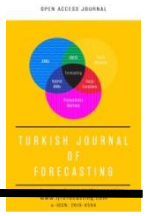
Turkish Journal of Forecasting vol. 8 no. 2 (2024) pp 54-64

Content list available at JournalPark

Turkish Journal of Forecasting

Journal Homepage: tjforecasting.com



Prediction for Türkiye's Tea Product With Machine Learning Algorithms

Mehmet Akif KARA[1]

[1]*Department of Business Administration, Faculty of Economics and Administrative Science, Giresun University, Gure Campus, 28200 Giresun, Türkiye*

Abstract

This study predicts tea production in Türkiye using machine learning algorithms. The analysis utilized data from 2001 to 2022, including tea production quantity, fresh tea prices, tea production area, temperature, and humidity. The study was conducted using the MATLAB 2023b Regression Learner toolbox. Initially, the obtained data were normalized, and then prediction performances were evaluated using various machine learning algorithms. The metrics used in the study included R^2, MAE, RMSE, and MSE. As a result, the Gaussian Process Regression algorithm emerged as the best-performing machine learning method.

Keywords: Machine Learning, Prediction, Tea, Agriculture, ANNs.

1. Introduction

In recent years, artificial intelligence applications have been increasingly utilized and developed across various scientific fields, particularly in statistics. In time series and prediction problems, artificial neural networks (ANNs) and machine learning algorithms are employed, demonstrating superior performance compared to classical forecasting methods (Eğrioğlu, Yolcu, & Baş, 2019; İslamoğlu, 2020; Tosunoğlu, 2021). Machine Learning (ML) is recognized as a subfield of Artificial Intelligence. The primary goal of the algorithms developed in this domain is to derive a mathematical model that accurately fits the data. Once the model effectively represents known data, it can be employed for making predictions with new data. Consequently, the learning process comprises two essential steps: estimating the unknown parameters of the model based on a given dataset, and, generating output predictions using new data alongside the parameters previously obtained (Cifuentes, et al., 2020).

The objective of Machine Learning (ML) methods aligns closely with that of statistical methods; both aim to enhance prediction accuracy by minimizing a loss function, typically the sum of squared errors. However, the key distinction lies in the approach to minimization: ML methods utilize non-linear algorithms, whereas statistical methods typically rely on linear processes. Additionally, ML methods are more computationally intensive, necessitating a greater reliance on computer science for implementation. This positions them at the intersection of statistics and computer science (Makridakis, et al., 2018).

Tea is cultivated in 30 countries, with seven multinational companies from the UK and Netherlands controlling 85% of the industry. India, China, and Sri Lanka produce 60% of global tea (Şahin Yıldırım, 2020: 371). Turkey is the largest per capita consumer, seventh largest producer, and third largest consumer of tea globally (Kara & Genç, 2022: 91). The tea sector is concentrated in the Eastern Black Sea region, particularly in Rize, Artvin, Trabzon, and Giresun, where companies like Çaykur and various cooperatives operate. The Turkish government sets floor prices and quotas to strengthen its position in the international market. As a crucial agricultural product, tea provides livelihoods for many in the Eastern Black Sea region. Agricultural growth underpins Türkiye's economy, fostering job creation and industrial development.

However, competition in agricultural trade has intensified due to geographic, climatic, and internal factors. To achieve sustainable success, Turkey must improve its agricultural performance, making effective prediction essential for strategic planning. A robust prediction model can mitigate risks in this process. Many tea farmers struggle with supply and demand analysis, leading to difficulties in predicting seasonal demand and resulting in price fluctuations and reduced income (Kara & Genç, 2022). Accurate prediction of agricultural production quantities is vital for developing strategic policies for enterprises

[1] Corresponding Author.
E-mail addresses: akifkara28@gmail.com

ORCID ID:
Mehmet Akif KARA: 0000-0003-4308-9933

and consumers, especially given the interplay between production levels and prices. Prediction involves estimating potential outcomes over time and, due to its forward-looking nature, carries inherent uncertainty. By analysing past data, prediction aids both businesses and individuals in making informed decisions about the future.

