Ratio Mathematica Volume 46, 2023

Wheat crop yield forecasting using various regression models

Shakila C V*
Khadar Babu SK †

Abstract

The prediction of crop yield, particularly paddy production is a challenging task and researchers are familiar with forecasting the paddy yield using statistical methods, but they have struggled to do so with greater accuracy for a variety of factors. Therefore, machine learning methods such as Elastic Net, Ridge Regression, Lasso and Polynomial Regression are demonstrated to predict and forecast the wheat yield accurately for all India-level data. Assessment metrics such as coefficient of determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) are used to evaluate the performance of each developed model. Finally, while evaluating the prediction accuracy using evaluation metrics, the performance of the Polynomial Regression model is shown to be high when compared to other models that are already accessible from various research in the literature.

Keywords: Elastic Net; Ridge Regression; Lasso Regression; Polynomial Regression; Ordinary Least Squares; forecast
2020 AMS subject classifications: 62J05, 62J07, 62-04, 62-06. [1]

*Department of Mathematics, Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India-632014; ajjimaths@gmail.com.

†Department of Mathematics, Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India-63201; Khadar.babu36@gmail.com.

[1]Received on September 15, 2022. Accepted on December 15, 2022. Published on March 20, 2023. DOI: 10.23755/rm.v46i0.1085. ISSN: 1592-7415. eISSN: 2282-8214.

1 Introduction

Agriculture has been the economic backbone of many nations. There are more than 118.9 million farmers in India, and as the population expands, there will be a demand for food. As a result, we need new methods to produce more food products in a shorter amount of time. However, since agriculture is not a profitable industry, not many people choose it as a career. Bhosale et al. [2018]. Agriculture has always been recognized as a vital and great culture that India has traditionally practiced. In the past, people used the land where they lived and made crop choices based on the local weather and conditions. However, due to the greenhouse effect and changes in climatic conditions, farmers today cannot predict when it will rain, snow, or whether there will be water available for their crops, among other things. With this serious issues and demand, it is important to estimate sustainable agricultural production using a system that can accurately measure crop conditions, crop type, and yield. Freie et al. [1999].

There are a few approaches to dealing with constructing the suitable improvement in the agricultural industry. There are several methods to approach these issues by utilizing some of the most significant technological advancements. One of the greatest and simplest technologies we can employ is AI-based and machine learning prediction principles. Furthermore, recently developed machine learning (ML) algorithms are more capable than statistical techniques to find yield estimations.

Artificial neural networks, Decision trees, Regression analysis, Clustering, Bayesian networks, Time series analysis, and Markov chain models are just a few of the mathematical and statistical techniques used in machine learning (ML) approaches for crop prediction. Due to the availability of multiple data from various sources to expose hidden information, the use of these machine learning techniques in crop production demonstrates even more significant advantages.

2 Literature survey

Machine learning addresses problems when the relationship between information and yield variables is uncertain or difficult to comprehend (Shaik et al. [1999] and Veenadhari et al. [2014]). Unlike traditional measuring techniques, machine learning explicitly displays the data whose trade-mark is beneficial to conduct modeling of complex and non-linear practices, such as a capacity for crop output forecasting (Praveen and Rama [2019] and Kumar et al. [2019]). Using supervised learning, the majority of machine learning algorithms are successfully used to predict crop yields (Praveen et al. [2017] and RaviKumar et al. [2019]). The preparation measures will continue until the model attained the desired level of

accuracy on the preparation data.

The majority of research in the past have developed statistical agricultural production prediction models using multiple linear regressions (MLRs) (Rai et al. [2013];Kumar et al. [2014]; Dhekale et al. [2014]).

Das et al. [2017] has studied about the Statistical approaches for feature selection or feature extraction, such as Least Absolute Shrinkage and Selection Operator (LASSO), Stepwise Multiple Linear Regression (SMLR) or Elastic Net (ENET) method, can be utilized to address these issues. Yousefi et al. [2015] discussed to forecast the output energy of rice production in Iran, several researchers used the polynomial and radial basis function kernels of support vector regression (SVR). Paidipati et al. [2021] developed a model using SVR Approach with Various Non-Linear Patterns for Forecasting Rice Cultivation in India.

