ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

Thailand's Maize Prices Forecasting Using Ensemble Technique

Sararith Mao[1] and Nuanwan Soonthornphisaj[2*]

[1] Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, 10900, Thailand
[2] Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, 10900, Thailand
* Correspondence: nuanwan.s@ku.th

Citation:
Mao, S.; Soonthornphisaj, N. Thailand's Maize prices forecasting using ensemble technique. *ASEAN J. Sci. Tech. Report.* **2024**, *27*(4), e252279. https://doi.org/10.55164/ajstr.v27i4.252279

Article history:
Received: January 4, 2024
Revised: May 16, 2024
Accepted: June 18, 2024
Available online: June 30, 2024

Abstract: Crop price forecasting is crucial for farmers, policymakers, and investors. This paper aims to propose suitable machine learning models for forecasting Thailand's maize prices by implementing and comparing various machine learning algorithms, including regression trees (RT), support vector regression (SVR), ensemble bagging with RT and SVR as the base learner (Bag-RT and Bag-SVR), and random forest (RF). The dataset used in this study is collected from two main sources: the Office of Agricultural Economics in Thailand (OAE) and the investing.com website from January 2002 to August 2023, consisting of 260 records and 53 features. Given the dataset's numerous independent variables, we applied the recursive feature elimination combined with the Pearson correlation feature selection method to reduce feature dimensions by focusing on the most relevant features. The prediction models were trained using 10-fold cross-validation and evaluated using three metrics: R-squared (R^2), mean absolute error (MAE), and root mean square error (RMSE). The top-performing model, Bag-SVR, achieved the best R^2 value of 0.961, MAE of 0.234, and RMSE of 0.315, followed by the SVR model with R^2 value of 0.959, MAE of 0.251, and RMSE of 0.333. In contrast, the RT model demonstrated the lowest performance scores with an R^2 value of 0.846, MAE of 0.44, and RMSE of 0.617. In conclusion, our study emphasizes the influence of feature selection on model performance and showcases the potential of machine learning models for accurate maize price forecasting in Thailand.

Keywords: Price Prediction; Feature Selection; Regression Problem; Ensemble Technique

1. Introduction

The agricultural sector is important for global economic stability and food security. The cultivation and trade of agricultural crops are particularly crucial, as they substantially impact international markets and complex food supply networks [1]. Thailand is among the Southeast Asian nations whose economies are still heavily dependent on agriculture. In 2020, it significantly contributed THB 1.36 trillion to the country's gross domestic product (GDP), representing 8.65% of the overall GDP [2]. Additionally, around 30% of the total labor force, which includes 6.4 million households, is involved in agricultural activities, emphasizing its role in providing rural employment and generating income [3]. Thailand is known for its diverse agricultural landscape and substantial maize production. Maize is one of the most widely cultivated crops in Thailand. It serves two important roles as feed livestock and as a resource for food and industrial applications. Forecasting maize prices is a complex and challenging task due to the volatile nature of agricultural markets and the influence of various factors such as weather patterns, supply and demand,

import and export, and other economic conditions. Therefore, accurate maize prices forecasting is not just an economic concern but also an urgent task.

Traditionally, forecasting maize prices in Thailand relies on farmers' experience, historical data, and expert opinion, which often fail to capture the complex patterns and non-linear relationships that define the country's current agricultural markets. But with the advance of modern technology like artificial intelligence and machine learning, along with the power to analyze vast datasets using high-performance computing, this limitation has led to a growing interest in utilizing machine learning algorithms to enhance the accuracy and timely agricultural price predictions [4, 5].

This research paper aims to tackle this issue by investigating the utilization of machine learning algorithms to construct predictive models that can improve the accuracy of maize price forecasts in Thailand. These models could benefit more than farmers and agribusinesses but also make food more secure, assist the government in creating better policies, and maintain market stability. The following two main objectives drive this paper:

1) Utilize the feature selection method to identify the most relevant features. This process helps to reduce the dataset's dimensionality, leading to improved computational efficiency and enhancing the effectiveness of the prediction models.
2) Employ various regression machine learning models to determine the one that demonstrates the most robust and effective performance in predicting maize prices in Thailand.

This section reviews the literature relevant to our machine learning-based maize price forecasting research. We aim to situate our work within the broader context of previous studies while highlighting the gaps and opportunities that motivate our research.

In the study related to the agricultural landscape, understanding factors influencing maize prices is really important. Exploring different studies helps us see how weather patterns, trade dynamics, and other factors contribute to the prediction and comprehension of maize pricing. Climate variables like temperature and rainfall were used to forecast the yield and price of corn and soybeans for Hancock County in Illinois, United States [6]. Significant information concerning how fluctuations in the prices of rice and wheat can influence corn prices, even though these crops can be used interchangeably as essential food sources for various needs [7]. The correlations between production and consumption, import and export volume, and supply and demand were examined to identify the main factors influencing maize prices in Chinese markets [8]. In addition to economic and environmental factors, currency rate fluctuations were used to impact agricultural goods trade between China and Africa [9]. These collective studies highlight the necessity of economic and environmental data to provide a complete understanding of the factors influencing maize prices.