Prediction with machine learning algorithms has gained significant traction in various fields due to its ability to handle complex patterns and large datasets. Studies highlight the effectiveness of algorithms such as artificial neural networks (ANNs), support vector machines (SVM), and decision trees in improving prediction accuracy compared to traditional statistical methods. For instance, ANNs are praised for their capacity to model nonlinear relationships and their adaptability in time series prediction. SVMs are noted for their robustness in high-dimensional spaces, while decision trees offer interpretability and ease of implementation. Research also emphasizes the importance of feature selection and pre-processing in enhancing model performance. Recent advancements include the integration of ensemble methods, which combine multiple models to produce more reliable forecasts. Applications span diverse domains, including finance, agriculture, energy, and healthcare, demonstrating the versatility and effectiveness of machine learning in prediction tasks. Overall, the literature suggests that machine learning algorithms provide substantial improvements in prediction accuracy, particularly in complex and dynamic environments.

While there are no studies specifically addressing the quantity or price of tea products in the literature, national and international research has been conducted on prediction the production quantities of other agricultural products. In these prediction studies, new approaches such as artificial neural networks and machine learning are utilized alongside traditional methods. Gür and Ali (2024) conducted a comprehensive evaluation of LSTM, MLP, Random Forest, SVM, XGBoost, and linear regression models to prediction Turkey's scrap iron and steel imports. The performance of the models was measured using RMSE, MSE, MAE, MAPE, and R^2 metrics. The LSTM model demonstrated the best prediction performance on the training set, achieving an RMSE of 0.0387, an MSE of 0.0014, an MAE of 0.0297, a MAPE of 0.1261, and an R^2 of 0.9631. Bayyurt and Deveci Kocakoç (2023) predicted hazelnut production using the NARX model. In their study, Turkey's hazelnut production was treated as the dependent variable, while the independent variables included the simple price index of walnuts as a substitute product, the simple price index of hazelnuts, the number of fruit-bearing trees, temperature, and precipitation. The study analysed data from 1991 to 2021 and identified an optimal NARX model with 10 neurons in the hidden layer and a lag length of 4, concluding that the ANNs NARX model produced successful results in predicting hazelnut production. Sivaranjani and Vimal (2023) employed the NARX model to address the problem of predicting suitable crops for cultivation, using Mean Squared Error (MSE) and correlation coefficient (R^2) as performance metrics. Yıldırım and Karaatlı (2022) conducted a study applying the NARX model to forecast apple production, utilizing annual data from 1966 to 2018. In this study, apple production was the dependent variable, while the independent variables included the simple price index of apples, the simple price index of oranges, apple cultivation area, temperature, and technology. The results demonstrated that the ANNs NARX model could be effectively used to forecast apple production. Can and Gerşil (2018) predicted cotton prices using time series and artificial neural network techniques, comparing methods through MAE, MAPE, and RMSE values, ultimately concluding that the artificial neural network outperformed in predictions and forecasts. Karahan (2015) utilized an artificial neural network model to predict export quantities of dried apricots from Malatya, employing a feedforward backpropagation network known for its successful results in nonlinear problems. This model was also compared with ARIMA, showing superior performance. In their studies, Granata et al. (2017) employed Support Vector Regression and Regression Trees for the forecasting of wastewater quality indicators, demonstrating that both models exhibited robustness, reliability, and high generalization capability. As a result, when considering the coefficient of determination R^2 and mean square error, Support Vector Regression showed better performance than Regression Trees in predicting TSS, TDS, and COD. Khamis and Abdullah (2014) used backpropagation neural networks and NARX models to forecast wheat prices, incorporating the prices of oats, barley, and soybeans as variables. Their analysis revealed that the NARX model with 8 nodes in the hidden layer and a 4-lag delay line yielded better results.

2. Data and Method

2.1. Data

The dataset used in the study was compiled from reports by ÇAYKUR, the Meteorological Directorate, and TÜİK for 2023. The research utilized annual data on the purchase price of fresh tea, tea production area, tea production quantity, average humidity rate, and average temperature for provinces where tea is produced. The data covers the years 2001 to 2022. The table presents the dataset used in the study.