There are few studies comparing the accuracy of feature selection, feature extraction, and both approaches combined for agricultural yield forecasting. With the following objectives: (i) to develop overall crop yield prediction models using various multivariate models; and (ii) to assess the analytical performance of the developed models, our study has found scope to develop and select a statistical forecasting model for rice using various regression techniques for the India level. Elastic Net Regression, Ridge Regression, Lasso Regression, and Polynomial Regression are some of the techniques we employed to construct this work. There are many comparable projects on the market, but what sets our project apart from the competition is how we've integrated Python with machine learning to cut down on the number of lines of code and production costs while still producing accurate results (Pramod et al. [2019]; Tutun et al. [2016]).

3 Material and methods

3.1 Ridge regression

In Ridge regression μ is a penalty term and that penalty function is equal to the squared root of the coefficient. The square of the coefficients magnitude corresponds to the L_2 term. To regulate that penalty term, we additionally incorporate the coefficient μ. In this instance, if μ is zero, the formula is the fundamental OLS; however, if μ is more than zero, a constraint will be added to the coefficient. This constraint makes the quantity of the coefficient tend towards zero as we raise the amount of μ. This results in a tradeoff between smaller variance and increased bias.

$$L_R = arg.min_{\hat{\alpha}} \left(\|Y - \alpha * X\|^2 + \mu * \|\alpha\|^2 \right).$$

where μ is regularization penalty.

Because it never reaches a coefficient of zero but only minimises it, ridge regression lowers a model's complexity without lowering the number of variables. As a result, this model is unable to achieve feature reduction.

3.2 Lasso regression

Least Absolute Shrinkage and Selection Operator is short for lasso regression. It extends the cost function's penalty term. The whole sum of the coefficients is represented by this phrase. When the value of the coefficients increases from 0 to 1, this term penalises, causing the model to lower the value of variables in order to minimise loss. While lasso regression usually makes the value of the coefficient to absolute zero, ridge regression never does.

$$L_lasso = arg.min_{\hat{\alpha}} \left(\|Y - \alpha * X\|^2 + \mu * \|\alpha\|_1 \right).$$

With various data types, Lasso occasionally has difficulties.If the number of predictors (p) is more than the number of observations, Lasso will choose at most n predictors as non-zero even if all of the predictors are significant (n). The LASSO regression method chooses one of the highly collinear variables at random when there are two or more, which is bad for data interpretation.

3.3 Elastic net

To address the drawbacks of Ridge and Lasso regression, **?** formed an elastic net regression. In general, ridge regression performs best with highly correlated variables, whereas Lasso regression performs well with less correlated variables. However, there are many models that represent a significant number of variables but lack information on attributes like correlation. Ridge regressions and Lasso are not very helpful in these circumstances. To get away from this problem, the function is estimated using ENR since it takes into account the consequences of both Lasso and Ridge regressions. L_1 and L_2 norms can be used to define the Lasso and ridge regression penalties, respectively. For accurate prediction, ENR take into account the L_1 and L_2 penalties by the following equations.

$$L_{ENR} = arg. \min_{\hat{\alpha}} \sum_i (Y_i - \alpha X_i)^2 + \beta^1 \sum_{k=1}^{1} |\alpha_k| + \beta^2 \sum_{k=1}^{1} \alpha_{k^2}.$$

where $\|L_1\| = \beta^1 \sum_{k=1}^{1} |\alpha_k| and \|L_2\| = beta^2 \sum_{k=1}^{1} \alpha_{k^2}$. L_1 is sum of the weights and L_2 is the sum of the square of the weights.

3.4 Polynomial regression

In polynomial regression, a kind of linear regression, the relationship between the random variable x and the dependent variables y is represented as a n^{th}-degree polynomial. Polynomial regression is used to fit a nonlinear relationship between the value of x and the corresponding dependent mean of y, denoted by the notation $E(y|x)$.

Here is a polynomial regression model's general equation.

$$L = s_0 + s_1x_1 + s_2x_12 + s_2x_13 + ... + s_nx_1n$$

Some correlations may be curvy, according to a researcher's hypothesis. Such scenarios will undoubtedly have a polynomial term. The assumption in common multiple linear regression analysis is that every independent variable is independent of every other independent variable. In the case of polynomial regression models, this assumption is incorrect.

3.5 Model validation

3.5.1 Mean absolute percentage error (MAPE)

To determine the mean absolute percentage error (MAPE), the absolute error for each period is subtracted from predicted values then as follows the procedures.

