10701.43	103.45
70:30	0.52	82.54	9493.26	97.43
60:40	0.34	88.38	10512.33	102.53
Random Forest Model Constructed Using Group-4 : Year + Rainfall + Temp Min Max + Vapor vs. Yield				
Quantity Shared	R^2	M A E	M S E kg per ha	R M S E kg per ha
90:10	0.78	61.28	6779.85	82.34
80:20	0.51	93.16	11029.44	105.02
70:30	0.50	86.64	9950.11	99.75
60:40	0.34	89.06	10522.07	102.58
Random Forest Model Constructed Using Group-5 : Year + Rainfall + Temp Min Max + Dew Point vs. Yield				
Quantity Shared	R^2	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.78	60.32	6721.96	81.99
80:20	0.53	91.82	10714.50	103.51
70:30	0.52	83.84	9429.81	97.11
60:40	0.37	86.57	10017.71	100.09

Graphically comparing test protocols. All factors vs. yield in Group-1 of the random forest regression model predicted yield, including year, rainfall, lowest and highest temperatures, and shortest and longest sunlight days. The lowest and greatest relative humidity The Vapour and Dew Point In Figure 2, the 70:30 percentage ratio had the greatest coefficient of determination for predictor factors (R-square= 0.53 and Root Mean Square Error = 96.37 kg per ha). For groups-2, 3, 4, and 5 parameters, random forest regression model fared poorly in all key aspects. R^2, MAE, MSE, and RMSE for several splitting models tested appear in Table-8.

3.5 Extra Tree Regression: Extremely Randomized Trees (Extra Trees) is a regression like Random Forest. Feature splits are random. Mixing the test and train datasets, the Extra tree regression model was created using 100 DTs from diverse informative categories.

Table 9: Performance evaluation of the Extra Tree Regression model for coffee yield prediction using variable abiotic factors across five groups, with results for different data split ratios. Performance indicators include R^2, MAE, MSE, and RMSE

Group-1 Extra Tree Model: All Parameters(Year + Rainfall + Temp Min Max + Sunshine Min Max + Relative Humidity Min Max + Vapor + Dew Point) vs. Yield				
Quantity Shared	R^2	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.78	59.79	6860.77	82.83
80:20	0.46	94.86	12221.13	110.55
70:30	0.47	93.29	10520.19	102.57

60:40	0.34	91.67	10518.58	102.56
Extra Tree Model Constructed Using Group-2 : Year + Rainfall + Temp Min Max + Sunshine Min Max vs. Yield				
Quantity Shared	R^2	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.79	57.19	6275.86	79.22
80:20	0.38	103.37	14183.45	119.09
70:30	0.44	88.62	11097.62	105.35
60:40	0.27	95.50	11631.80	107.85
Extra Tree Model Constructed Using Group-3: Year + Rainfall + Temp Min Max + Relative Humidity Min Max vs. Yield				
Quantity Shared	R^2	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.78	58.80	6635.44	81.46
80:20	0.43	99.66	12998.17	114.01
70:30	0.53	88.25	9360.95	96.75
60:40	0.37	92.00	10020.85	100.10
Extra Tree Model Constructed Using Group-4: Year + Rainfall + Temp Min Max + Vapor vs. Yield				
Quantity Shared	R^2	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.80	56.42	6107.72	78.15
80:20	0.41	99.94	13339.82	115.50
70:30	0.48	86.75	10198.86	100.99
60:40	0.32	94.32	10801.73	103.93
Extra Tree Model Constructed Using Group-5: Year + Rainfall + Temp Min Max + Dew Point vs. Yield				
Quantity Shared	R^2	M A E kg per ha	M S E kg per ha	R M S E kg per ha
90:10	0.80	56.95	6223.36	78.89
80:20	0.44	98.23	12820.74	113.23
70:30	0.51	83.17	9638.42	98.18
60:40	0.32	93.99	10827.54	104.06

The variables in Group 3 are: Season + Average + Extreme Conditions (Temperatures and RH) yields five groupings, with minimum and maximum being most relevant. The additional tree regression model with group-3 parameters (Year, Year + Relative Humidity Minimum and Maximum + Rainfall + Minimum and Maximum Temperature) had the best coefficient of determination (R-square = 0.53 and Root Mean Square Error = 96.75 kg per ha). Comparing measured yield to expected yield for 70:30 split ratio showed this. Performance indicators showed the extra tree regression model performed badly for groups 1, 2, 4, and 5 parameters. Using the same inputs and group-3 parameters, the Bayesian Ridge regression model and Lasso regression model fail to extract predictive features from multi-parameter data of coffee yield. Table-9 displays R^2, MAE, MSE, and RMSE results for several splitting models during testing.

3.6 Gradient Boosting Regression: Here, Boosting is carried out using 100 weak learners, and the base_estimator is conceptualized as a random forest. Weak learners are increased at every step to make up for the weak learners already there. Gradients are used to identify the combined model's flaws.

