Gaziosmanpaşa Üniversitesi Ziraat Fakültesi Dergisi
Journal of Agricultural Faculty of Gaziosmanpasa University
https://dergipark.org.tr/tr/pub/gopzfd

Araştırma Makalesi/*Research Article*

JAFAG
ISSN: 1300-2910
E-ISSN: 2147-8848
(2023) 40(2) 74-80
doi:10.55507/gopzfd.1320542

Estimating Corn Yield Using Statistical, Machine Learning and Deep Learning Methods

Cevher OZDEN[1*]
[1]Department of Agricultural Economy, Faculty of Agriculture, Cukurova University, Adana
***Corresponding author's email**: efeozden@gmail.com

Alındığı tarih (Received): 27.06.2023 Kabul tarihi (Accepted): 06.08.2023.

Abstract: Yield estimation is an important field of study in agriculture. Forecasting yields provides producers, consumers, traders and policymakers with important preliminary information and time to take necessary action. Corn is an important product in terms of international trade and is widely used in human and animal nutrition throughout the world. Adana produces the highest amount of corn sown both as main and secondary product in Türkiye. Therefore, in this study, corn yield was tried to be estimated by using various meteorological parameters and plant fertilizer usage amounts. For this purpose, statistical (Auto-ARIMA), machine learning (Random Forest) and deep learning (CNN, LSTM) methods were used. The study findings showed that all models used predicted maize yield highly accurately. However, the highest accuracy LSTM model estimated the yield of first corn crop.

Keywords: Yield estimation, deep learning, corn,

İstatistiksel, Makine Öğrenmesi ve Derin Öğrenme Yöntemleri ile Mısır Verimi Tahmini

Öz: Tarımda verim tahmini önemli bir çalışma alanıdır. Verimin önceden tahmin edilmesi üreticilere, tüketicilere, tüccarlara ve politika yapıcılara önemli ön bilgiler sunmakta ve gerekli tedbirlerin alınması için zaman sağlamaktadır. Mısır, dünya genelinde insan ve hayvan beslenmesinde yaygın olarak kullanılan, uluslararası ticaret açıdan da önemli bir üründür. Adana ülkemizde mısır üretiminin hem ana ürün olarak hem de ikincil ürün olarak en yüksek miktarda yetiştirilen ildir. Bu nedenle, bu çalışmada, çeşitli meteorolojik parametreler ve bitki gübre kullanım miktarları kullanılarak mısır verimi tahmin edilmeye çalışılmıştır. Bu amaç doğrultusunda, istatistiksel (Auto-ARIMA), makine öğrenmesi (Random Forest) ve derin öğrenme (CNN, LSTM) yöntemleri kullanılmıştır. Çalışma bulguları, kullanılan tüm modellerin mısır verimini yüksek oranda doğru tahmin ettiğini göstermiştir. Bununla birlikte en yüksek doğruluk LSTM modeli ile birinci mısır ürünü verimini tahminde bulunmuştur.

Anahtar Kelimeler: Verim tahmini, derin öğrenme, mısır

1. Introduction

Turkey is among the countries that produce a significant amount of corn worldwide. According to the data of the Turkish Ministry of Agriculture and Forestry, annual corn production in Turkey has been changing between 5-7 million tons on average in recent years (TSI, 2023). Turkey has an important ranking in world corn production. According to the amount of production, Turkey is generally among the top 20 countries in world corn production (FAOSTAT, 2023). However, the exact ranking may vary depending on the year and the amount of production. Turkey is a favorable country for corn production in terms of climate and soil characteristics. Corn is a plant that can be grown in various regions of Turkey. Corn production is common especially in Marmara, Aegean, Central Anatolia and Southeastern Anatolia regions.

As a fundamental agricultural product, corn is an important food source for human consumption and is widely used as animal feed. For Turkey's livestock sector, corn is a key ingredient for feed production. Corn production contributes to the provision of products such as maize flour and maize silage used as animal feed. The agricultural sector in Turkey is an important part of the economy. Corn is a product that plays an important role in agricultural exports and contributes to the agricultural economy. Increasing the domestic production of corn encourages the growth of the agricultural sector and reduces foreign dependency.