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{Predicted_i - Actual_i}{Actual_i}\right|$$

4 Results and discussion

4.1 Dataset

Directorate Of Economics and Statistics Department of Agricultural and Farmers Welfare, and Government of India provided the time series data of wheat yield at India level (1966 to 2017). The research used characteristics like Area Under Cultivation (Thousand / Hectares), Production (Thousand / Tons), and Yield (KG / Hectare) to evaluate data from all of India. Elastic Net, Ridge Regression, Lasso Regression, and Polynomial Regression were constructed and compared to determine the best-fit model.

4.2 Overview of wheat parameter statistics

The Elastic Net, Ridge Regression, Lasso Regression, and Polynomial Regression models were separately applied to the wheat yield data to investigate

Table 1: Statistical Measures

Measures	Elastic Net Regression	Lasso Regression	Ridge Regression	Polynomial Regression
R^2	0.9252	0.9252	0.9252	0.9688
MSE	38964456.45	40910114.34	42898765.78	17083450.77
RMSE	5996.1034	6396.1015	6876.3675	4133.2131
MAE	4894.87	5194.88	5794.86	3302.65
MAPE	20.435	55.405	69.203	8.232

their connection, and the effectiveness of each model was evaluated using MSE, R^2, and MAPE.

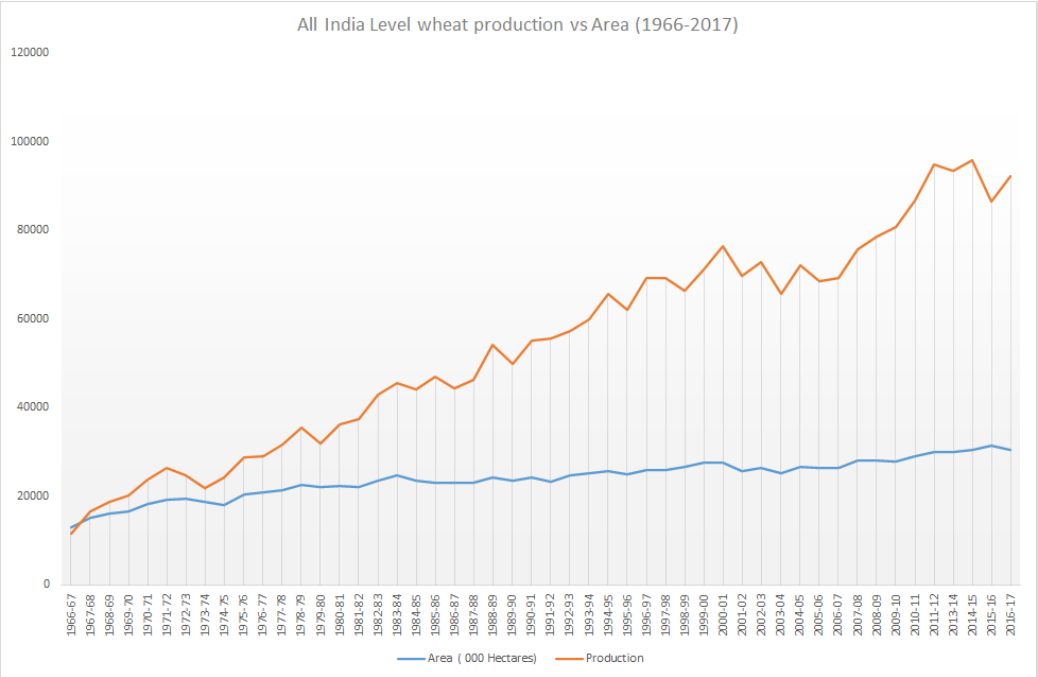

Figure 1: All India Level wheat production vs Area (1966-2017)

Figure 1shows that the all-India level wheat production wise Area. The stronger correlation and the error higher form will be regarded as the most effective method for predicting agricultural production (kg/acre). The statistical approach's results are shown in the first case R^2 values were verified among all the particular regression models (Table 1).

4.3 Regression models

Numerous research showed that machine learning techniques could forecast wheat production. For a consistent wheat yield, it is necessary to increase prediction accuracy. To ensure a consistent wheat production, it is necessary to improve prediction accuracy. The accuracy of the suggested Elastic Net Regression, Ridge Regression, Lasso Regression, and Polynomial Regression for wheat yield prediction is assessed using the R2, RMSE, MAE, MSE, and MAPE metrics, as was previously stated.