Table 10: Performance evaluation of the Gradient Boosting Regression model for coffee yield prediction using abiotic factors across five groups, with testing phase results at different split ratios. Performance metrics include R², MAE, MSE, and RMSE

Group-1 Gradient Boosting Model: All Parameters (Year + Rainfall + Temp Min Max + Sunshine Min Max + Relative Humidity Min Max + Vapor + Dew Point) vs. Yield				
Quantity Shared	R²	MAE kg per ha	MSE kg per ha	RMSE kg per ha
90:10	0.78	59.82	6867.02	82.87
80:20	0.07	121.14	21165.33	145.48
70:30	0.40	92.62	11919.26	109.18
60:40	0.26	93.73	11752.02	108.41
Gradient Boosting Model Constructed Using Group-2: Year + Rainfall + Temp Min Max + Sunshine Min Max vs. Yield				
Quantity Shared	R²	MAE kg per ha	MSE kg per ha	RMSE kg per ha
90:10	0.76	62.36	7461.82	86.38
80:20	0.40	102.36	13569.17	116.49
70:30	0.38	96.06	12251.49	110.69
60:40	0.22	94.46	12438.62	111.53
Gradient Boosting Model Constructed Using Group-3: Year + Rainfall + Temp Min Max + Relative Humidity Min Max vs. Yield				
Quantity Shared	R²	MAE kg per ha	MSE kg per ha	RMSE kg per ha
90:10	0.76	61.73	7311.97	85.51
80:20	0.38	103.76	14048.69	118.53
70:30	0.46	87.84	10649.94	103.20
60:40	0.39	87.27	9764.08	98.81
Gradient Boosting Model Constructed Using Group-4: Year + Rainfall + Temp Min Max + Vapor vs. Yield				
Quantity Shared	R²	MAE kg per ha	MSE kg per ha	RMSE kg per ha
90:10	0.73	66.18	8405.47	91.68
80:20	0.43	99.33	12941.39	113.76
70:30	0.42	92.73	11408.94	106.81
60:40	0.47	78.54	8392.54	91.61
Gradient Boosting Model Constructed Using Group-5: Year + Rainfall + Temp Min Max + Dew Point vs. Yield				
Quantity Shared	R²	MAE kg per ha	MSE kg per ha	RMSE kg per ha
90:10	0.79	58.10	6476.56	80.48
80:20	0.40	101.80	13587.32	116.56
70:30	0.45	89.15	10795.64	103.90
60:40	0.43	82.08	9036.93	95.06

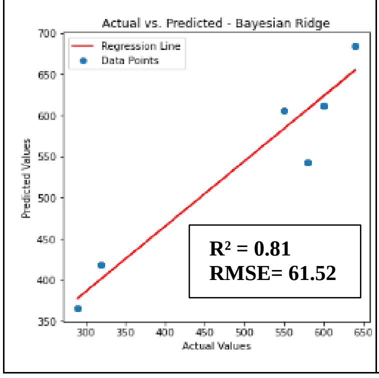
Each parameter group had distinct results, and Table-8 indicates how consistently each split performed. This model's most relevant variables were group-3: Year + Relative Humidity Minimum and Maximum + Rainfall + Temperature vs. Yield. Model findings fall into five groups. The gradient boosting regression model using Year, Rainfall, low and high temperatures, and relative humidity as predictor variables has the greatest coefficient of determination (R-square = 0.46 and Root Mean Square Error = 103.20 kg per ha) employing all group-3 factors (Figure-3). Comparing measured yield to expected yield for 70:30 split ratio showed this. Table-10 examines the performance of the 70:30 split model testing phase, including R², MAE, MSE, and RMSE.

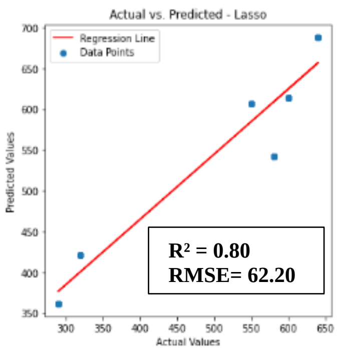
Table-11 shows the predicted coffee yield data with respect to the model error variance in maximum, minimum, skewness, kurtosis, standard deviation, p25, p50 (median), and p75 for all six methods for the five optimal input combinations (groups 1, 2, 3, 4, and 5. Overall expected yield for ideal testing phase. Similar conclusions were derived using the minimum, standard deviation, and quartiles. Two deductions possible:

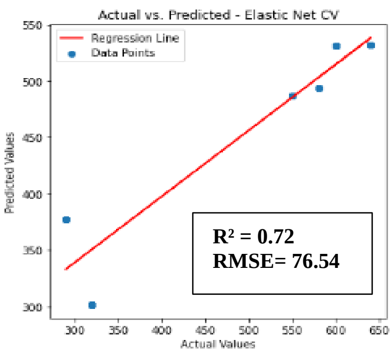
i. A comprehensive and robust statistical dependency analysis of inputs and the target variable must choose the most essential input variables to forecast coffee crop production based on abiotic properties (table 1); and
ii. Bayesian ridge models outperformed with a R² = 0.81 compared to Lasso with a R²= 0.80, Elastic net with a R²=0.72, Random forest with a R²= 0.53, Extra Tree with a R²= 0.53, and Gradient Boosting with a R²= 0.46 models in forecasting coffee yields and identifying abiotic factor correlations.

Rainfall models have revealed that this abiotic component affects coffee crop prediction. Effective variable selection enhances simulation accuracy. Bayesian ridge model predictors for 10 abiotic factors were relevant to coffee yield estimate.