The Mediterranean Region, especially Çukurova, ranks first in Turkey's grain corn production. Last year, Adana ranked first with 1.036.130 tons of grain corn production in Turkey, which was 5 million 900 thousand tons. Adana province produces approximately 17.5% of Turkey's total corn production. Corn production in Adana can be done as the main crop or as a secondary crop. Since 2011, corn cultivation area and production as the primary product has followed a fluctuating but increasing course. As a second crop, a contraction is observed in its cultivation (Table 1).

Table 1. Corn cultivation in Adana Province
Çizelge 1. *Adana ili mısır üretimi*

	Sown Area (da) Primary Product	Sown Area (da) Second Product	Yield (kg/da) First Product	Yield (kg/da) Second Product	Production (ton) First Product	Production (ton) Second Product
2011	564.470	330.622	947	685	534.295	226.449
2012	540.837	251.452	923	729	499.148	183.314
2013	652.373	246.119	1.108	785	722.769	192.515
2014	720.870	198.710	1.180	780	850.720	154.931
2015	772.689	185.131	1.109	857	856.808	158.620
2016	811.721	155.649	1.145	1.010	929.455	157.151
2017	761.871	169.937	1.145	964	872.314	163.816
2018	611.651	127.778	1.162	1.031	710.940	131.757
2019	568.470	97.174	1.106	917	628.713	89.089
2020	624.954	76.596	1.188	1.011	742.503	77.475
2021	571.112	93.075	1.294	765	738.956	71.189
2022	780.283	86.700	1.051	788	820.050	68.298

Source: TSI, 2023

Yield estimation in agricultural activities provides producers with the opportunity to predict the amount of future production. These forecasts allow producers to plan production and effectively manage resources (water, fertilizer, seeds, etc.). Agricultural enterprises can determine the sowing time, plan the harvesting processes and adjust the production amount according to market demands in line with yield estimates. Agriculture is exposed to many risk factors. Factors such as climate changes, diseases, pests, natural disasters can negatively affect yield. Yield forecasting guides producers in developing risk management strategies by evaluating the effects of these risk factors. For example, in the case of low expected yields, producers may consider alternative product options or resort to risk reduction tools such as insurance. Yield forecasting plays an important role in the marketing and trading processes of agricultural products. Estimation of production quantity can be matched with market demand and adjusted to trade processes. It provides basic information on yield estimation, export planning and negotiation, especially in export-oriented agricultural products. Yield estimation helps to manage agricultural resources (water, soil, energy, etc.) efficiently. Forecasts contribute to making more informed decisions on issues such as water resources and irrigation schedules, fertilizer use, pesticide applications, and management of other agricultural inputs. This ensures the adoption of sustainable agricultural practices and the protection of natural resources.

Agricultural crop yield estimation has been studied as an important issue in the agricultural sector. Research in this field consists of the use of various methods and analyzes on different agricultural products. Matsuura et al. (2014) discuss a comparison of regression and machine learning models for maize yield prediction in Jilin, China. Analyzes using various datasets focus on identifying the ANN model with the best forecasting performance. Sharifi (2020) deal with yield prediction of field crops using satellite imagery and machine learning techniques. The data obtained from satellite images are combined with various machine learning algorithms to evaluate the performance of the models used in yield estimation. Fathima et al. (2020) present an agricultural crop yield estimation method based on climate data and data mining techniques. Relationships between climate data and agricultural data are analyzed using data mining techniques and models are built to predict future crop yields. Joshi et al. (2023) discuss a deep learning-based agricultural crop yield estimation method. Field images obtained using remote sensing data are processed with deep learning algorithms and models used to predict agricultural crop yields are developed. Kundu et al. (2022) consider yield estimation of field crops using machine learning methods and model coupling techniques. It is aimed to obtain more accurate and reliable yield estimates by using various machine learning algorithms and model fusion techniques. Paudel et al. (2022) make regional agricultural crop yield prediction based on ensemble machine learning methods through a case study in Shandong Province, China. The current study employs various methods from statistical, machine learning and deep learning fields. In this way, the study results will shed light on the applicability of different methodologies commonly used for prediction purposes in separate studies. The results are given in a comparative way to determine the best efficient method for prediction corn yield in Turkey. By combining different machine learning algorithms, it is tried to produce more accurate and reliable agricultural product yield predictions.

