Random forest model for forecasting vegetable prices: a case study in Nakhon Si Thammarat Province, Thailand

Sopee Kaewchada¹, Somporn Ruang-On¹, Uthai Kuhapong², Kritaphat Songsri-in³

¹Creative Innovation in Science and Technology Program, Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Nakhon Si Thammarat, Thailand

²School of Science, Walailak University, Nakhon Si Thammarat, Thailand

³Computer Science Program, Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Nakhon Si Thammarat,

Thailand

Article Info

Article history:

Received Oct 20, 2022 Revised Jan 12, 2023 Accepted Feb 4, 2023

Keywords:

Dataset Forecasting Machine learning Random forest model Vegetable price

ABSTRACT

The objectives of this research were developing a model for forecasting vegetable prices in Nakhon Si Thammarat Province using random forest and comparing the forecast results of different crops. The information used in this paper were monthly climate data and average monthly vegetable prices collected between 2011 – 2020 from Nakhon Si Thammarat meteorological station and Nakhon Si Thammarat Provincial Commercial Office, respectively. We evaluated model performance based on mean absolute percentage error (MAPE), root mean squared error (RMSE), and mean absolute error (MAE). The experimental results showed that the random forest model was able to predict the prices of vegetables, including pumpkin, eggplant, and lentils with high accuracy with MAPE values of 0.09, 0.07, and 0.15, with RMSE values of 1.82, 1.46, and 2.33, and with MAE values of 3.32, 2.15, and 5.42, respectively. The forecast model derived from this research can be beneficial for vegetable planting planning in the Pak Phanang River Basin of Nakhon Si Thammarat Province, Thailand.

This is an open access article under the <u>CC BY-SA</u> license.



Corresponding Author:

Somporn Ruang-On Creative Innovation in Science and Technology Program, Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University Tha Ngio Subdistrict, Muang District, Nakhon Si Thammarat 80280, Thailand Email: somporn_rua@nstru.ac.th

1. INTRODUCTION

Nakhon Si Thammarat is a province in the south of Thailand, where most of the population is engaged in agriculture. The main problems found in vegetable cultivation in the province are droughts. According to the statistics, Nakhon Si Thammarat experienced a total of 5 droughts during 2013 to 2019. In 2016, there were 12 districts with the highest drought level, and the agriculture was damaged by 883.54 square kilometres [1]. Besides the unfavourable climate, farmers face the problem of plant disease, pest infestation, and low consumer prices as farmers cannot set desired prices [2].

Although the price of vegetables has a large impact on the population, it is volatile and changes quickly. This makes it more difficult to predict future prices consistently. Nonetheless, vegetable price prediction is necessary for the general public to recognize the price of vegetables in advance [3].

There is currently a lot of research focusing on improving forecasting models to be more accurate by using modern statistical and computing methods such as machine learning (ML) and artificial intelligence (AI) depending on the goals and nature of the problem [4]. ML is a subdomain of AI [5]. It is a science of training computers to act without giving any command to it [6]. In AI, we make computers artificially more intelligent

as they perform tasks on their own. These systems are highly accurate and fast in doing their tasks. While in machine learning, we create and train a model using various techniques such as supervised learning, unsupervised learning, and reinforcement learning [6]. The data in machine learning is made up of examples, and each example is described by a set of attributes. These characteristics are also known as variables [7], [8]. There are two types of supervised learning: classification and regression. In particular, the dependent variable in the classification problem is discrete but continuous in the regression problem [9].

Random forest is a machine learning technique that employs a large number of classifications or regression sub-trees. It is a popular prediction algorithm because it is a versatile algorithm for analyzing large datasets. Furthermore, it has a high prediction accuracy and provides information on important variables for classification [10].