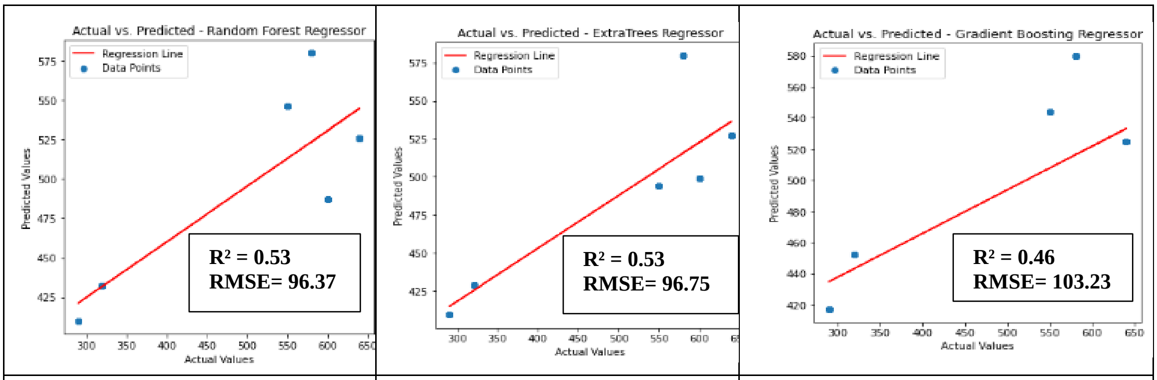
Figure-3 compares expected and actual coffee yields. Bayesian ridge models had a minor variation between observed and forecast coffee yield, whereas Lasso, ENet, ETR, RFR, and GBR models had significantly larger disparities (table-11). BRR models had relatively tiny disparities between the top and lower quartiles of actual data, but the higher quartile was below forecasted for all 5 BRR optimal models. For BRR models, actual and projected bottom quartiles were similar. Other models in this study demonstrated consistent yield under forecast.

<table>
<tr><th>Bayesian Ridge Model</th><th>Lasso Model</th><th>Elastic Net Model</th></tr>
<tr>
<td>


</td>
<td>


</td>
<td>


</td>
</tr>
</table>



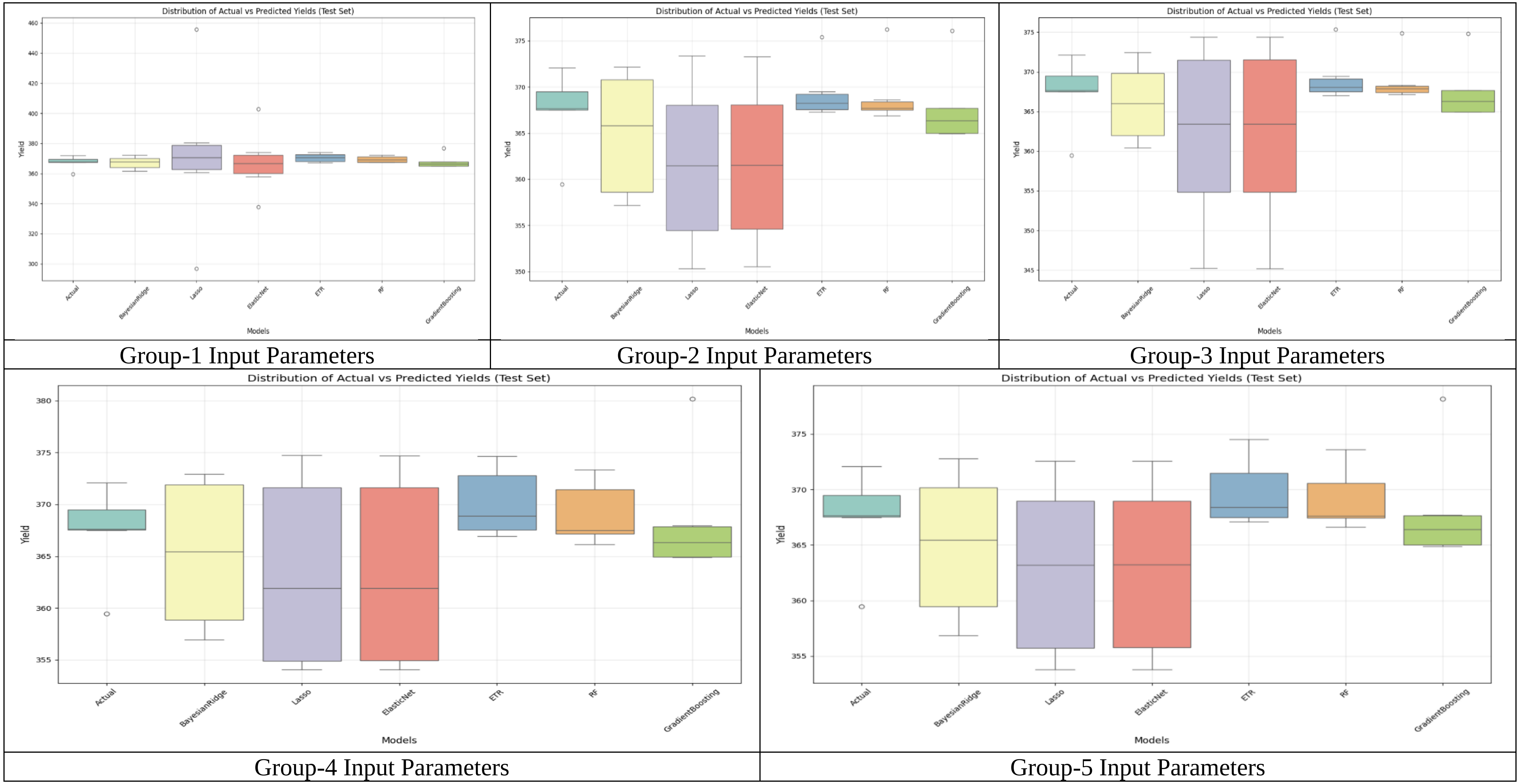
Random Forest Model | Extra Tree Model | Gradient Boosting Model