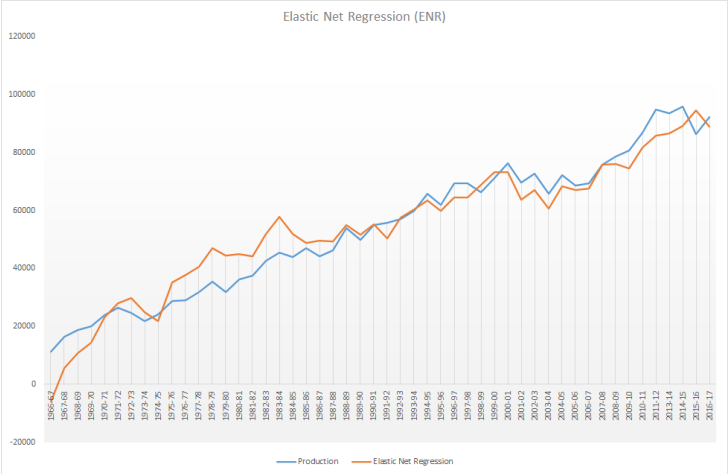

Figure 2: Forecasting using Elastic Net Regression

Analysising the data by using Elastic Net Regression, we have got the values of R^2, MSE, RMSE, MAE, and MAPE of around 0.9252, 38964456.45, 5996.1034, 4894.87, and 20.435, respectively, the forecasting utilizing elastic net regression is shown in Fig.2 Here 92% of data are used to fit the model.

Fig. 3 shows the Forecasting using the Ridge regression, which is evaluation metrics with values of R^2, MSE, RMSE, MAE, and MAPE of about 0.9252, 42898765.78, 6876.3675, 5794.86 and 69.203 respectively. Here 92% of data are used to fit the model. And the MAPE values is 69.2 which large in model fitting.

By using Lasso Regression model the values of R^2, MSE, RMSE, MAE, and MAPE of around 0.9251, 40910114.34, 6396.10, 5194.88 and 55.405, respectively, the forecasting utilizing Lasso regression is shown in Fig 4. Here also 92% of data are used to fit the model. But the MAPE values are huge as 55% for the model fit.

In Fig. 5 forecasting the wheat yield data with Polynomial regression, estimate the R^2, MSE, RMSE, MAE, and MAPE of around is 0.9687, 17083450.77,