The process of selecting key factors holds significant importance, directly impacting the precision of crop price prediction. Several studies have explored different feature selection methods, showcasing their effectiveness in finding the most important features from large datasets to enhance the accuracy of predicting crop prices and market trends. The variables that affect agricultural prices in India, including total area for planting, supply forecasts, government regulations, consumer needs, and producer supply for products derived from agriculture, were investigated. With many features in the dataset, the author employed the feature concatenation approach to select only the feature representative from weather data, data quality, and Agmarknet data [10]. A modified recursive feature elimination (MRFE) technique was introduced, proving highly effective in selecting relevant features. Combined with the bagging ensemble technique, this approach achieved an impressive 95% accuracy rate in predicting land suitability for crop cultivation [11]. The effectiveness of combining the recursive feature elimination (RFE) technique with the adaptive bagging classifier for precise crop suitability prediction was emphasized in the study by [12]. Various factors, including plant population, planting dates, environmental elements, and partial in-season weather knowledge, were analyzed to forecast corn yield in three US Corn Belt states: Illinois, Indiana, and Iowa. The study employed a three-stage feature selection process involving consultation with domain experts, utilization of random forest feature importance, and Pearson correlation analysis, effectively reducing the initial set of 597 features to 72. An optimized weighted ensemble technique was also applied to those selected features, achieving the best performance with a relative root mean square error (RRMSE) of 9.5% [13]. Leave One Out Cross-Validation (LOOCV) was employed on a relatively small dataset consisting of 168 samples and 10 variables, utilizing principal component analysis (PCA) to reduce the feature set from 10 to 7. This study found that the ensemble

models with two, three, or four base learners outperformed individual models for predicting future corn prices [14]. These approaches highlight feature selection's crucial role in improving crop price-prediction models.

Agricultural price forecasting is a vibrant field of research that encompasses diverse scenarios, including stock price prediction [15], energy load forecasting [16], traffic forecasting [17], and crop yield prediction [18]. Within the scope of time series prediction, numerous traditional machine learning models have been introduced, such as Ridge and LASSO regression [19, 20], Gaussian processes [21], support vector regression [22], as well as modern deep learning techniques like LSTMs [23]. Beyond these single models, various ensemble models were proposed to improve prediction accuracy by combining the strengths of multiple base learners [24, 25]. Recent studies have explored traditional and ensemble machine learning models to predict commodity prices. These investigations highlight the effectiveness of diverse techniques in accurately forecasting prices within different agricultural markets. A combination of econometrics and ensemble machine learning models was employed to predict corn and sugar price in Brazil. Their findings revealed that the SVR model outperformed other models because of the small dataset with a remarkable R2 of 0.99 and 0.979 and a low MAE of 0.287 and 0.430 for corn and sugar, respectively [26]. Multiple linear, Ridge, and Lasso regression models were implemented to predict maize prices in Thailand. The authors introduced Pearson correlation analysis and stepAIC function, reducing the initial feature set from 47 to 27. Utilizing the selected features, the multiple linear regression model outperformed other models, achieving an R^2 value of 0.94, MAE of 0.31, and RMSE of 0.50 [20].

From a literature review, our study explores the complex landscape of maize prices forecasting in Thailand, considering multiple factors. We aim to fill these research gaps by underscoring the significance of feature selection, emphasizing its capacity to improve accuracy while reducing the number of features. Furthermore, we explore various machine learning algorithms, including individual models and ensemble bagging techniques, to effectively address the challenges of constructing maize price-prediction models. The structure of this paper is as follows: In Section 2, we provide detailed data used in this study and explain the key theories supporting our research. The results and discussion are presented in Section 3. Finally, in Section 4, we conclude.

2. Materials and Methods

2.1 Dataset

The dataset is essential to our study since it provides insightful information and ensures our research findings' robustness. We present an overview of the dataset used in our study, which was collected from two primary sources.

2.1.1 Office of Agricultural Economics Dataset

We acquired historical data on maize prices in Thailand from January 2002 to August 2023 through the Office of Agricultural Economics (OAE). This dataset includes a wide array of monthly historical data encompassing numerous variables. This study focuses on the dependent variable, which is the price of maize sold by Thai farmers in Thai baht per kilogram. The independent variables cover various aspects of maize and cassava, such as the total planting land area, crop price, crop yield, rainfall, import and export volumes, import and export values, and the price change of both crops. Overall, this dataset contains 48 variables and 260 observations.

2.1.2 Investing.com Dataset

The dataset from the Investing website (https://www.investing.com) is a comprehensive repository containing various crop prices and additional relevant data. It offers records at hourly, daily, and monthly intervals, making it suitable for our research needs. In the existing literature, [7] have examined the correlation among agricultural commodity prices, such as wheat, rice, and corn, which can serve as alternative crops. Our study aims to identify potential factors influencing maize prices, explicitly looking at soy and sugar prices. Additionally, the exchange rate of USD and Thai Baht data was collected, considering its relevance to import and export activities in the global market [9]. This dataset has 260 records and 5 independent variables, covering January 2002 to August 2023.

The descriptive measure of the target variable (maize prices) alongside the prices of alternative crops is shown in Table 1. In contrast, Figure 1 illustrates the trend of monthly maize prices in Thailand over the entire period covered by the dataset.

Figure 1. Graph of monthly maize prices in Thailand

Table 1. Descriptive measure of maize prices and alternative crop's price

Descriptive Measure	Maize	Cassava	Wheat	Soy	Sugar	Rice
Count	260	260	260	260	260	260
Min	3.88	0.74	261.75	430.25	5.67	3.50
Q1	5.99	1.23	427.06	848.87	10.94	9.83
Q3	8.31	2.18	667.97	1305.37	18.66	14.82
Max	11.60	3.15	1088	1754.37	33.97	21.48
Mean	7.15	1.77	555	1026.92	15.14	11.97
Std	1.71	0.57	177.55	316.56	5.85	3.65

2.2 Methodology

Investigating and using methods that improve forecasting accuracy become highly relevant in this setting. To enhance the clarity and dependability of decision-making, increasing forecasting precision is a key focus in our study. As such, this section will explore the key theories crucial to understanding how this research has developed.

2.2.1 Data Preprocessing

The data preprocessing step is a key component of getting data ready for training the machine learning model. First, due to the diverse scale of values among numerous independent variables, we applied the standard scaler technique to normalize and rescale all input variables to a range between 0 and 1. Second, we employed recursive feature elimination using a random forest model to select only the most significant features to maximize overall accuracy. Finally, Pearson correlation was applied to select only the strong linear relationship between input features and target variables. All these data preprocessing steps help to enhance the efficiency and effectiveness of machine learning algorithms and reduce the computational workload for building the prediction model.