In this study, it is aimed to make future yield estimations by determining external factors such as climate and fertilization that affect corn yield in Adana province. For this purpose, Auto-ARIMA model, which is the most widely used statistical method, Random Forest, which is one of the most powerful machine learning methods, and Convolutional Neural Network (CNN) and Long Short-Term Memory neural networks, which are deep learning methods that have been increasingly used in recent years, were compared. has been applied. The data set and the codes of the applications produced within the scope of the study were shared in the public github repository for the benefit of other researchers.

2. Material and Methods

When corn is planted as the main product in Adana, it is planted in April - May and harvested in July-August. As a second crop, it is planted at the beginning of July at the latest and is harvested in September-October. For this reason, meteorological parameters of the April-October months measured in Adana province were taken from the General Directorate of Meteorology. These data include monthly average 10 cm soil temperature, average minimum temperature, average temperature, average wind speed and monthly total precipitation. In addition to these data, the amounts of nitrogen (N), phosphorus (P2O2) and potash (K2O) used in Adana province were included in the study as plant nutrition inputs. Thus, the amount of yield (kg/da) was tried to be estimated with 39 inputs (Table 2). Some important meteorological parameters such as evapotranspiration, global radiation etc. could be included in the study due to the large missing observations in their times series.

Table 2. Input variables
Çizelge 2. *Girdi değişkenleri*

Inputs	Input Category and Metric Unit
Nitrogen (N)	Numeric – tons
Phosphorus (P2O5)	Numeric - tons
Potas (K2O)	Numeric – tons
Total Plant Nutrient	Numeric - tons
Mean Soil Temperature at 10cm x 7 months (April-October)	Numeric - ^{0}C
Monthly Mean Minimum Temperature x 7 months (April-October)	Numeric – ^{0}C
Monthly Mean Wind Speed (April-October) x 7 months (April-October)	Numeric - m/sec
Monthly Mean Temperature x 7 months (April-October)	Numeric - ^{0}C
Monthly Total Precipitation x 7 months (April-October)	Numeric - mm

2.1. ARIMA

For the Auto-ARIMA method, it was tested for stationarity with the Augmented Dickey-Fuller test and the first differences of the non-stationary series were made stationary by taking the first difference and then standardized with the Min-Max scaling method. For CNN and LSTM methods, stationarity tests were not performed on the data.

ARIMA (Autoregressive Integrated Moving Average) is a model for analyzing a time series data, making predictions and estimating its future values. Auto-ARIMA chooses the best model by trying different ARIMA models to analyze time series data (Yermal & Balasubramanian, 2017). This method helps the model to automatically determine the optimal p, d, and q parameters. where p is the number of autoregressive terms; d is the degree of difference of the data; and q is the number of moving average terms.

$$(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p)(1 - B)^D Y_t = (1 + \varphi_1 B + \varphi_2 B^2 + \cdots + \varphi_Q B^Q)\varepsilon_t$$

Where:

- Y_t is the value of the time series at time t,
- B is the backshift operator that represents the lag operator,
- θ_1, θ_2, … θ_p are the autoregressive parameters where P is the maximum order of autoregression,
- D is the maximum order of differencing,
- φ_1, φ_2, …, φ_Q are the moving average parameters where Q is the maximum order of the moving average,
- ε_t represents white noise, a random error term with mean zero and constant variance

The working principle of Auto-ARIMA starts by making the data stationary by determining the degree of difference applied to the data. A series of ARIMA models are then automatically generated using different p, d, and q values, and these models are applied to the residuals of the data. This process relies on various statistical criteria (eg AIC, BIC) to evaluate the performance of the models and select the best model. The best model is considered the one that provides the best fit and the lowest information loss.

2 2. CNN (Convolutional Neural Network)

CNN is an artificial neural network model that is widely used in the field of deep learning and is especially effective in visual data analysis such as image recognition, object detection and image classification. This model is inspired by biological neural networks and uses convolutional layers to detect local patterns of data. It consists of convolutional

layers, activation functions, pooling and fully-connected layers.