Table 1. Dataset

Years	Purchase Price($)	Production Quantity(kg)	Production Area(sqm.)	Average Humidity (%)	Average Temperature (C)
2001	0.25	824946	766530	69.21	15.53
2002	0.32	791700	766450	69.33	15.49
2003	0.40	869000	766400	70.74	15.56
2004	0.46	1105000	766320	71.02	15.44
2005	0.51	1192004	766250	71.26	15.41

2006	0.57	1121206	766136	71.35	15.39
2007	0.64	1145321	765808	71.15	15.37
2008	0.73	1100257	758257	70.85	15.35
2009	0.79	1103340	758513	70.59	15.34
2010	0.88	1305566	758641	70.36	15.33
2011	0.98	1231141	758895	71.03	15.28
2012	1.10	1250000	758566	69.57	15.27
2013	1.23	1180000	764255	69.02	15.24
2014	1.38	1266311	760494	68.94	15.22
2015	1.58	1327934	762073	69.57	15.19
2016	1.77	1350000	763609	68.84	15.16
2017	2.00	1300000	821079	68.97	15.14
2018	2.32	1480534	781334	69.55	15.13
2019	2.90	1407448	785693	69.51	15.10
2020	3.27	1450556	786813	69.67	15.08
2021	3.87	1453964	789001	69.69	15.05
2022	6.70	1269546	791285	70.04	15.03

Normalization techniques can vary widely, including methods such as Min-Max, Median, Sigmoid, and Z-Score. In this study, the commonly used D-Min-Max method was applied, normalizing the data to a range between 0.1 and 0.9. The choice of this method is supported by findings in various studies indicating that it yields better results (Yavuz & Deveci, 2012).

Normalization is performed using Equation 5:

D-Min-Max Normalized Value (x')= $0.8*\frac{x_i - x_{min}}{x_{max} - x_{min}} + 0.1$ (1)

In this equation:

(x'): the normalized value,
(x_i): the input value,
(x_{min}): the smallest number in the input set,
(x_{max}): the largest number in the input set.

The metrics used to evaluate model performance play a critical role in determining the reliability and applicability of the obtained results (Engin & İlter Fakhouri, 2024). In addition, various evaluation metrics such as MSE (Mean Squared Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), RMSE (Root Mean Squared Error), and R^2 (Coefficient of Determination) have been used to assess the models' performance during training, testing, and cross-validation.

$$RMSE = \sqrt{\frac{1}{ntest}\sum_{t=1}^{ntest}(x_t - \hat{x}_t)^2} \quad (2)$$

$$MSE = \frac{1}{ntest}\sum_{t=1}^{ntest}(x_t - \hat{x}_t)^2 \quad (3)$$

$$MAE = \sum_{i=1}^{ntest}\frac{|x_t - \hat{x}_t|}{ntest} \quad (4)$$

$$R^2 = 1 - \frac{\sum_{t=1}^{ntest}(x_t - \hat{x}_t)^2}{\sum_{t=1}^{ntest}(x_t)^2} \quad (5)$$

2.2. Method

The study employs several machine learning algorithms, including Robust Linear Regression, Regression Tree, Support Vector Machine (SVM), Ensemble Bagged Tree, and Gaussian Process Regression. The study utilized the Regression Learner Toolbox in MATLAB 2023b.

Robust Linear Regression: This method is designed to enhance the resilience of classical linear regression against outliers in the data. Outliers can negatively impact the predictive capability of the model; thus, robust linear regression is less sensitive to such values (Yu & Yao, 2017).

Regression Tree: Decision trees are used to predict continuous target variables in regression problems. They partition the data into a tree structure, making decisions at each node based on specific criteria. This approach is particularly effective in capturing the complexities within the dataset (Loh, 2011).
Support Vector Machine (SVM): While SVM is commonly used for classification tasks, it can also be applied to regression problems. Support Vector Regression (SVR) seeks to find a hyperplane that best separates the data points, focusing on minimizing the margin of error during this process (Pisner & Schnyer, 2020).
Ensemble Bagged Tree: This method relies on combining multiple decision trees to achieve improved predictive performance. The trees are trained on different subsets of the data, which enhances the overall model performance and reduces the risk of overfitting (Dietterich , 2000)
Gaussian Process Regression: This statistical model is used to model a continuous distribution of the data. It takes into account uncertainties in its predictions, presenting the results as a distribution. It is particularly effective in determining confidence intervals for the predictions (Williams & Rasmussen, 1995).
Each of these algorithms offers distinct advantages depending on the characteristics of the dataset and the problem being addressed. The selection of the appropriate algorithm can vary based on these factors, and typically requires a process of trial and error, as well as comparison, to identify which algorithm yields the best results.