1) Standard Scaler: Standard Scaler is a technique used for feature scaling, specifically mean centering and variance scaling. This process does not change the shape of the feature's distribution but ensures that they all have the same scale to avoid potential issues where the magnitudes of certain features could mislead the machine learning models. This operation is performed independently for each feature in the dataset. As a result, the mean of each feature becomes 0, and the standard deviation becomes 1. The formula of the Standard Scaler function for each feature is shown in equation (1).

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

- x is the input feature
- μ is Mean
- σ is the Standard Deviation.

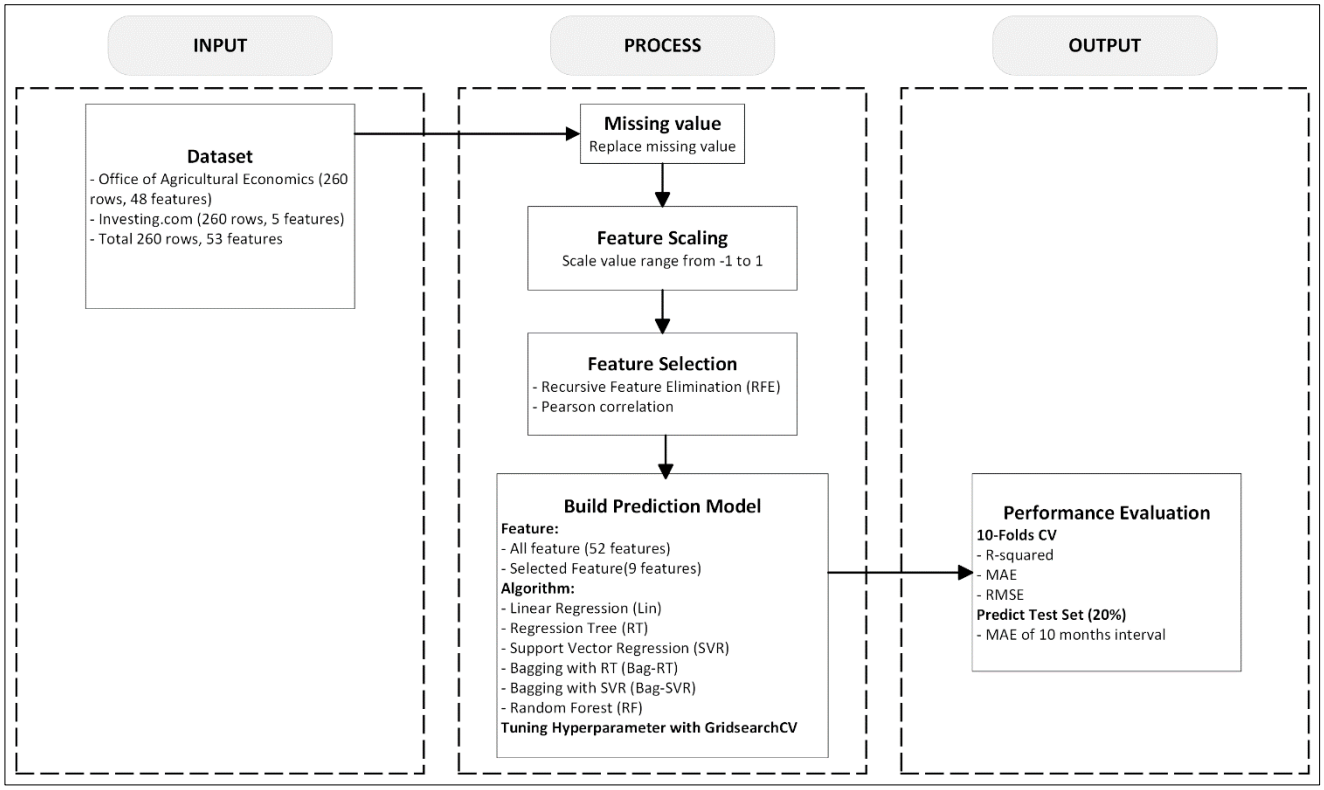

Figure 2. Conceptual framework in this study

2) Feature Selection: As discussed in the dataset description, the presence of numerous variables in the dataset can potentially lead to overfitting, reducing the ability of the predictive model to generalize to new observations. To solve this problem, we proposed two feature selection methods to reduce dataset dimensionality and ensure only the most relevant variables were selected for building the prediction model. The recursive feature elimination (RFE) method based on Random Forest was implemented in the initial stage. The RFE technique systematically evaluates the significance of each feature. It eliminates those that contribute the least to model performance, meaning to keep the feature set that leads to achieving the highest accuracy [11]. In the second stage, we applied a filter-based feature selection method called Pearson correlation, specifically when the input and target variables are quantitative. Pearson correlation is a statistical measure designed to quantify the strength and direction of a linear relationship between two continuous variables. It produces a correlation coefficient, represented by the letter "r" with values between -1 and 1. An "r" value close to 0 denotes a weak or nonexistent linear relationship, whereas an "r" value close to 1 or -1 indicates a strong positive or strong negative linear relationship. The formula produces a number between -1 and 1 by dividing the covariance of the two variables (the numerator) by the product of their standard deviations (the denominator), as shown in equation (2).

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2(Y_i - \bar{Y})^2}} \quad (2)$$

- r represents the Pearson correlation coefficient.
- X_i and Y_i are data points from the two variables being correlated.
- $\bar{X}$ *and* $\bar{Y}$ represent the mean of the respective variables.