Pooling Layers: Pooling layers are used to reduce the spatial dimensions of the feature maps, reducing the number of parameters in the model and aiding in translation invariance. The most common pooling operation is max-pooling, which takes the maximum value within a small region of the feature map.

Activation Functions: Non-linear activation functions, such as ReLU (Rectified Linear Unit), are applied after convolutional and pooling layers to introduce non-linearity into the network, enabling it to learn complex patterns in the data.

Fully Connected Layers: After several convolutional and pooling layers, the feature maps are flattened and connected to one or more fully connected layers, which act as a traditional neural network for making final predictions.

Weight Sharing: One key feature of CNNs is weight sharing, which allows the same set of learnable filters to be applied to different parts of the input image. This helps reduce the number of parameters in the model and enables CNNs to generalize well to different locations of features in the input.

The CNN model is created by iteratively combining these basic components. Usually multiple convolutional layers and pooling layers are followed by fully connected layers (Kim, 2017). This architecture can produce effective results by learning hierarchical properties of data and offering advanced pattern recognition capabilities. Convolutional Layers: The fundamental building blocks of CNNs are convolutional layers. These layers apply convolution operations to the input image to extract various features. Convolution involves sliding a small filter (also called a kernel) over the input image and computing element-wise multiplications and summations to produce feature maps.

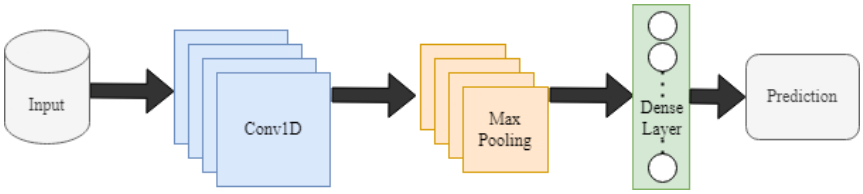

Figure 1. CNN architecture
Şekil 1. *CNN mimarisi*

2.3. LSTM (Long Short-Term Memory)

LSTM is a type of recursive neural network that is particularly effective in modeling data with long-term dependencies, such as time series data. It is mainly designed to address the vanishing gradient problem and effectively capture long-term dependencies in sequential data (Hochreiter & Schmidhuber, 1997). The key feature of LSTM is that it can more effectively handle long-term addictions compared to traditional recursive neural networks (Figure 2). This is accomplished using the cell state. The cell state is like a line that carries information inside the network and is maintained or changed over time. Another important component of LSTM are three gates used to control input data and make decisions: Forget Gate, Input Gate and Output Gate. These components of the LSTM enable the model to make better predictions by effectively capturing long-term dependencies and preserving important information.

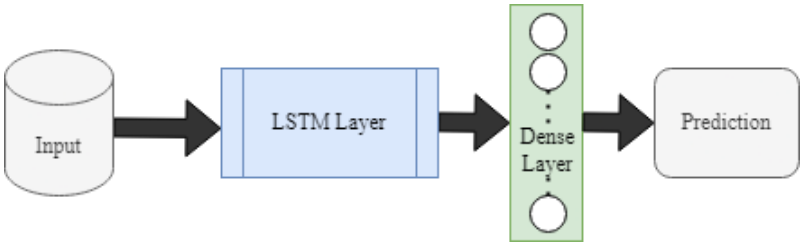

Figure 2. LSTM architecture
Şekil 2. *LSTM mimarisi*

2.4. Random Forest

Random Forest is an ensemble learning method that is widely used in the field of machine learning. This method is an assembly created by combining many decision trees. Each tree is trained on random samples of the dataset and makes predictions independently (Biau & Scornet, 2016). The formula for building a Random Forest model involves two main components: the process of creating individual decision trees and the voting mechanism to combine their predictions.

1. Building individual trees:

a. Selecting a Subset of Features: Given a dataset with "n" samples and "m" features, at each node of the decision tree, a random subset of features (typically denoted as "k") is selected to split the data. The value of "k" is usually much smaller than "m," and it remains constant throughout the tree-building process.

b. Data Bootstrapping: For training each tree, a random subset of the original data is sampled with replacement. This process, known as bootstrapping, creates a new dataset of the same size as the original but with some duplicate and missing samples.

c. Building Decision Trees: Using the selected features and the bootstrapped dataset, a decision tree is built recursively by selecting the best feature and split point at each node, based on criteria such as Gini impurity (for classification) or mean squared error (for regression).