In previous research, a variety of machine learning techniques have been applied to data analysis in order to identify patterns and trends. For example, one study compared the performance of random forest and multiple regression models in predicting apartment prices [11], while another used linear regression and random forest regression to forecast ticket prices for public transportation [12]. In addition, decision trees and random forest models were utilized to predict crop prices [13], and machine learning methods were employed to forecast the prices of agricultural products [8] and used cars [14]. A comparison was also conducted on the efficiency of machine learning models for predicting bird's eye chili prices in Nakhon Si Thammarat province [15]. Moreover, deep learning has been applied to forecasting in some cases [16], [17]. However, using machine learning models with a small dataset to predict vegetable prices may overfit the dataset and might not be efficient. Therefore, we propose using random forest models to forecast vegetable prices in Nakhon Si Thammarat Province and comparing the results for different crops. As a result, we propose to i) use random forests to forecast vegetable prices in Nakhon Si Thammarat Province and comparing the results for different crops.

2. METHOD

2.1. Dataset

The Meteorological Station and the Provincial Commercial Office in Nakhon Si Thammarat province provided historical data on the climate and vegetable prices between 2011 and 2020 for this study in commaseparated values (CSV) file format. The dataset consists of 7 attributes, namely month, temperature (degree Celsius), rainfall (mm.), humidity (%), seasons, average price per month (Bath), and average price per year (Bath). The dataset contains no missing data nor any significant outliers. Table 1 displays the attributes and their data type of the dataset.

Table 1. List of attributes				
No	Attribute	Data Type		
1	Month	Date		
2	Temperature	Number		
3	Rainfall	Number		
4	Humility	Number		
5	Season	Number		
6	Average price per month	Number		
7	Average price per year	Number		

2.2. Research tools

In this study, we chose to run the experiments with Scikit-learn [18], Python's most comprehensive and open-source machine learning package. Scikit-learn covers four major machine learning topics: data transformation, supervised learning, unsupervised learning, and model evaluation and selection. Scikit-learn provides various ready-to-use pre-processing algorithms and machine learning models which can be directly applied to the collected dataset.

2.3. Research process

We followed the setup in [19] and divided the dataset into two parts for this study: the training set and the test set. The training set, which contains 84 data points (70%), is used to train the model. The test set, which contains 36 samples (30%), is reserved for measuring the performance of the models. Figure 1 [20] depicts a more detailed overview of how machine learning models are trained and tested.

2.4. Accuracy measures for forecasting

The performances of the models were measures with three metrics that are commonly used for regression problems. Particularly, we used mean absolute error (MAE), root mean squared error (RMSE), and

mean absolute percentage error (MAPE) [8], [21]. To formally quantify the metrics, let L_i and P_i be the observed price and the forecasted price of a data point i, respectively.

The MAE determines the average size of error in a series of forecasts without taking into account their direction. It is the test sample's average of the absolute disparities between prediction and actual observation, with all individual deviations given equal weight. It can be formally defined as (1).

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |L_i - P_i|$$
(1)



Figure 1. The overview of how the machine learning models is trained and tested [20]

The RMSE is the square root of the average of the error squares. It is, in other words, the average squared difference between the estimated and actual values. Because of its square design, serious mistakes are amplified and have a significantly greater effect on the value of the performance indicator. Simultaneously, the impact of relatively minor mistakes will be significantly reduced. This element of the squared error is sometimes referred to as penalizing excessive errors or being susceptible to outliers. It is mathematically defined as (2).

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (L_i - P_i)^2}$$
(2)

The MAPE is the extension of the MAE that satisfies the criteria of reliability, ease of interpretation, and clarity of presentation. It is formally defined as (3). Interpretation criteria to evaluate the performance of the predictive model using the MAPE are shown in Table 2 [22].

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{L_i - P_i}{L_i} \right| x 100\%$$
(3)

Table 2. Interpretation of typical MAPE values

1	~ 1
MAPE	Interpretation
<10	Highly accurate forecasting
10 to 20	Good forecasting
20 to 50	Reasonable forecasting
>50	Inaccurate forecasting

2.5. Random forest model

Random forest is an ensemble machine learning methodology that is a mixture of several tree-based predictors. It is a supervised method that can handle both regression (problems with continuous dependent variables) and classification (problems with categorical dependent variables) tasks. The core concept of the method is to integrate many decision trees to decide the final output rather than depending on individual decision trees, which reduces model variance [23]–[26]. Random forest constructs numerous versions of

decision trees by sampling different subsets of the given training data. These tree predictions are combined with a majority vote to get the final projection. As a consequence, over-fitting is reduced, and predicted accuracy is improved [27]. An overview of how the algorithms work is depicted in Figure 2. The random forest training algorithm is mainly defined as follows.