2.2.2 Regression Trees

Regression trees are a type of machine learning model employed in regression tasks. They operate by recursively partitioning the dataset into subsets based on input features, with each split chosen to minimize the variance of the target variable within each subset. Predictions are made at the tree's leaf nodes, typically by calculating the mean or median value of the target variable for the data points in that node. One of the key advantages of Regression Trees is their interpretability and visualizability, which allows for a clear understanding of how input features relate to the target variable. However, they are prone to overfitting, particularly when the tree becomes too deep [27].

2.2.3 Support Vector Regression

Initially, support vector machines (SVMs) were introduced to solve classification problems and were later extended to support vector regression (SVR) to address the regression issue. The core concept behind this methodology involves identifying data points close to a hyperplane (known as support vectors) that maximize the margin between two classes of data points – those above and below the target variable. The difference between the target value and a certain threshold is used to calculate this margin. This approach aims to reduce structural risk in regression problems by minimizing the upper bound of generalization error rather than focusing on lowering training error [28]. Recognizing that many real-world problems exhibit non-linear characteristics, SVMs can incorporate the concept of Kernel functions, which enable the transformation of data into a higher-dimensional space and capture its inherent features. Various kernel functions are available, including Gaussian, Polynomial, Linear, and radial basis function (RBF) kernels. Support vector regression (SVR) offers advantages in this study when the dataset consists of several samples. However, a notable drawback when choosing the wrong kernel function can potentially lead to misleading or incorrect conclusions [29].

2.2.4 Ensemble Bagging Regressor

In machine learning, the ensemble bagging regressor is a robust and adaptable approach frequently used with various base learners to significantly boost the performance of regression tasks. The key concept behind ensemble bagging is the creation of multiple base learners that collectively outperform the individual base model [30]. In our research, we take advantage of the power of ensemble bagging by combining it with two base learners named regression trees and support vector regression.

1) RT as a base learner (Bag-RT): Regression Trees are a foundational and widely used base learner in machine learning. They excel at capturing nonlinear relationships between features and target variables, making them well-suited for cases where data exhibits intricate and non-obvious patterns. By integrating ensemble bagging with Regression Trees, we aim to harness the advantages of both techniques. Ensemble bagging lowers the risk of overfitting and model variance by using diverse subsets of the training data (bootstrap samples) to train several instances of Regression Trees. The combination of these methods not only enhances model stability but also enables our model to generalize unseen data more effectively. The adaptability of ensemble bagging with regression trees is valuable when dealing with diverse and complex datasets, ultimately contributing to more accurate results.

2) SVR as a base learner (Bag-SVR): SVR is a robust and versatile base learner known for its capability to model linear and nonlinear relationships between features and target variables. This versatility makes SVR an excellent choice when dealing with datasets with a wide range of complexities. Ensemble bagging with SVR is particularly effective in scenarios characterized by noisy data, complex feature interactions, and challenges in feature selection. The ensemble bagging technique mitigates the pitfalls of overfitting and variance associated with individual models. At the same time, SVR's ability to adapt to various data patterns ensures that our model remains adaptable and resilient.

The combination of ensemble bagging that integrates regression trees and support vector regression as the base learner illustrates our commitment to a comprehensive approach to predicting maize prices. Our goal is to maximize the predictive model's performance through ensemble bagging techniques combined with two strong base learners. Our model generates remarkably accurate forecasts through this strategy while effortlessly capturing complex data patterns. This creates a model that maintains adaptability and robustness, improving accuracy in making predictions for our research.

3) Random Forest (RF): Random Forest Regressor is a robust ensemble machine learning algorithm for regression tasks. It leverages an ensemble of decision trees, each trained on a random subset of the data and features. Aggregating predictions from these trees provides robust and accurate predictions while also handling noisy data and overfitting. It's widely used across various domains and offers the benefit of feature importance analysis to understand the data [31] better.

2.2.5 Hyperparameter Tuning

Hyperparameter tuning is a vital step in developing robust machine-learning models. The choice of hyperparameters can significantly impact the performance and generalizability of the models. In our study, we employed Gridsearch CV, a popular optimization technique, to systematically explore hyperparameter combinations available from each machine learning model. In addition, K subsets (or "folds") with approximately the same size were created within the dataset using K-fold cross-validation. A different fold is used as the validation set, and the remaining K-1 folds are used for each iteration's training set. This process ensures that every data point is used for validation exactly once. The results from each fold are averaged to obtain a more robust performance and reduce the risk of overfitting. Typical values for K include 5, 10, and 20, depending on the size of the dataset and the computational resources available. In our study, the value of K was set to 10 to evaluate these hyperparameter combinations through the Gridsearch CV function as part of the cross-validation process. This method allowed us to identify and select the most effective hyperparameters to improve model performance and accuracy. All Grid values and the hyperparameters chosen for each model in this study are shown in **Table 8**.

2.2.6 Performance Measures

Performance measures are critical tools used to evaluate the effectiveness and quality of machine learning models. These measures provide insights into how well a model performs, how accurately it makes predictions, and its overall reliability. In this paper, the model's performance is assessed using three metrics: R-squared (R^2), mean absolute error (MAE), and root mean squared error (RMSE). The mathematical expressions for each of these metrics used for model performance evaluation are shown in the equation (3), (4), and (5).

$$R^2 = 1 - \frac{RSS}{TSS} \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \bar{y}_i| \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{5}$$

Where R^2 is the coefficient of determination, RSS represents the residual sum of squares, and TSS represents the total sum of squares. n is the number of total observations, y_i is the ith value observed for series and $\bar{y}_i$ is the ith value predicted for the model.

3. Results and Discussion

3.1. Feature Scaling

In the dataset in our study, due to the different scale of various features which make machine learning difficult for training, we apply Standard Scaler function, a standardized method to make all input data at the

same scale. Figure 3 shows the data distribution plot of some independent variables after scaling with the standard scaler function.