2. Voting mechanism:

a. Classification: For classification tasks, each tree in the Random Forest independently predicts the class label of a sample. The final prediction is determined through majority voting, i.e., the class that receives the most votes from the individual trees is considered the final predicted class.

b. Regression: For regression tasks, the output of each tree in the Random Forest is a numerical prediction. The final prediction is obtained by averaging the predictions from all the individual trees.

Pmdarima Python library is used for the application of the Auto-ARIMA method. The implementation of the Random Forest model is made with the help of the Scikit-Learn library. The simplest possible architectural structure was created for the CNN and LSTM models. The CNN architecture is composed of 1 Conv1D layer with 63 neurons, 1 MaxPooling with pool size 2, 1 Dense Layer with 50 neurons and 1 Dense layer as output (Figure 1). On the other hand, LSTM architecture contains 1 LSTM layer with 50 neurons and 1 Dense layer as output (Figure 2). ReLu is chosen as activation and Adam is used as optimizer with a learning rate of 0.005, and both models were trained for 1000 epochs.

3. Results and Discussion

The data used in the study consists of monthly observations between 2011 and 2022. A total of 39 inputs including weather parameters, plant fertilizer and nutrients are used as inputs in order to predict the corn yield cultivated as first and second product in Adana province. The correlations between inputs and yield are shown in Figure 1. When there is no correlation between 2 variables (when correlation is 0 or near 0) the color is gray. The darkest red means there is a perfect positive correlation, while the darkest blue means there is a perfect negative correlation. Further details can be seen at the public Github repository (https://github.com/cevher/corn-yield-prediction/tree/master).

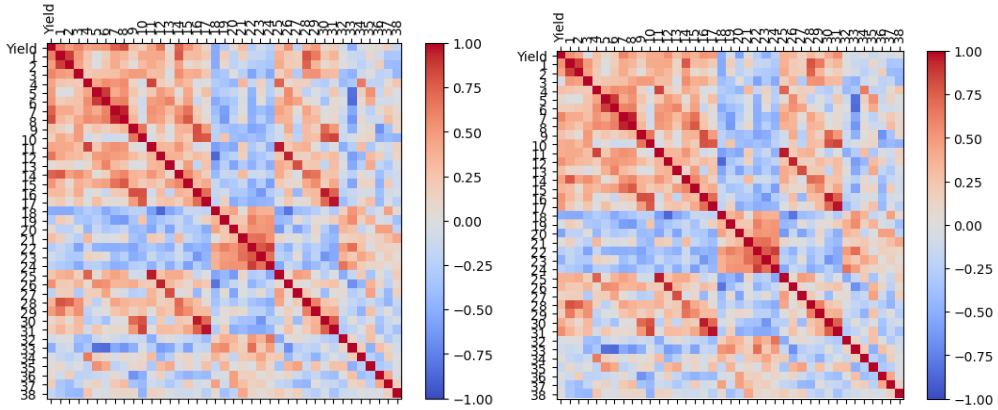

Figure 3. Correlation results between inputs and yield (1st and 2nd product, respectively)
Şekil 3. *Girdi değişkenleri ile verim (ilk ve ikinci ürün) arasındaki korelasyon sonuçları*

Table 3. Analysis results
Çizelge 3. *Analiz sonuçları*

	Corn Yield (First Product)			Corn Yield (Second Product)		
	MAE	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**
Auto-ARIMA	0.2996	0.1024	0.3201	0.3151	0.1512	0.3888
Random Forest	0.3028	0.1040	0.3226	0.4082	0.1668	0.4084
CNN	0.2679	0.0724	0.2691	0.4362	0.2310	0.4806
LSTM	0.2471	0.0612	0.2475	0.5063	0.3379	0.5813

Prior to the application of Auto-ARIMA model, data was checked for stationarity using Augmented Dickey-Fuller test and first difference of data provided stationarity. Then, Min-Max scaler was used to standardize variables in order to eliminate any possible bias among data. Subsequently, data was split into 90% training and 10% test set. All models were applied separately to predict corn yield of the first and second cultivated products. Mean Squared Error, Mean Absolute Error and Root Mean Squared Error metrics were used to evaluate the accuracy of models. The results are summarized in Table 3.