Figure 2. Group -3 served as input parameters for the following models shown in above scatterplots based on stochastic models based on the Bayesian ridge, lasso, elastic net, random forest, extra tree, and gradient boost algorithms. Above are scatterplots depicting the highest expected and actual coffee yields during the experimentation period (70:30 Splits).

Table 11: Statistical summary of model performance errors during the testing period for five input combinations, presented as absolute values in kilograms per hectare

Multivariate distribution statistics for performance error	Model and Input Parameters (Group – 1,2,3,4 and 5)														
	BRR					Lasso					Enet				
	Group-1	*Group-2*	*Group-3*	*Group-4*	*Group-5*	*Group-1*	*Group-2*	*Group-3*	*Group-4*	*Group-5*	*Group-1*	*Group-2*	*Group-3*	*Group-4*	*Group-5*
Minimum	361.51	357.19	360.40	356.96	356.85	296.83	350.31	345.20	354.07	353.78	337.80	350.54	345.18	354.08	353.79
Lower Quartile : p25	364.01	358.59	361.98	358.84	359.45	362.63	354.44	354.85	354.92	355.75	360.02	354.58	354.85	354.93	355.76
Median : p50	367.72	365.81	366.01	365.43	365.46	370.67	361.46	363.43	361.90	363.19	366.77	361.53	363.45	361.91	363.23
Upper Quartile : p75	370.17	370.81	369.81	371.89	370.17	378.69	368.04	371.47	371.62	368.98	372.16	368.09	371.53	371.61	368.98
Maximum	372.24	372.19	372.46	372.94	372.77	355.59	373.39	374.37	374.73	372.55	302.86	373.28	374.37	374.71	372.55
Standard Deviation	3.89	6.27	4.55	6.65	6.10	46.28	8.34	10.74	8.77	7.41	19.44	8.26	10.76	8.76	7.40
Skewness	-0.15	-0.15	0.07	-0.04	-0.07	0.23	0.05	-0.29	0.13	0.01	0.37	0.05	-0.29	0.13	0.00
Kurtosis	-1.52	-1.71	-1.64	-1.72	-1.67	-0.08	-1.50	-1.47	-1.82	-1.76	-0.27	-1.53	-1.47	-1.82	-1.75
Multivariate distribution statistics for performance error	Model and Input Parameters														
	ETR					RF					GBR				
	Group-1	*Group-2*	*Group-3*	*Group-4*	*Group-5*	*Group-1*	*Group-2*	*Group-3*	*Group-4*	*Group-5*	*Group-1*	*Group-2*	*Group-3*	*Group-4*	*Group-5*
Minimum	367.33	367.28	367.01	366.92	367.12	367.38	366.87	367.11	366.15	366.61	364.91	364.92	364.91	364.91	364.91
Lower Quartile : p25	368.02	367.58	367.50	367.55	367.50	367.55	367.53	367.43	367.18	367.43	364.91	364.96	364.91	364.94	365.02
Median : p50	370.46	368.25	368.04	368.88	368.38	369.10	367.70	367.90	367.51	367.62	366.27	366.35	366.26	366.31	366.41
Upper Quartile : p75	372.46	369.22	369.13	372.76	371.49	371.03	368.41	368.23	371.42	370.56	367.68	367.69	367.67	367.86	367.67
Maximum	373.92	375.40	375.33	374.63	374.53	372.17	376.22	374.86	373.32	373.60	376.85	376.11	374.80	380.15	378.14
Standard Deviation	2.55	2.80	2.84	3.06	2.79	1.92	3.23	2.70	2.85	2.56	4.22	3.94	3.50	5.39	4.65
Skewness	0.05	1.54	1.52	0.51	0.75	0.20	1.69	1.70	0.64	0.82	1.47	1.44	1.33	1.58	1.54
Kurtosis	-1.70	0.72	0.68	-1.48	-1.11	-1.69	1.02	1.04	-1.45	-1.02	0.60	0.54	0.35	0.80	0.74

Received 22 June 2025; Accepted 22 September 2025



Group-1 Input Parameters

Group-2 Input Parameters

Group-3 Input Parameters

Group-4 Input Parameters

Group-5 Input Parameters

Figure-3: Boxplots comparing the real and redirected coffee yields from Group-1,2,3,4,5 parameters using ideal BRR, Lasso, Enet, ETR, RFR and GBR models.