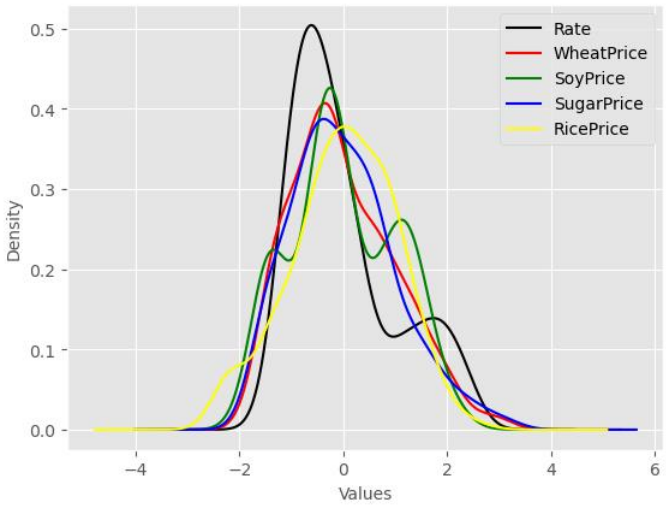

Figure 3. The data distribution after scaling process

3.2 Feature Selection

To the feature selection process discussed in Section 2, first, we applied the recursive feature elimination based on random forest to identify the optimal number of features required to achieve the highest score. Additionally, it has identified and selected the top 6 performing features from an initial set of 52 features, which form the appropriate datasets for training the machine learning models. The 6 chosen features by the RFE model include: 'MaiFutPrice,' 'CasPrice,' 'CasPlant,' 'MaiChg4', 'Rate,' and 'RicePrice.' Using these 6 features, the model achieved the highest R^2 score of 0.928, as shown in

Table **2**. Second, we applied the Pearson correlation feature selection technique, in which the input variable correlates with the target variable. In this case, features with a correlation coefficient (r) value more excellent than 0.6 or less than -0.7 were selected, and all p-values of those correlated features must be less than 0.05 to present a statistically significant linear relationship between the variables. This feature selection step resulted in the selection of 8 features considered strong positive and negative linear relationships with the target variable. Table *3* shows the Pearson correlation coefficient value and the p-value between target variable with strong linear relationship variables selected through this technique.

Figure 4. Graph of R^2 score by number of features using RFE feature selection method

Table 2. R^2 score of RFE method by number of features

#	R^2	#	R^2	#	R^2	#	R^2
52	0.913	39	0.913	26	0.915	13	0.919
51	0.915	38	0.914	25	0.914	12	0.918
50	0.912	37	0.912	24	0.915	11	0.919
49	0.913	36	0.913	23	0.915	10	0.922
48	0.913	35	0.913	22	0.915	9	0.921
47	0.913	34	0.912	21	0.915	8	0.924
46	0.913	33	0.912	20	0.915	7	0.924
45	0.913	32	0.913	19	0.915	**6**	**0.928**
44	0.912	31	0.916	18	0.916	5	0.9
43	0.913	30	0.914	17	0.918	4	0.888
42	0.912	29	0.914	16	0.917	3	0.877
41	0.912	28	0.914	15	0.918	2	0.814
40	0.913	27	0.916	14	0.918	1	0.649

Table 3. Correlation value and p-value of selected feature by Pearson correlation

No	Feature	Corr Value	p-value
1	CasPrice	0.85	$2e^{-74}$
2	MaiFutPrice	0.81	$6e^{-62}$
3	SoyPrice	0.80	$8e^{-59}$
4	RicePrice	0.79	$7e^{-56}$
5	WheatPrice	0.74	$3e^{-47}$
6	CasPlant	0.73	$2e^{-45}$
7	SugarPrice	0.60	$1e^{-24}$
8	Rate	-0.70	$6e^{-40}$

By integrating the RFE and Pearson correlation feature selection stages, we aim to optimize the dataset by keeping only the most relevant features for building the prediction model. We combine the features selected by those two methods together, resulting in 9 variables such as 'MaiFutPrice,' 'CasPrice,' 'CasPlant,' 'MaiChg4', 'Rate,' 'RicePrice,' 'SoyPrice,' 'WheatPrice,' and 'SugarPrice.' This approach reduces dataset dimensions, prevents the model from overfitting, and strengthens the model's ability to generate accurate forecasts while managing a concise set of features.

3.3 Experiment #1

In the first experiment, we employed various machine learning algorithms, including individual models like regression trees (RT) and support vector regression (SVR), as well as the ensemble bagging regressor (Bag-RT and Bag-SVR), and random forest (RF). We also utilized a multiple linear regression model (Lin), the proposed model by recent study of [20] on the same OAE dataset, to compare the performance with our proposed models. Our study utilized 260 complete records from the dataset and training using 10-fold cross-validation to ensure each fold was used as a test set exactly once. Each model was trained using the parameters outlined in Table 8 and tuned through the GridSearch CV function to achieve the best score. This table summarizes the results of hyperparameter tuning performed using the GridSearchCV function. It includes columns for the Model, Hyperparameter, Parameter List, and Best Parameter. Each row represents a different model, and the corresponding hyperparameters are tuned to enhance model performance. By systematically testing various hyperparameter configurations and selecting the ones that maximize performance metrics, GridSearchCV helps fine-tune models for better predictive accuracy and generalization.

Furthermore, we compared the performance of each model using "All Features" against "Selected Features" set to see the difference between these selections. Evaluation metrics, including R^2, MAE, and RMSE, were used to compare the model's performance, as shown in Table 4.

Table 4. Performance score comparison by each model on "All feature" and "Selected feature."