3. Applications

As a result of the application, the performance of the machine learning algorithms was evaluated based on the RMSE, MAE, MSE, and R^2 metrics. Table 2 presents the data related to these values.

Table 2. Prediction Method Results

Methods	Criteria				
	RMSE	**R^2**	**MSE**	**MAE**	**Training Time (Sec.)**
Linear Regression	0.14828	0.71	0.021986	0.1165	45.532
Robust Linear R.	0.1521	0.70	0.023134	0.11823	36.764
Regression Tree	0.19723	0.49	0.038899	0.16723	22.959
SVM	0.14914	0.71	0.022242	0.10784	51.618
Ensemble Bagged Tree	0.21147	0.41	0.044719	0.14424	42.655
Gaussian Process Regression	**0.11316**	**0.83**	**0.012806**	**0.089363**	**71.392**

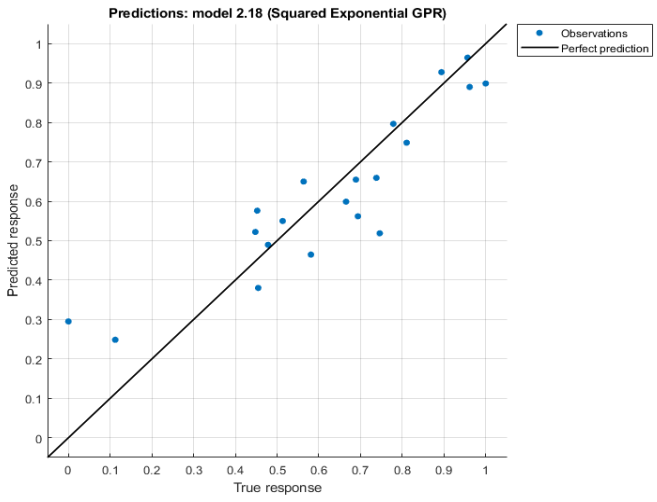


Figure 1. Gaussian Process Regression Predictions and True responses graph

The best result is obtained from Gaussian Process regression and the results are presented in table 2. Rational Quadratic function is used as kernel function in Gaussian Process regression. The explanatory power of the model is quite high and its prediction performance is better than the other methods according to all criteria.

4. Conclusions and Discussions

In this study, machine learning methods were employed to predict the future values of tea production in Turkey. The Gaussian Process Regression model demonstrated the best prediction performance among the time series data. These results have the potential to serve as a significant information source for decision-makers by providing insights into the future growth trends of tea production. This study conducts research on prediction tea production using machine learning algorithms, utilizing monthly data on tea production quantity, price, area, humidity, and temperature from 2001 to 2022. The MATLAB 2023b Regression Learner toolbox was employed in the study for ease of application.

Funding

No funding was received for this work

Credit authorship contribution statement

Author's full name: Conceptualization, Methodology, Software. **Author's full name:** Data curation, Writing, Original draft preparation. **Author's full name:** Visualization, Investigation. **Author's full name:** Supervision. **Author's full name:** Software, Validation. **Author's full name:** Writing, Reviewing, Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

If you have any (potential) conflict of interest, please state it here.

Data availability

References

[1] Bayyurt, D., & Deveci Kocakoç, İ. (2023). Yapay sinir ağlari narx ile Türkiye findik üretim miktari tahmini. *Giresun Üniversitesi İktisadi Ve İdari Bilimler Dergisi*, 9(1), 15-35. https://doi.org/10.46849/guiibd.1271782

[2] Can, Ş. & Gerşil, M. (2018). Manisa pamuk fiyatlarının zaman serisi analizi vey apay sinir ağır teknikleri ile tahminlenmesi ve tahmin performanslarının karşılaştırılması. *Yönetim ve Ekonomi Dergisi,* 25(3), 1071-1031. https:/doi.org/10.18657/yonveek.457761

[3] ÇAYKUR (2023). Çay İşletmeleri Genel Müdürlüğü, 2022 Yılı Çay Sektörü Raporu, Mayıs 2023. Erişim tarihi. 16.01.2023.