4 Discussion

Coffee yield is strongly influenced by abiotic conditions such as rainfall, minimum and maximum temperature, sunshine duration, vapor pressure, dew point, and relative humidity, which vary annually and contribute to production uncertainty. This study evaluated six stochastic machine learning models—Bayesian Ridge, Lasso, Elastic Net, Random Forest, Extra Trees, and Gradient Boosting regressions—using long-term abiotic factor data (2004–2022) to predict coffee yields in India.

Among the models tested, Bayesian Ridge regression exhibited the strongest predictive performance, particularly when restricted to Group-3 parameters (year, rainfall, temperature min–max, and relative humidity). At the 70:30 train–test split, it achieved the highest explanatory power ($R^2 = 0.81$) with the lowest prediction error (RMSE = 61.52 kg per ha; MAE = 54.81 kg per ha). Similarly, Lasso regression demonstrated near-identical performance using the same parameter group ($R^2 = 0.80$; RMSE = 62.20 kg per ha; MAE = 55.60 kg per ha). Both models highlight the principle of statistical parsimony, whereby fewer but highly relevant climatic predictors reduce multicollinearity and enhance generalization accuracy. Expanded parameter sets, such as Group-1 (all variables), reduced efficiency in both models, with Bayesian Ridge declining to $R^2 = 0.64$; RMSE = 84.07 kg per ha and Lasso to $R^2 = 0.65$; RMSE = 83.11 kg per ha.

By contrast, Elastic Net regression showed moderate predictive capability, performing best with Group-1 (all parameters) at the 70:30 split ($R^2 = 0.72$; RMSE = 76.54 kg per ha; MAE = 71.09 kg per ha). However, reduced predictor sets (e.g., Groups 2 and 4) yielded weaker fits ($R^2 \leq 0.64$; RMSE $>$ 83 kg per ha), suggesting over-penalization limited its capacity to capture nonlinear interactions. Ensemble tree-based models fared less effectively: Random Forest regression achieved only $R^2 = 0.53$; RMSE = 96.37 kg per ha; MAE = 81.41 kg per ha (Group-1, 70:30 split), while Extra Trees regression performed similarly ($R^2 = 0.53$; RMSE = 96.75 kg per ha; MAE = 88.25 kg per ha). Both models displayed instability across groups and splits, with R^2 rarely exceeding 0.52 and RMSE consistently above 97 kg per ha, indicating sensitivity to noise and parameter redundancy. Finally, Gradient Boosting regression showed the weakest generalization, with its best case (Group-3, 70:30 split) achieving only $R^2 = 0.46$; RMSE = 103.20 kg per ha; MAE = 87.84 kg per ha. Although higher R^2 values were observed at 90:10 splits (~0.76‑0.79), these deteriorated sharply under larger test sets ($R^2 \leq 0.43$; RMSE $>$ 106 kg per ha), confirming overfitting.

These results emphasize that parsimonious linear models (Bayesian Ridge and Lasso) outperformed nonlinear ensemble regressors for coffee yield prediction when limited to key abiotic predictors. This aligns with previous agricultural modeling studies that identified climatic factors, especially rainfall and temperature, as dominant predictors of crop yields in soybean, maize, rice, and sugarcane [22], [35]–[37][38]. Comparable findings in other domains demonstrate the utility of machine learning in agriculture, including pest infestation forecasting [24], groundwater level prediction [39], irrigation management, precision farming, and farmland mapping [40]–[43]. However, unlike prior works, which often relied on soil nutrient data or disease incidence to predict coffee outcomes in regions such as Vietnam, Mexico, Uganda, and Zimbabwe, the present study is the first to focus exclusively on abiotic climatic predictors using long-term Indian

Received 22 June 2025; Accepted 22 September 2025

datasets collected from the Central Coffee Research Institute (CCRI), Balehonnur, Karnataka.

While Bayesian Ridge and Lasso regression demonstrated strong predictive power, limitations remain. Abiotic factors alone cannot fully capture yield variability, as soil fertility, pest and disease pressure, and fertilizer application are major confounders. The reliance on a single geographic dataset also constrains generalizability. Future work should expand to multi-regional datasets, integrate soil and biotic stress variables, and evaluate hybrid approaches such as ensemble Bayesian frameworks. Random sampling ensembles of Bayesian Ridge could also be applied to quantify prediction uncertainty and reduce statistical error bounds. Such improvements would strengthen biophysical yield models and support more reliable decision-making for smallholder coffee farmers facing increasing climatic variability.

6 Conclusion

The use of stochastic machine learning models as a robust data-driven method for analyzing predictive characteristics in abiotic factor data to optimize coffee crop productivity was assessed at a coffee research station in Balehonnur, Karnataka. Among the six stochastic machine learning models employed in the current work, two stand out: Bayesian Ridge regression with R^2=0.81 and RMSE = 61.52 kg per ha and Lasso Regression with R^2 = 0.80 and RMSE = 62.30 kg per ha. These models take a look at a wide range of abiotic factors—including year, rainfall, temperature, sunshine, relative humidity, vapor, and dew points—and use them as predictors of the objective variable, coffee crop yield (Y). The results show that when it comes to predicting coffee yield using multiple inputs, Bayesian ridge and Lasso regression models are more reliable and efficient at extracting features between abiotic factors and crop yields than random forest with R^2 = 0.53 and RMSE = 96.37 kg per ha, extra tree with R^2 = 0.53 and RMSE = 96.75 kg per ha, gradient boosting with R^2 = 0.46 and RMSE = 103.23 kg per ha, or elastic net models with R^2 = 0.72 and RMSE = 76.54 kg per ha respectively. In order to intentionally increase yield in coffee research stations using a set of meticulously curated datasets for abiotic factors, the present study validated the possible value of integrating AI algorithms with biophysical crop models in decision support systems that employ precision agriculture.