Algorithm:

Step	1:	From the dataset, pick M random records.
Step	2:	Based on M records, build a decision tree.
Step	3a:	From your algorithm, choose the number of trees and repeat steps 1 and $.2$
Step	3b: In	case of a regression problem, for a new record, each tree in the forest predicts
a val	lue for	Y (output).



Figure 2. General structure of a random forest [28]

For each sub-tree, the prediction function f(x) is defined as formulas (4) and (5) [29]

$$\mathbf{f}(\mathbf{x}) = \sum_{m=1}^{M} c_m \prod(\mathbf{x}, R_m) \tag{4}$$

where M is the number of regions in the feature space, R_m is a region corresponding to m, c_m is a constant corresponding to m:

$$\prod(\mathbf{x}, \mathbf{R}_m) = \{ 1, \text{ if } \mathbf{x} \in \mathbf{R}_m \text{ 0, otherwise}$$
(5)

The final classification decision is made from the majority a vote of all trees.

3. RESULTS AND DISCUSSION

3.1. Results

This study developed a random forest model for predicting vegetable prices in Nakhon Si Thammarat province using scikit-learn (random forest regressor). Six hyper-parameter combinations were investigated, specifically three estimator values 50, 100, and 150) and two max depth values 5 and 10). Table 3 displays the model's predicted outcomes.

The forecast model development results are shown in Table 3. Setting the number of estimators option to 50 and the maximum depth to 10 consistently results in the least amount of error in terms of MAE, RMSE, and

MAPE. According to Table 4, the MAPE for prediction accuracy was less than 10, indicating that the random forest model forecast was highly accurate for pumpkin and eggplant, while the result for lentils was good.

No	n_estimators	max_depth	Accuracy	Pumpkin	Eggplant	Lentils
			measures			
1	50	5	MAE	3.41	2.18	5.98
			RMSE	1.84	1.47	2.44
			MAPE	0.10	0.07	0.16
2	100	5	MAE	3.44	2.15	6.07
			RMSE	1.85	1.47	2.46
			MAPE	0.10	0.07	0.16
3	150	5	MAE	3.41	2.17	5.67
			RMSE	1.84	1.47	2.38
			MAPE	0.10	0.07	0.15
4	50	10	MAE	3.32	2.15	5.42
			RMSE	1.82	1.46	2.33
			MAPE	0.09	0.07	0.15
5	100	10	MAE	3.39	2.21	6.33
			RMSE	1.84	1.48	2.51
			MAPE	0.10	0.07	0.17
6	150	10	MAE	3.33	2.16	6.39
			RMSE	1.82	1.47	2.53
			MAPE	0.09	0.07	0.17

Table 3. The results of the development of the forecast model using the random forest

Table 4. Accuracy measures for forecasting pumpkin, eggplant, and lentils

Accuracy measures	Pumpkin	Eggplant	Lentils
for forecasting			
MAE	3.32	2.15	5.42
RMSE	1.82	1.46	2.33
MAPE	0.09	0.07	0.15

Table 5 compares the actual and expected costs of pumpkin, eggplant, and lentils over a 12-month period. Setting the number of estimators to 50 and the maximum depth to 10 yields the least error model. Figure 3 shows that anticipated vegetable prices were nearly identical to actual prices for the values of pumpkin in Figure 3(a), eggplant in Figure 3(b), and lentils in Figure 3(c).