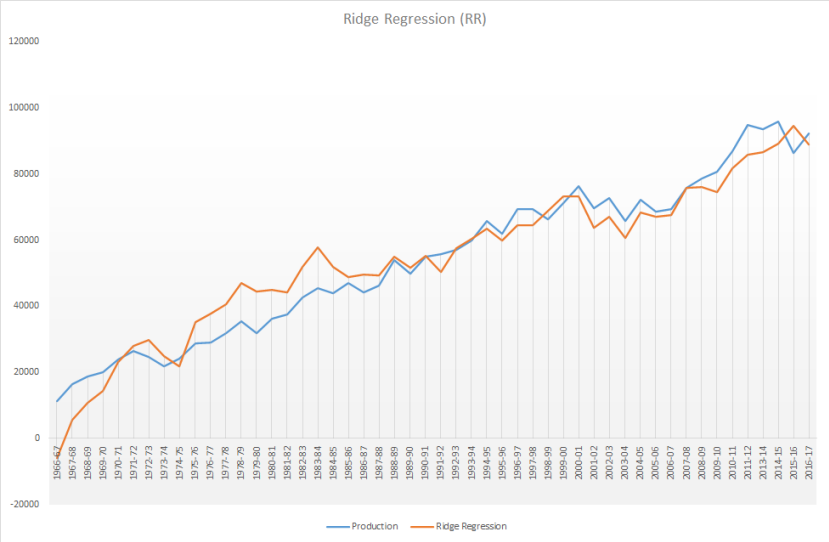

Figure 3: Forecasting using Ridge Regression

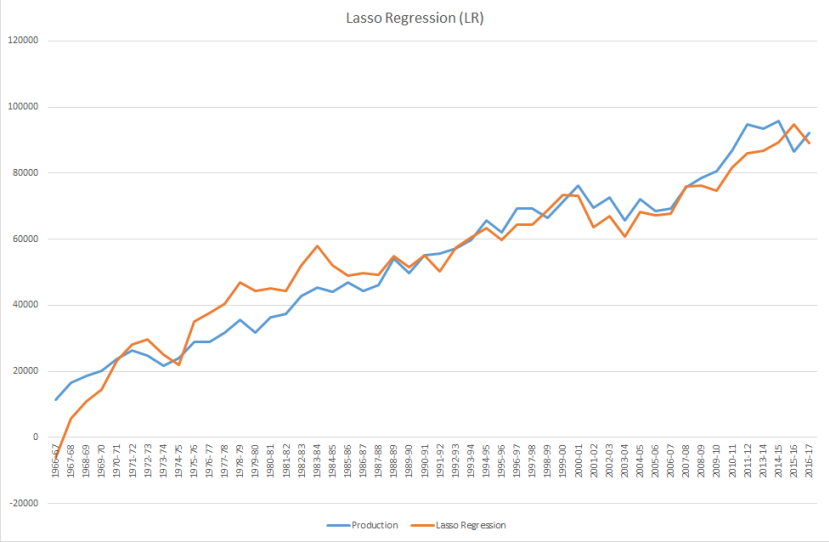

Figure 4: Forecasting using Lasso Regression

4133.213, 3302.674 and 8.23 respectively. Comparing the all above various measure value the R^2 is 96% data using to fit the model. Mean-while, the MAPE values is 8.23 which is highly acceptably accurate level.

Fig. 6 shows the forecasting of India-wide level wheat crop production using various regression models. The accuracy of the Polynomial regression model exhibits superior scale than other chosen machine learning models, as like regression

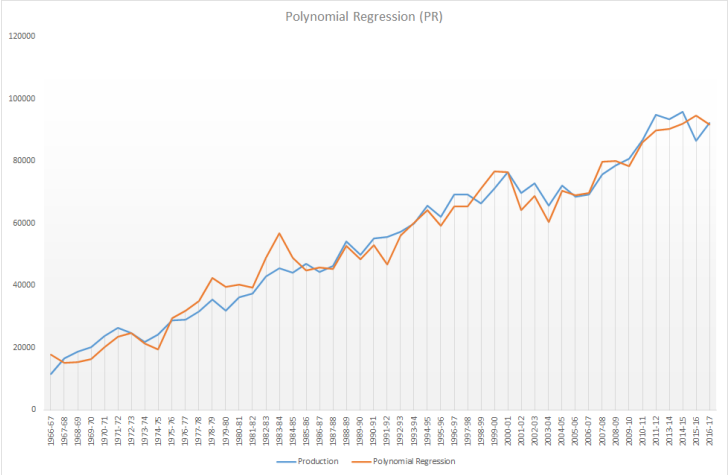

Figure 5: Forecasting using Polynomial Regression

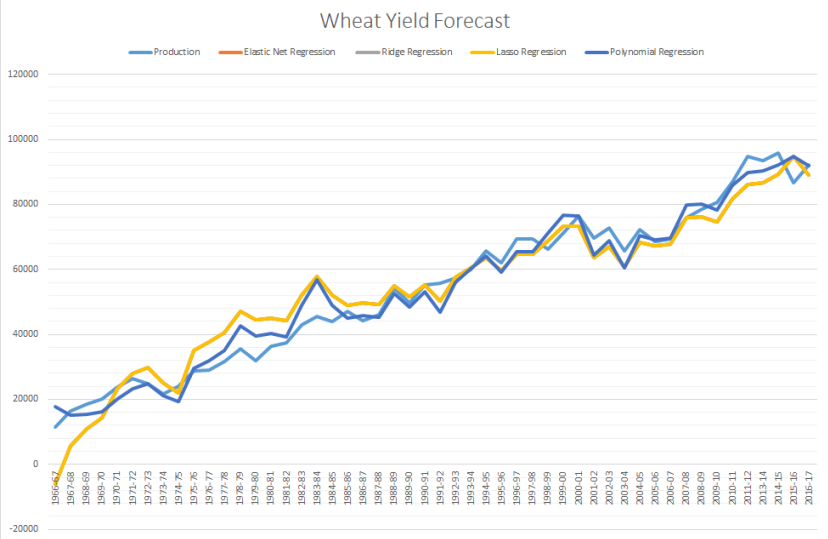

Figure 6: Forecasting using Various Regression Models.

models.