ACKNOWLEDGEMENTS

The authors appreciate the support received in the form of historical coffee records from the Central Coffee Research Institute (CCRI) in Balehonnur, Chikkamagaluru District, Karnataka, India, provided through JSS Science and Technology University, Mysuru. These records were crucial in facilitating the anticipation of coffee harvest yields and in understanding the various factors influencing coffee production.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work the author(s) used QuillBot AI tool to paraphrase self-plagiarism. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1]. Martins, L. D., Eugenio, F. C., Rodrigues, W. N., Brinati, S. V. B., Colodetti, T. V., Christo, B. F., Olivas, D. B. L., Partelli, F. L., Amaral, J. F. T. do, Tomaz, M. A., Ramalho, J. D. C., Santos, A. R. dos, Martins, L. D., Eugenio, F. C., Rodrigues, W. N., Brinati, S. V. B., Colodetti, T. V., Christo, B. F., Olivas, D. B. L., … Santos, A. R. dos. (2018). Adaptation to Long-Term Rainfall Variability for Robusta Coffee Cultivation in Brazilian Southeast. American Journal of Climate Change, 7(4), 487–504. https://doi.org/10.4236/AJCC.2018.74030

[2]. Campanha, M. M., Santos, R. H. S., De Freitas, G. B., Martinez, H. E. P., Garcia, S. L. R., & Finger, F. L. (2004). Growth and yield of coffee plants in agroforestry and monoculture systems in Minas Gerais, Brazil. Agroforestry Systems, 63(1), 75–82. https://doi.org/10.1023/B:AGFO.0000049435.22512.2D/METRICS

[3]. Wu, M., Shi, Z., Zhang, H., Wang, R., Chu, J., Liu, S. Q., Zhang, H., Bi, H., Huang, W., Zhou, R., & Wang, C. (2025). Predicting the flavor potential of green coffee beans with machine learning-assisted visible/near-infrared hyperspectral imaging (Vis-NIR HSI): Batch effect removal and few-shot learning framework. Food Control, 175, 111310. https://doi.org/10.1016/J.FOODCONT.2025.111310

[4]. Wang, N., Jassogne, L., van Asten, P. J. A., Mukasa, D., Wanyama, I., Kagezi, G., & Giller, K. E. (2015). Evaluating coffee yield gaps and important biotic, abiotic, and management factors limiting coffee production in Uganda. European Journal of Agronomy, 63, 1–11. https://doi.org/10.1016/J.EJA.2014.11.003

[5]. P.G., C., & C.M., D. (2016). Climate Variability Concerns for the Future of Coffee in India : An Exploratory Study. International Journal of Environment, Agriculture and Biotechnology, 1(4), 819–826. https://doi.org/10.22161/ijeab/1.4.27

[6]. Della Peruta, R., Mereu, V., Spano, D., Marras, S., Vezy, R., & Trabucco, A. (2025). Projecting trends of arabica coffee yield under climate change: A process-based modelling study at continental scale. Agricultural Systems, 227, 104353. https://doi.org/10.1016/J.AGSY.2025.104353

[7]. Armando, L., Navarro, A., Rivera Hernández, B., López Castañeda, A., Jesús, D., López, P., Mancillas, R. G., Francisco, J., & López, J. (2018). Potential areas and vulnerability of the robust coffee crop (Coffea canephora P.) to climate change in the state of Tabasco, Mexico. Nova Scientia, 10(20), 369–396. https://doi.org/10.21640/NS.V10I20.1379

[8]. Rahn, E., Vaast, P., Läderach, P., van Asten, P., Jassogne, L., & Ghazoul, J. (2018). Exploring adaptation strategies of coffee production to climate change using a process-based model. Ecological Modelling, 371, 76–89. https://doi.org/10.1016/J.ECOLMODEL.2018.01.009

[9]. Gines, K. R. S., Garcia, E. V., Sagum, R. S., & Bautista VII, A. T. (2025). Geographical origin differentiation of Philippine Robusta coffee (C. canephora) using X-ray fluorescence-based elemental profiling with chemometrics and machine learning. Food Chemistry, 478, 143676. https://doi.org/10.1016/J.FOODCHEM.2025.143676

[10]. Byrareddy, V., Kouadio, L., Kath, J., Mushtaq, S., Rafiei, V., Scobie, M., & Stone, R. (2020). Win-win: Improved irrigation management saves water and increases yield for robusta coffee farms in Vietnam. Agricultural Water Management, 241, 106350. https://doi.org/10.1016/J.AGWAT.2020.106350

[11]. Kittichotsatsawat, Y., Tippayawong, N., & Tippayawong, K. Y. (2022). Prediction of annual coffee production yield using artificial neural network and multiple linear regression techniques. https://doi.org/10.21203/RS.3.RS-1504007/V1

[12]. de Freitas, C. H., Coelho, R. D., de Oliveira Costa, J., & Sentelhas, P. C. (2025). Equationing Arabica coffee: Adaptation, calibration, and application of an agrometeorological model for yield estimation. Agricultural Systems, 223, 104181. https://doi.org/10.1016/J.AGSY.2024.104181

[13]. Varshitha, D. N., & Choudhary, S. (2022). Soil fertility and yield prediction of coffee plantation using machine learning technique. Res J Agric Sci, 13, 514–518.