Month	Pumpkin		Eggplant		Lentils	
wonun	Actual	Predicted	Actual	Predicted	Actual	Predicted
January	42.81	42.11	41.88	42.50	46.25	52.67
February	38.44	37.16	36.88	36.64	40.31	41.81
March	31.56	35.39	31.25	33.47	42.81	43.22
April	26.88	30.57	35.63	36.95	48.75	47.71
May	25.31	26.14	39.38	38.83	53.13	50.91
June	26.88	26.55	40.63	39.94	48.75	47.01
July	25.94	27.45	39.38	39.97	36.25	40.91
August	30.63	35.43	38.13	40.77	41.25	45.18
September	39.38	38.69	43.75	41.22	44.69	44.18
October	48.75	43.99	46.25	44.66	54.69	52.90
November	45.31	43.15	48.75	48.33	57.50	59.88
December	38.75	40.44	50.63	47.98	76.56	66.44

Table 5. Actual and predicted values of three vegetables in random forest model

3.2. Discussion

In this study, a random forest model was developed to predict vegetable prices in the province of Nakhon Si Thammarat. The results showed that the random forest model was an appropriate model for forecasting crop price because the forecasted outcomes were quite accurate. The findings are consistent with previous research, which found that random forest makes predictions with low RMSE and performs well with a high R-squared value [14]. Another study showed that random forest was a suitable model for predicting bird's eye chili prices in Nakhon Si Thammarat province [15]. A random forest approach for real-time price forecasting was discovered to be suitable and predict consistent results in the New York power market [30].

Furthermore, the random forest is used to predict house prices, with an error margin of 5 compared between anticipated and actual prices [31].



Figure 3. Actual and predicted values of three vegetables in random forest model; (a) actual and predicted values of pumpkin, (b) actual and predicted values of eggplant, and (c) actual and predicted values of lentils.

4. CONCLUSION

Forecasting vegetable prices is essential for farmers who want to know the price of their crops in advance. In this study, the random forest model was used to forecast vegetable prices. The study's data set, in particular, included seven characteristics. The prediction results showed that the random forest model was capable of accurately forecasting vegetable prices for pumpkin, eggplant, and lentils with MAPE values of 0.09, 0.07, and 0.15; RMSE values of 1.82, 1.46, and 2.33, and MAE values of 3.32, 2.15, and 5.42, respectively.

However, the model developed in this study was only applicable to climate and vegetable price data from Nakhon Si Thammarat Province. Additionally, the model user must consider additional factors such as soil conditions, pests, plant diseases, vegetable varieties, and so on. For future work, other types of vegetable can be studied. Additional independent variables can be used. To further improve prediction accuracy, different supervised learning approaches can also be explored.

ACKNOWLEDGEMENTS

The author would like to thank the Meteorological Station, Provincial Commercial Office, Nakhon Si Thammarat province, and Graduate School Nakhon Si Thammarat Rajabhat University.