Feature	Model	R^2	MAE	RMSE
All Feature	Lin	0.865	0.456	0.599
	RT	0.822	0.476	0.638
	SVR	0.918	0.360	0.475
	Bag-RT	0.927	0.318	0.426
	Bag-SVR	0.926	0.344	0.459
	RF	0.91	0.375	0.487
Selected Feature	Lin	0.893	0.423	0.537
	RT	0.846	0.44	0.617
	SVR	**0.959**	**0.251**	**0.333**
	Bag-RT	0.94	0.292	0.396
	Bag-SVR	**0.961**	**0.234**	**0.315**
	RF	0.932	0.309	0.421

From a comparison of performance between using the "Selected Feature" set and a complete set of 52 features in Table 4, the "Selected Feature" set consistently demonstrates superior performance across all models. This suggests improved predictive accuracy and efficiency with the smaller feature set. The Bag-SVR and SVR model within the "Selected Feature" achieved the highest R^2 value of 0.961 and 0.959 as well as the lowest MAE value of 0.234 and 0.251 and RMSE value of 0.315 and 0.333, respectively, compared to all other models applied on the "All Feature" set.

3.4 Experiment #2

The second experiment involved dividing the data into different training sizes (70%, 75%, 80%, 85%, 90%, and 95%) to assess the model's predictive capability for each division aim to find the best training and testing set ratio which make the model achieved the best R^2 score to predict the future maize prices in the unseen data. The samples of each training size were trained using 10-fold cross-validation, and the rest were kept as a test set for future prediction. The outcomes of this analysis are illustrated in Table 5, and the graph of the prediction accuracy of each model in different training sizes is shown in Figure *5*.

Table 5. R^2 score by different training sets from 70 to 95 percent

Training		Lin	RT	SVR	Bag-RT	Bag-SVR	RF
%	Records						
70	182	0.898	0.899	0.942	0.901	0.951	0.9
75	195	0.894	0.863	0.938	0.931	0.95	0.91
80	208	0.895	0.870	**0.947**	0.927	**0.956**	0.921
85	221	0.894	0.863	0.941	0.932	0.953	0.927
90	234	0.869	0.856	0.909	0.916	0.929	0.903
95	247	0.858	0.863	0.847	0.863	0.868	0.871

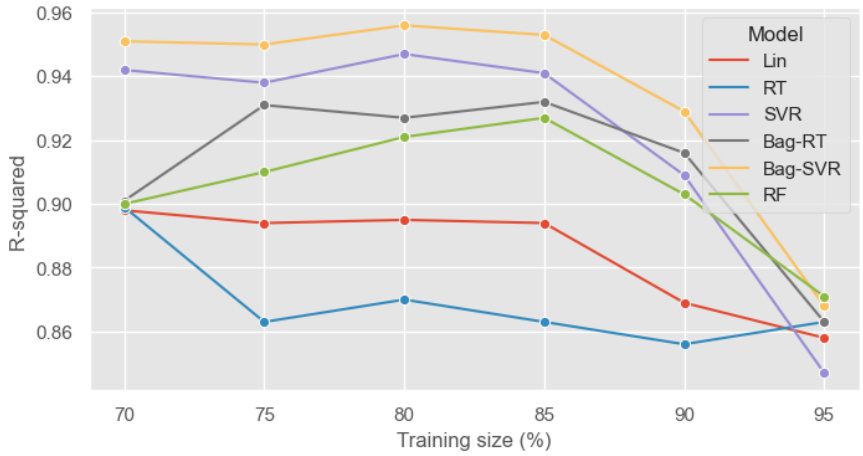

Figure 5. R^2 score of test set from different training size

The experiment's result in Table 5 and Figure *5* revealed that the Bag-SVR model using the "Selected Feature" set consistently demonstrated superior performance expressly when the training size was set to 80%, resulting in the highest R^2 score value of 0.956, followed by SVR model with R^2 value of 0.947. Based on this conclusion, we conducted 80% of the dataset as a training set using 10-fold cross-validation while allocating 20% in the most recent months for the testing set. This approach was employed to evaluate the mean absolute error (MAE) score, aiming to see the average price difference between the actual and predicted prices generated by each model and find the model that accurately forecasts the future price of maize. Figure *6* displays a graph visualizing the correlation between actual and predicted prices. This analysis incorporates all machine learning models applied to the "Selected Feature" by using each model's best hyperparameters obtained from the Gridsearch CV function. These models were utilized to predict the 20% of unseen data (test set), covering a range from May 2019 to August 2023, totaling 52 months.

Based on the graph in Figure 6, the models have superior predictive capabilities compared to the actual values within the first 10-month period of the test set. So, we segmented the test set into 10-month intervals and computed the MAE score, as outlined in Table 6. The first segment, from May 2019 to February 2020, notably stands out for its consistently low MAE values across multiple models. During this period, the SVR model demonstrated superior predictive capabilities with an impressive MAE score of 0.23, closely followed by the Bag-RT and Bag-SVR models, which had MAE values of 0.25 and 0.26, respectively. This exceptional performance is due to the fact that the test data is closely aligned with the previous training set. However, as we progress beyond this initial phase into the period from March 2020 to December 2020, while some models maintain relatively low MAE scores (e.g., RT and Bag-RT), for others, the MAE scores go up a bit, suggesting that maybe the data is changing or the models need a little fine-tuning. Moving further along the timeline, the MAE scores exhibit a gradual upward trend, particularly in the intervals from January 2021 to October 2021 and November 2021 to August 2022. During these periods, the models encountered more significant challenges in accurately predicting outcomes, as reflected in the higher MAE values recorded across all models. By the final segment, spanning from September 2022 to August 2023, the MAE scores peak, suggesting a significant divergence between the test data and the original training set, posing considerable challenges for the models in making accurate predictions.