[14]. Byrareddy, V., Kouadio, L., Mushtaq, S., & Stone, R. (2019). Sustainable Production of Robusta Coffee under a Changing Climate: A 10-Year Monitoring of Fertilizer Management in Coffee Farms in Vietnam and Indonesia. Agronomy 2019, Vol. 9, Page 499, 9(9), 499. https://doi.org/10.3390/AGRONOMY9090499

[15]. Pambudi, S., Jongyingcharoen, J. S., & Saechua, W. (2024). Machine learning based prediction and iso-conversional assessment of oxidatively torrefied spent coffee grounds pyrolysis. Renewable Energy, 237, 121657. https://doi.org/10.1016/J.RENENE.2024.121657

[16]. Muliasari, A. A., & Dewi, H. (2022). Estimated Yield Potential of Robusta Coffee (Coffea canephora Pierre ex A. Froehner) at Bogor District. E3S Web of Conferences, 348, 00020. https://doi.org/10.1051/E3SCONF/202234800020

[17]. Veenadhari, S., Misra, B., & Singh, C. D. (2014). Machine learning approach for forecasting crop yield based on climatic parameters. 2014 International Conference on Computer Communication and Informatics: Ushering in Technologies of Tomorrow, Today, ICCCI 2014. https://doi.org/10.1109/ICCCI.2014.6921718

[18]. Pambudi, S., Jongyingcharoen, J. S., & Saechua, W. (2025). Explainable machine learning for predicting thermogravimetric analysis of oxidatively torrefied spent coffee grounds combustion. Energy, 320, 135288. https://doi.org/10.1016/J.ENERGY.2025.135288

[19]. Natarajan, R., Subramanian, J., & Papageorgiou, E. I. (2016). Hybrid learning of fuzzy cognitive maps for sugarcane yield classification. Computers and Electronics in Agriculture, 127, 147–157. https://doi.org/10.1016/J.COMPAG.2016.05.016

[20]. Sirsat, M. S., Mendes-Moreira, J., Ferreira, C., & Cunha, M. (2019). Machine Learning predictive model of grapevine yield based on agroclimatic patterns. Engineering in Agriculture, Environment and Food, 12(4), 443–450. https://doi.org/10.1016/J.EAEF.2019.07.003

[21]. KUMAR, S., ATTRI, S. D., & SINGH, K. K. (2019). Comparison of Lasso and stepwise regression technique for wheat yield prediction. Journal of Agrometeorology, 21(2), 188–192. https://doi.org/10.54386/JAM.V21I2.231

[22]. Kaul, M., Hill, R. L., & Walthall, C. (2005). Artificial neural networks for corn and soybean yield prediction. Agricultural Systems, 85(1), 1–18. https://doi.org/10.1016/J.AGSY.2004.07.009

[23]. Astuti, S. D., Wicaksono, I. R., Soelistiono, S., Permatasari, P. A. D., Yaqubi, A. K., Susilo, Y., Putra, C. D., & Syahrom, A. (2024). Electronic nose coupled with artificial neural network for classifying of coffee roasting profile. Sensing and Bio-Sensing Research, 43, 100632. https://doi.org/10.1016/J.SBSR.2024.100632

[24]. Kim, Y. H., Yoo, S. J., Gu, Y. H., Lim, J. H., Han, D., & Baik, S. W. (2014). Crop Pests Prediction Method Using Regression and Machine Learning Technology: Survey. IERI Procedia, 6, 52–56. https://doi.org/10.1016/J.IERI.2014.03.009

[25]. Shastry, K. A., Sanjay, H. A., & Deshmukh, A. (2016). A parameter based customized artificial neural network model for crop yield prediction. Journal of Artificial Intelligence, 9(1–3), 23–32. https://doi.org/10.3923/JAI.2016.23.32

[26]. Shakoor, M. T., Rahman, K., Rayta, S. N., & Chakrabarty, A. (2017). Agricultural production output prediction using Supervised Machine Learning techniques. 2017

1st International Conference on Next Generation Computing Applications, NextComp 2017, 182–187. https://doi.org/10.1109/NEXTCOMP.2017.8016196

[27]. Mishra, S., Mishra, D., & Santra, H. (n.d.). Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper. https://doi.org/10.17485/ijst/2016/v9i38/95032

[28]. Chen, H., Wu, W., & Liu, H. Bin. (2016). Assessing the relative importance of climate variables to rice yield variation using support vector machines. Theoretical and Applied Climatology, 126(1–2), 105–111. https://doi.org/10.1007/S00704-015-1559-Y/TABLES/2

[29]. González Sánchez, A., Frausto Solís, J., Ojeda Bustamante, W., Gonzalez-Sanchez, A., FRAUSTO SOLIS, J. 31308, & OJEDA BUSTAMANTE, W. 33681. (2014). Predictive ability of machine learning methods for massive crop yield prediction. Spanish Journal of Agricultural Research (2171-9292), 12(2), 12(2), 57–79. https://doi.org/10.16/CSS/JQUERY.DATATABLES.MIN.CSS