5 Conclusions

Statistical and machine learning methods are used to predict agricultural yield. The statistical analysis of various Regression approaches and machine learning, specifically Elastic Net Regression, Ridge Regression, Lasso Regression, and Polynomial Regression are among the techniques that are evaluated to acquire

higher accurate crop yield forecast. To evaluate the level of accuracy of the various methods, model performance measures are updated. The main findings are drawn from the results seen:

- Assessment metrics such as coefficient of determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) are used to evaluate the performance of each developed model.
- The obtained study showed that the Polynomial Regression method produces superior evaluation metrics with values of R2, MSE, RMSE, MAE, and MAPE of about 0.9687, 17083450.77, 4133.21, 3302.67, and 8.2326 respectively.
- The R^2 metrics of the Polynomial regression is 4.498 percent better than other existing models from other regression models.
- Hence, its improved performance metrics, the suggested machine learning algorithm in particular, Polynomial Regression reduces the risk factor for Wheat yield.

References

S. Bhosale, R. Thombare, P. Dhemey, and A. Chaudhari. Crop yield prediction using data analytics and hybrid approach. *In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA),IEEE*, pages 1–5, 2018.

B. Das, R. Sahoo, S. Pargal, G. Krishna, R. Verma, V. Chinnusamy, V. Sehgal, and V. Gupta. Comparison of different uni- and multi-variate techniques for monitoring leaf water status as an indicator of water-deficit stress in wheat through spectroscopy. *Biosystems Engineering*, 160:69–83, 2017. doi: 10.1016/j.biosystemseng.2017.05.007.

B. Dhekale, PKS, and TPU. Weather based pre-harvest forecasting of rice at kolhapur (maharashtra). *Trends Biosci*, page 39–41, 2014.

U. Freie, S. Sporri, O. Stebler, and F. Holecz. Rice field mapping in sri lanka using ers sar data. *Earth Observ. Q.*, 63:30—-35, 1999.

N. Kumar, R. Pisal, SP, Shukla, and K. Pandye. Regression technique for south gujarat. *MAUSAM*, 65:361–364, 2014.

R. Kumar, M. Reddy, and P. Praveen. Text classification performance analysis on machine learning. *Int.J.AdvSciTechnol*, 20:691–697, 2019.

K. Paidipati, C. Chesneau, N. B M, K. Kumar, P. Kalpana, and C. Ku-rangi. Prediction of rice cultivation in india—support vector regression approach with various kernels for non-linear patterns. *AgriEngineering*, 3:182–198, 2021. doi: 10.3390/agriengineering3020012.

K. Pramod, S. Naresh Kumar, V. Thirupathi, and C. Sandeep. Qos and security problems in 4g networks and qos mechanisms offered by 4g. *International Journal of Advanced Science and Technology*, 20:600–606, 2019.

P. Praveen and B. Rama. An efficient smart search using r tree on spatial data. *Journal of Advanced Research in Dynamical and Control Systems*, 4:1943–1949, 2019.

P. Praveen, B. Rama, and T. S. Kumar. An efficient clustering algorithm of minimum spanning tree. *Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, pages 131–135, 2017.

K. Rai, N. P. V, B. Bharti, and S. K. Pre harvest forecast models based on weather variable. *Adv Biores*, 4:118—-122, 2013.

R. RaviKumar, M. B. Reddy, and P. Praveen. An evaluation of feature selection algorithms in machine learning. *Int J SciTechnol Res*, 12:2071–2074, 2019.

M. A. Shaik, T. S. Kumar, P. Praveen, and R. Vijayaprakash. Research on multiagent experiment in clustering. *International Journal of Recent Technology and Engineering (IJRTE)*, 8:1126–1129, 1999.

S. Tutun, M. Bataineh, M. Aladeemy, and M. Khasawneh. *The Optimized Elastic Net Regression Model for Electricity Consumption Forecasting*, 2016.

S. Veenadhari, B. Misra, and C. D.Singh. Machine learning approach for forecasting crop yield based on climatic parameters. *In 2014 International Conference on Computer Communication and In-formatics*, 8:1– 5, 2014.

M. Yousefi, B. Khoshnevisan, S. Band, S. Motamedi, M. Md Nasir, M. Arif, and R. Ahmad. Support vector regression methodology for prediction of output energy in rice production. *Stochastic Environmental Research and Risk Assessment*, 29, 2015. doi: 10.1007/s00477-015-1055-z.