Table 6. The MAE score of the test set split by 10 months range

Date Range	Lin	RT	SVR	Bag-RT	Bag-SVR	RF
May 19 - Feb 20	0.41	0.36	**0.23**	**0.25**	**0.26**	0.42
Mar 20 - Dec 20	0.34	0.26	0.40	0.25	0.62	0.32
Jan 21 - Oct 21	1.21	0.47	1.12	0.51	1.57	0.49
Nov 21- Aug 22	0.96	0.78	1.30	0.65	1.08	0.91
Sep 22 - Aug 23	1.49	2.08	3.24	2.05	3.02	2.22

Furthermore, we conducted training using all the observations while reserving only the most recent month, August 2023, with a value of 9.56 Baht per kilogram as a test set. This approach allowed us to observe and compare the predictions made by each model. Table 7 provides an analysis of different models' performance in predicting maize prices one month ahead by displaying the prices, along with the price difference between the predicted and actual values.

Table 7. The predicted price and price difference of each model on the test set of the last month

Model	Lin	RT	SVR	Bag-RT	Bag-SVR	RF
Predicted	8.71	7.73	9.9	8.68	9.8	9.12
Different	0.85	1.83	**0.34**	0.88	**0.24**	0.44

From the result of Table *7*, both the Bag-SVR and SVR models exhibited notable predictive accuracy in the last month of the test set. The Bag-SVR model showcased a closer estimation of the actual price, differing by only 0.24 baht/kg, followed by the SVR model, which demonstrated differing by 0.34 baht/kg.

4. Conclusions

This study aimed to enhance maize price forecasts in Thailand using various machine learning algorithms, targeting benefits for farmers, agribusinesses, and governmental policies. The Recursive Feature Elimination (RFE) combined with Pearson correlation successfully identified 9 key variables for training the machine learning models for feature selection. In building the prediction models, individual and ensemble techniques were employed to compare their accuracy in predicting maize prices. Notably, the Bag-SVR demonstrated superior predictive accuracy with an R^2 score of 0.961, MAE of 0.234, and RMSE of 0.315, followed closely by the SVR model with an R^2 score of 0.959, MAE of 0.251, and RMSE of 0.333. In contrast, the RT model achieved the lowest score, with an R^2 score of 0.846, MAE of 0.44, and RMSE of 0.617. In conclusion, this research underscores the importance of feature selection in refining model efficiency and emphasizes the potential of machine learning in enhancing maize price forecasts in Thailand.

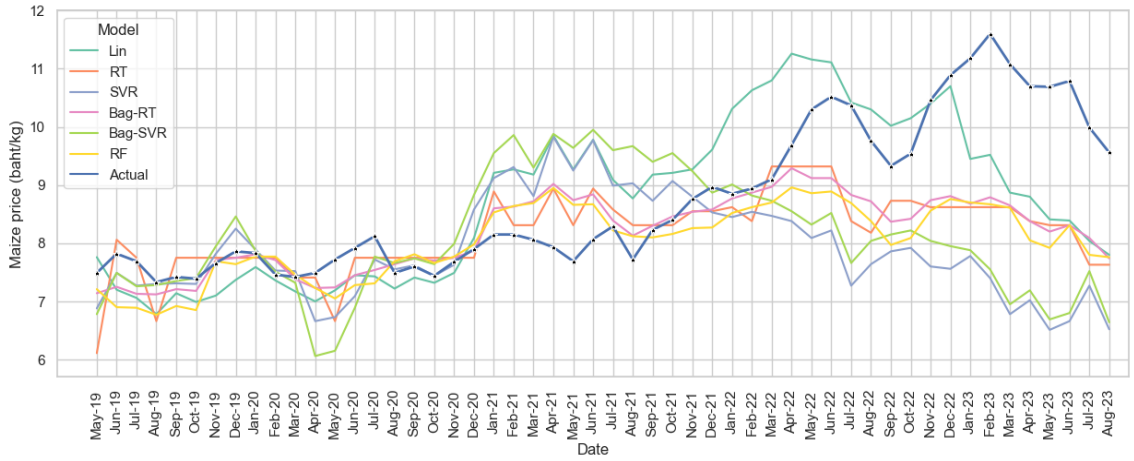

Figure 6. Actual maize price and predicted price by each model on test set

Table 8. The details of hyperparameter tuning of each model

Model	Hyperparameter	Parameter List	Best Parameter
RT	max_depth	[2, 5, 10, 20, 30, 40, None]	20
	min_samples_split	[1,2,3,4,5,6,7,8,9,10]	1
	min_samples_leaf	[1,2,3,4,5,6,7,8,9,10]	5
SVR	C	[10,20,30,40,50,60,70,80,90,100]	20
	gamma	['scale', 'auto']	scale
	kernel	['rbf','linear','poly']	'rbf'
Bag-RT	Bootstrap	[True, False]	False
	bootstrap_features	[True, False]	True
	max_features	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]	0.9
	max_samples	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]	0.9
	n_estimators	[10, 50, 100, 200]	50
Bag-SVR	Bootstrap	[True, False]	False
	bootstrap_features	[True, False]	False
	max_features	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]	0.8
	max_samples	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]	0.9
	n_estimators	[10, 50, 100, 200]	200
RF	max_features	[' sqrt', 'log2']	'sqrt'
	max_depth	[10, 20, 30, 40, None]	20
	min_samples_split	[1,2,3,4,5,6,7,8,9,10]	2
	min_samples_leaf	[1,2,3,4,5,6,7,8,9,10]	1
	n_estimators	[10, 50, 100, 200]	100

5. Acknowledgement

This work has been supported by the Royal Scholarship under Her Royal Highness Princess Maha Chakri Sirindhorn Education Project to the Kingdom of Cambodia.

Author Contributions: Conceptualization, N.S.; methodology, N.S, S.M.; software, S.M.; validation, N.S., S.M. data curation, N.S.; writing—original draft preparation, S.M.; writing—review and editing, N.S.; visualization, S.M.; supervision, N.S. All authors have read and agreed to the published version of the manuscript."