[30]. García-Nieto, P. J., García-Gonzalo, E., & Paredes-Sánchez, J. P. (2021). Prediction of the critical temperature of a superconductor by using the WOA/MARS, Ridge, Lasso and Elastic-net machine learning techniques. Neural Computing and Applications, 33(24), 17131–17145. https://doi.org/10.1007/S00521-021-06304-Z/TABLES/8

[31]. Kouadio, L., Deo, R. C., Byrareddy, V., Adamowski, J. F., Mushtaq, S., & Phuong Nguyen, V. (2018). Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. Computers and Electronics in Agriculture, 155, 324–338. https://doi.org/10.1016/J.COMPAG.2018.10.014

[32]. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3–42. https://doi.org/10.1007/S10994-006-6226-1/METRICS

[33]. Wijaya, D. R., Afianti, F., Arifianto, A., Rahmawati, D., & Kodogiannis, V. S. (2022). Ensemble machine learning approach for electronic nose signal processing. Sensing and Bio-Sensing Research, 36, 100495. https://doi.org/10.1016/J.SBSR.2022.100495

[34]. Singh, U., Rizwan, M., Alaraj, M., & Alsaidan, I. (2021). A Machine Learning-Based Gradient Boosting Regression Approach for Wind Power Production Forecasting: A Step towards Smart Grid Environments. Energies 2021, Vol. 14, Page 5196, 14(16), 5196. https://doi.org/10.3390/EN14165196

[35]. Drummond, S. T., Sudduth, K. A., Joshi, A., Birrell, S. J., & Kitchen, N. R. (2003). STATISTICAL AND NEURAL METHODS FOR SITE–SPECIFIC YIELD PREDICTION. Transactions of the ASAE, 46(1), 5-. https://doi.org/10.13031/2013.12541

[36]. Görgens, E. B., Montaghi, A., & Rodriguez, L. C. E. (2015). A performance comparison of machine learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics. Computers and Electronics in Agriculture, 116, 221–227. https://doi.org/10.1016/J.COMPAG.2015.07.004

[37]. Fieuzal, R., Marais Sicre, C., & Baup, F. (2017). Estimation of corn yield using multi-temporal optical and radar satellite data and artificial neural networks. International Journal of Applied Earth Observation and Geoinformation, 57, 14–23. https://doi.org/10.1016/J.JAG.2016.12.011

[38]. Neethirajan, S. (2020). The role of sensors, big data and machine learning in modern animal farming. Sensing and Bio-Sensing Research, 29, 100367. https://doi.org/10.1016/J.SBSR.2020.100367

[39]. Sahoo, S., Russo, T. A., Elliott, J., & Foster, I. (2017). Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. Water

Resources Research, 53(5), 3878–3895. https://doi.org/10.1002/2016WR019933

[40]. Cheviron, B., Vervoort, R. W., Albasha, R., Dairon, R., Le Priol, C., & Mailhol, J. C. (2016). A framework to use crop models for multi-objective constrained optimization of irrigation strategies. Environmental Modelling & Software, 86, 145–157. https://doi.org/10.1016/J.ENVSOFT.2016.09.001

[41]. Dimitriadis, S., & Goumopoulos, C. (2008). Applying machine learning to extract new knowledge in precision agriculture applications. Proceedings - 12th Pan-Hellenic Conference on Informatics, PCI 2008, 100–104. https://doi.org/10.1109/PCI.2008.30

[42]. Chemura, A., & Mutanga, O. (2017). Developing detailed age-specific thematic maps for coffee (Coffea arabica L.) in heterogeneous agricultural landscapes using random forests applied on Landsat 8 multispectral sensor. Geocarto International, 32(7), 759–776. https://doi.org/10.1080/10106049.2016.1178812

[43]. Ansari, G., Pal, A., Srivastava, A. K., & Verma, G. (2023). Machine learning approach to surface plasmon resonance bio-chemical sensor based on nanocarbon allotropes for formalin detection in water. Sensing and Bio-Sensing Research, 42, 100605. https://doi.org/10.1016/J.SBSR.2023.100605

Notes on contributors

Santhosh C S is an assistant professor in department of Computer applications JSS Science and Technology University, Mysuru, Karnataka, India. His main teaching and research interests include data mining and machine learning in agricultural applications. He has published several research articles in international conferences and journals in the field of Computer science, applications and information technology.

Dr. Umesh K K is an associate professor in the department of information science and engineering, JSS Science and Technology University, Mysuru, Karnataka, India. His main teaching and research interests include data mining, retrieval systems and machine learning. He has published several research articles in international conferences and journals in the field of Computer science and information technology.

Dr. Narendra Khatri, SM IEEE, LM ISTE, is Assistant Professor (Senior Scale) at Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India. With 34+ SCI/SCIE publications, patents, and strong global collaborations, his research spans Artificial Intelligence, IoT, Embedded Systems, and Machine Learning. He plays a key editorial role as Academic Editor for PLOS ONE and Scientific Reports (Nature Portfolio), contributing to the advancement of high-quality scientific dissemination.