REFERENCES

- Nakhon Si Thammarat Provincial Agriculture and cooperative Office, "Agricultural disaster prevention and mitigation plan during [1] the drv season 2019/2020." 2020.
- S. Buakhao, "The transfer of knowledge and technology for chili growers for the good agricultural practices standard (GAP) which [2] consistent of the market demand." 2021.
- M. Subhasree and C. A. Priya, "Forecasting vegetable price using time series data," International Journal of Advanced Research [3] (IJAR), vol. 3, pp. 535-641, 2016.
- A. Nansaior and A. Apichottanakul, "Sugar and raw sugar products export volumes forecasting models based on wavelet-nonlinear [4] autoregressive neural network," Khon Kaen Agriculture Journal, vol. 49, no. 1, pp. 179-191, 2021.
- D. Saha and A. Manickavasagan, "Machine learning techniques for analysis of hyperspectral images to determine quality of food [5] products: a review," Current Research in Food Science, vol. 4, pp. 28-44, 2021, doi: 10.1016/j.crfs.2021.01.002
- A. Pandey, V. Rastogi, and S. Singh, "Car's selling price prediction using random forest machine learning algorithm," SSRN [6] Electronic Journal, 2020, doi: 10.2139/ssrn.3702236.
- K. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: a review," Sensors, vol. 18, no. 8, [7] Aug. 2018, doi: 10.3390/s18082674.
- S. Bayona-Oré, R. Cerna, and E. T. Hinojoza, "Machine learning for price prediction for agricultural products," WSEAS [8] Transactions on Business and Economics, vol. 18, pp. 969–977, Jun. 2021, doi: 10.37394/23207.2021.18.92.
- [9] M. Mukhtar et al., "Hybrid model in machine learning-robust regression applied for sustainability agriculture and food security," International Journal of Electrical and Computer Engineering (IJECE), vol. 12, no. 4, pp. 4457–4468, Aug. 2022, doi: 10.11591/ijece.v12i4.pp4457-4468.
- [10] R. S and S. K. J., "Performance evaluation of random forest with feature selection methods in prediction of diabetes," International Journal of Electrical and Computer Engineering (IJECE), vol. 10, no. 1, pp. 353–359, Feb. 2020, doi: 10.11591/ijece.v10i1.pp353-359
- [11] M. Ballings, D. Van den Poel, N. Hespeels, and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," Expert Systems with Applications, vol. 42, no. 20, pp. 7046–7056, Nov. 2015, doi: 10.1016/j.eswa.2015.05.013.
- [12] Aditi, A. Dutta, A. Dureja, S. Abrol, and A. Dureja, "Prediction of ticket prices for public transport using linear regression and random forest regression methods: A practical approach using machine learning," in Data Science and Analytics, 2020, pp. 140-150. doi: 10.1007/978-981-15-5827-6_12.
- M. Rakhra et al., "WITHDRAWN: Crop price prediction using random forest and decision tree regression:-A review," Materials [13] Today: Proceedings, Apr. 2021, doi: 10.1016/j.matpr.2021.03.261.
- [14] P. Gajera, A. Gondaliya, and J. Kavathiya, "Old car price prediction with machine learning," International Research Journal of Modernization in Engineering Technology and Science, vol. 3, no. 3, pp. 284–290, 2021.
- S. Kaewchada, S. Ruang-on, and U. Kuhapong, "Comparison of the efficiency of machine learning model for predicting bird's eye chili prices in Nakhon Si Thammarat province," *PKRU SciTech Journal*, vol. 6, no. 2, pp. 1–11, 2022. [15]
- Y. Hong, Y. Zhou, Q. Li, W. Xu, and X. Zheng, "A deep learning method for short-term residential load forecasting in smart grid," [16] IEEE Access, vol. 8, pp. 55785-55797, 2020, doi: 10.1109/ACCESS.2020.2981817.
- [17] S. Ittisoponpisan, C. Kaipan, S. Ruang-on, R. Thaiphan, and K. Songsri-in, "Pushing the accuracy of thai food image classification with transfer learning," *Engineering Journal*, vol. 26, no. 10, pp. 57–71, Oct. 2022, doi: 10.4186/ej.2022.26.10.57. J. Hao and T. K. Ho, "Machine learning made easy: A review of scikit-learn package in python programming language," *Journal*
- [18] of Educational and Behavioral Statistics, vol. 44, no. 3, pp. 348–361, Jun. 2019, doi: 10.3102/1076998619832248.
- [19] D. P. Lestari and R. Kosasih, "Comparison of two deep learning methods for detecting fire hotspots," International Journal of Electrical and Computer Engineering (IJECE), vol. 12, no. 3, pp. 3118-3128, Jun. 2022, doi: 10.11591/ijece.v12i3.pp3118-3128.
- [20] M. K. Uçar, M. Nour, H. Sindi, and K. Polat, "The effect of training and testing process on machine learning in biomedical datasets," Mathematical Problems in Engineering, vol. 2020, pp. 1-17, May 2020, doi: 10.1155/2020/2836236.
- K. Gajowniczek and T. Ząbkowski, "Two-stage electricity demand modeling using machine learning algorithms," Energies, vol. [21] 10, no. 10, Oct. 2017, doi: 10.3390/en10101547.
- [22] J. J. M. Moreno, A. P. Pol, A. S. Abad, and B. C. Blasco, "Using the R-MAPE index as a resistant measure of forecast accuracy," Psicothema, vol. 25, no. 4, pp. 500-506, 2013, doi: 10.7334/psicothema2013.23.
- M. Vijh, D. Chandola, V. A. Tikkiwal, and A. Kumar, "Stock closing price prediction using machine learning techniques," Procedia [23] Computer Science, vol. 167, pp. 599-606, 2020, doi: 10.1016/j.procs.2020.03.326.
- S. Singh, T. K. Madan, J. Kumar, and A. K. Singh, "Stock market forecasting using machine learning: Today and tomorrow," in [24] 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Jul. 2019, pp. 738-745. doi: 10.1109/ICICICT46008.2019.8993160.
- [25] T. Ayhan and T. Uçar, "Determining customer limits by data mining methods in credit allocation process," International Journal of Electrical and Computer Engineering (IJECE), vol. 12, no. 2, pp. 1910–1915, Apr. 2022. doi: 10.11591/ijece.v12i2.pp1910-1915.
- Y. Grichi, Y. Beauregard, and T. M. Dao, "A random forest method for obsolescence forecasting," in 2017 IEEE [26] International Conference on Industrial Engineering and Engineering Management (IEEM), Dec. 2017, pp. 1602-1606. doi: 10.1109/IEEM.2017.8290163.
- S. R. Polamuri, D. K. Srinivasi, and D. A. K. Mohan, "Stock market prices prediction using random forest and extra tree regression," [27] International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 3, pp. 1224–1228, Sep. 2019, doi: 10.35940/ijrte.C4314.098319.
- B. Dai, C. Gu, E. Zhao, and X. Qin, "Statistical model optimized random forest regression model for concrete dam deformation [28] monitoring," Structural Control and Health Monitoring, vol. 25, no. 6, Jun. 2018, doi: 10.1002/stc.2170.
- [29] C. Dewi and R.-R. Chen, "Random forest and support vector machine on features selection for regression analysis," International iournal of innovative computing, information & control, vol. 15, no. 6, pp. 2027-2037, 2019, doi: 10.24507/ijicic.15.06.2027.
- [30] J. Mei, D. He, R. Harley, T. Habetler, and G. Qu, "A random forest method for real-time price forecasting in New York electricity 2014 IEEE pp. 1–5. market." in PES General Meeting / Conference & Exposition, Jul. 2014, doi: 10.1109/PESGM.2014.6939932.
- [31] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House price prediction using random technique," Procedia 199, forest machine learning Computer Science, vol. pp. 806-813, 2022 doi: 10.1016/j.procs.2022.01.100.