Funding: The APC was funded by Department of Computer Science, Faculty of Science, Kasetsart University.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

[1] Sjah, T.; Zainuri, Z. Agricultural supply chain and food security. In *Zero Hunger*, Springer, **2020**, pp 79-88.

[2] Warr, P.; Suphannachart, W. Agricultural productivity growth and poverty reduction: Evidence from Thailand. *Journal of Agricultural Economics* **2021**, *72*(2), 525-546.

[3] Jaipong, P.; Sriboonruang, P.; Siripipattanakul, S.; Sitthipon, T.; Kaewpuang, P.; Auttawechasakoon, P. A review of intentions to use artificial intelligence in Big Data Analytics for Thailand agriculture. *Review of Advanced Multidisciplinary Science, Engineering & Innovation* **2022**, *1*(2), 1-8.

[4] Basso, B.; Liu, L. Seasonal crop yield forecast: Methods, applications, and accuracies. *advances in agronomy* **2019**, *154*, 201-255.

[5] Shahhosseini, M.; Hu, G.; Pham, H. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Machine Learning with Applications* **2022**, *7*, 100251.

[6] Kantanantha, N.; Serban, N.; Griffin, P. Yield and Price Forecasting for Stochastic Crop Decision Planning. *Journal of Agricultural, Biological, and Environmental Statistics* **2010**, *15*(3), 362-380. DOI: 10.1007/s13253-010-0025-7.

[7] Wang, L.; Duan, W.; Qu, D.; Wang, S. What matters for global food price volatility? *Empirical Economics* **2018**, *54*, 1549-1572.

[8] Ge, Y.; Wu, H. Prediction of corn price fluctuation based on multiple linear regression analysis model under big data. *Neural Computing and Applications* **2020**, *32*, 16843-16855.

[9] Ya, Z.; Pei, K. Factors influencing agricultural products trade between China and Africa. *Sustainability* **2022**, *14* (9), 5589.

[10] Jain, A.; Marvaniya, S.; Godbole, S.; Munigala, V. A framework for crop price forecasting in emerging economies by analyzing the quality of time-series data. *arXiv preprint arXiv:2009.04171* **2020**.

[11] Mariammal, G.; Suruliandi, A.; Raja, S.; Poongothai, E. Prediction of land suitability for crop cultivation based on soil and environmental characteristics using modified recursive feature elimination technique with various classifiers. *IEEE Transactions on Computational Social Systems* **2021**, *8*(5), 1132-1142.

[12] Suruliandi, A.; Mariammal, G.; Raja, S. Crop prediction based on soil and environmental characteristics using feature selection techniques. *Mathematical and Computer Modelling of Dynamical Systems* **2021**, *27* (1), 117-140.

[13] Shahhosseini, M.; Hu, G.; Archontoulis, S. V. Forecasting corn yield with machine learning ensembles. *Frontiers in Plant Science* **2020**, *11*, 1120.

[14] Ribeiro, M. H. D. M.; Ribeiro, V. H. A.; Reynoso-Meza, G.; dos Santos Coelho, L. Multi-objective ensemble model for short-term price forecasting in corn price time series. In *2019 International Joint Conference on Neural Networks (IJCNN)*, **2019**, pp 1-8.

[15] Nikou, M.; Mansourfar, G.; Bagherzadeh, J. Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management* **2019**, *26*(4), 164-174.
[16] Khwaja, A. S.; Anpalagan, A.; Naeem, M.; Venkatesh, B. Joint bagged-boosted artificial neural networks: Using ensemble machine learning to improve short-term electricity load forecasting. *Electric Power Systems Research* **2020**, *179*, 106080.
[17] Li, Y.; Yu, R.; Shahabi, C.; Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, **2017**.
[18] Paudel, D.; Boogaard, H.; de Wit, A.; Janssen, S.; Osinga, S.; Pylianidis, C.; Athanasiadis, I. N. Machine learning for large-scale crop yield forecasting. *Agricultural Systems* **2021**, *187*, 103016.
[19] Li, J.; Chen, W. Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models. *International Journal of Forecasting* **2014**, *30*(4), 996-1015.
[20] Jantankaew, P.; Soonthornphisaj, N. Data Analytics for Maize Price Prediction using Regression Algorithms. *KKU Research Journal (Graduate Studies)* **2023**, *23*(2), 92-106.
[21] Roberts, S.; Osborne, M.; Ebden, M.; Reece, S.; Gibson, N.; Aigrain, S. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2013**, *371*(1984), 20110550.
[22] Kim, K.-j. Financial time series forecasting using support vector machines. *Neurocomputing* **2003**, *55* (1-2), 307-319.
[23] Ouyang, H.; Wei, X.; Wu, Q. Agricultural commodity futures prices prediction via long-and short-term time series network. *Journal of Applied Economics* **2019**, *22*(1), 468-483.
[24] Cerqueira, V.; Torgo, L.; Pinto, F.; Soares, C. Arbitrated ensemble for time series forecasting. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part II 10*, **2017**, pp 478-494.
[25] Oliveira, M.; Torgo, L. Ensembles for time series forecasting. In *Asian Conference on Machine Learning*, **2015**, PMLR: pp 360-370.
[26] Silva, R. F.; Barreira, B. L.; Cugnasca, C. E. Prediction of Corn and Sugar Prices Using Machine Learning, Econometrics, and Ensemble Models. *Engineering Proceedings* **2021**, *9*(1), 31.
[27] Breiman, L. *Classification and regression trees*; Routledge, 2017.
[28] Chen, R.; Liang, C.-Y.; Hong, W.-C.; Gu, D.-X. Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm. *Applied Soft Computing* **2015**, *26*, 435-443.
[29] Cortes, C.; Vapnik, V. Support-vector networks. *Machine learning* **1995**, *20*, 273-297.
[30] Dietterich, T. G. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, **2000**, pp 1-15.
[31] Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5-32.