[4] Cifuentes, J., Geovanny M., Antonio B., and Javier R. (2020). Air Temperature Forecasting Using Machine Learning Techniques: A Review, *Energies*, 13(16), 4215. https://doi.org/10.3390/en13164215

[5] Dietterich, T.G. (2000). An experimental comparison of three methods for constructing ensembles of decision tress: Bagging, boosting, and randomization. *Machine Learning,* 40, 139-157.

[6] Eğrioğlu, E., Yolcu, U. & Baş, E. (2019). *Yapay sinir ağları.* Nobel Yayınları.

[7] Engin, E. & İlter Fakhouri, D. (2024). Comparison of machine learning algorithms for predicting financial risk in cash flow statements. *Turkish Journal of Forecasting*, 08(1), 1-12. https://doi.org/10.34110/forecasting.1403565.

[8] Granata, F., Papirio, S., Esposito, G., Gargano, R., & De Marinis, G. (2017). Machine learning algorithms for the forecasting of wastewater quality indicators. *Water*, 9(2), 105.

[9] Gür, Y. E., & Eşidir, K. A. (2024). Türkiye hurda demir çelik ithalatının gelecek değerlerinin derin öğrenme, makine öğrenmesi ve topluluk öğrenme yöntemleri ile öngörülmesi. *Alanya Akademik Bakış*, 8(3), 885-908. https://doi.org/10.29023/alanyaakademik.1497646.

[10] İslamoğlu, E. (2020). *Modern zaman serileri ve yöntemleri.* Nobel Yayıncılık.

[11] Kara, M.A. & Genç, K.Y. (2022). *Kooperatiflerde kurumsal yönetişim*. Eğiten Kitap.

[12] Karahan, M. (2015). Yapay sinir ağları metodu ile ihracat miktarlarının tahmini: ARIMA ve YSA metodunun karşılaştırmalı analizi. *Ege Akademik Bakış*, 15(2), 165-172.

[13] Khamis, A., & Abdullah, S. N. S. B. (2014). Forecasting wheat price using back propagation and NARX neural network. The International Journal of Engineering and Science, 3(11), 19-26.

[14] Loh, W.Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 14-23.

[15] Makridakis S, Spiliotis E, Assimakopoulos V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS ONE*, 13(3), e0194889. https://doi.org/10.1371/journal.pone.0194889

[16] Pisner, D.A. & Schnyer, D.M. (2020). Support vector machine. In *Machine Learning* (pp. 101-121). Academic Press.

[17] Şahin Yıldırım, E. (2020). Doğu Karadeniz'de Bir Kolektif Dayanışma: Hopa Çay Kooperatifi. *Journal of Sociological Research.* 23 (2), 357-391.

[18] Sivaranjani, T., & Vimal, S. P. (2023, January). Application of NARX Neural Network for Predicting Suitable crop for Cultivation-An Comparative analysis. In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 1333-1336). IEEE.

[19] Tarımsal Ekonomi ve Politika Geliştirme Enstitüsü (2023). Ürün Raporu Çay 2023. Erişim Tarihi: 16.01.2024.

[20] Tosunoğlu, N. (2021). *Zaman serilerinin öngörüsünde yapay sinir ağları.* Detay Yayıncılık.

[21] Williams, C., & Rasmussen, C. (1995). Gaussian process for regression. *Advances in Neural Information Processing Systems*, 8.

[22] Yavuz, S., & Deveci, M. (2012). İstatiksel normalizasyon tekniklerinin yapay sinir ağin performansina etkisi. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, (40), 167-187.

[23] Yildirim, H., & Karaatli, M. (2022). Yapay sinir ağlari narx modeli ile elma üretim miktarinin öngörülmesi. *Süleyman Demirel Üniversitesi Sosyal Bilimler Enstitüsü Dergisi* (42), 1-29.

[24] Yu, C. & Yao, W. (2017). Robust linear regression: A review and comparison. *Communications in Statistics-Simulation and Computation*, 46(8), 6261-6282.