BIOGRAPHIES OF AUTHORS



Sopee Kaewchada Sopee Kaewchada Sope Kae



Somporn Ruang-On Sec received the B.Sc. degree in computer science from Rajabhat Phetchaburi Institute, Thailand, in 1995, the M.Sc. degrees in information technology from Sripatum University, Thailand, in 2003 and Ph.D. degree in Quality information technology from Phetchaburi Rajabhat University, in 2013. Currently, he is an Assistant Professor at the Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Thailand. He can be contacted at somporn_rua@nstru.ac.th.



Uthai Kuhapong D X S P received the B. Sc. degrees in Computer Education from Bansomdejchaopraya Teachers College Thailand, in 1991, M.Sc. degree in Information Science from King Mongkut's Institute of Technology Ladkrabang, Thailand, in 2004 and the Ph.D. degree in Computational Science from Walailak University, Thailand, in 2013 Currently, he is an Assistant Professor at the School of Science, Walailak University, Thailand. He can be contacted at uthai.ku@wu.ac.th.



Kritaphat Songsri-in b K S c finished MEng and Ph. D. in computing from Imperial College London in 2011 and 2020, respectively. He has been a lecturer in the department of computer science at Nakhon Si Thammarat Rajabhat University since 2020. His research interests include Machine Learning, Deep Learning, and Computer Vision. He has published in and is a reviewer for multiple international conferences and journals such as IEEE Transactions on Image Processing and IEEE Transactions on Information Forensics & Security. Dr. Songsriin was a recipient of the Royal Thai Government Scholarship covering his undergraduate and postgraduate degrees in 2010. He received the Best Student Paper Awards at the IEEE 13th International Conference for Automatic Face and Gesture Recognition (FG2018) and the 6th National Science and Technology Conference (NSCIC2021). In 2021, his PhD thesis received an award from the National Research Council of Thailand (NRCT). He can be contacted at kritaphat_son@nstru.ac.th.