The results indicate that all models are robust in predicting on yield within statistically acceptable level. However, it is noteworthy that models are more capable of predicting the yield for the 1st product corn. The highest accuracy was obtained with LSTM model for the yield of the first product corn. This was closely followed by CNN, Auto-ARIMA and Random Forest models. The study results clearly show that corn yield can be predicted with any of the models using the selected weather and plant nutrient input variables. The previous studies mainly used either remote sensing images or deep architectures of neural networks. To our knowledge, LSTM and CNN have not been used in predicting agricultural product yield in comparison with other common statistical methods. The results of this study reveal that corn yield can be accurately estimated by using the simplest possible architectures of CNN and LSTM when there are enough meteorological and nutrient input observations collected between planting and harvesting periods.

4. Conclusion

This study makes a comparative analysis on the prediction of corn yield using various statistical, machine learning and deep learning models. The dataset contains the monthly observations of weather parameters and annual amount of plant nutrients, fertilizers to forecast the corn yield cultivated as 1st and 2nd product in Adana Province, which is the biggest corn producing city in Turkey. Corn yield is quite significant as corn is consumed by humans and used intensively in animal feed production. Statistical, machine learning and deep learning models are used for this prediction purpose in a comparative way. All models are found quite apt at learning the relations between input variables and corn yield and provided accurate yield predictions for both first and second product corn. However, the models provided slightly better results for the first product corn. Overall, LSTM yielded the highest accuracy in predicting the yield of the first product corn. It was closely followed by CNN, Auto-ARIMA and Random Forest models. Further studies can be implemented to generalize the findings of the current study over other corn production areas. Also, this methodology can be extended to other agricultural products and additional meteorological and agricultural observations can be included in the analysis, as well.

References

Biau, G., & Scornet, E. A. (2016). Random forest guided tour. *TEST* 25, 197–227 doi.org/10.1007/s11749-016-0481-7

FAOSTAT, (2023). Food and Agriculture Organization of the United Nations, *Crops and Livestock Products Statistics*, https://www.fao.org/faostat/en/#data/QCL

Fathima, M., Sowmya K., Barker, S., & Kulkarni, S. (2020). Analysis of Crop Yield Prediction using Data Mining Technique. *International Research Journal of Engineering and Technology.* 07(5) 10.13140/RG.2.2.14424.52482.

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-term Memory. *Neural computation*. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

Joshi, A., Biswajeet P., Shilpa, G, & Subrata C. (2023). Remote-Sensing Data and Deep-Learning Techniques in Crop Mapping and Yield Prediction: A Systematic Review. *Remote Sensing.* 15(8). 10.3390/rs15082014

Kim, P. (2017). Convolutional Neural Network. In: MATLAB Deep Learning. *Apress, Berkeley, CA.* doi.org/10.1007/978-1-4842-2845-6_6Kundu, S., Ghosh, A., Kundu, A. & Girish P. (2022) A ML-AI Enabled Ensemble Model for Predicting Agricultural Yield, *Cogent Food & Agriculture*, 8:1, 10.1080/23311932.2022.2085717

Matsuura, K., Gaitan, C., Hsieh, W., & Cannon, A. (2014). Maize yield forecasting by linear regression and artificial neural networks in Jilin, China. *The Journal of Agricultural Science*. 10.1017/S0021859614000392.

Paudel, D., Boogaard, H., Wit, A., Velde, M., Claverie, M., Nisini, L. Janssen, S., Osinga, S., & Athanasiadis, I. (2022). Machine learning for regional crop yield forecasting in

Europe, *Field Crops Research*, 276, 10.1016/j.fcr.2021.108377.

Sharifi, A. (2020). Yield prediction with machine learning algorithms and satellite images. *Journal of the Science of Food and Agriculture*. 101. 10.1002/jsfa.10696.

TSI, (2023). Turkish Statistical Institute, *Crop Production Statistics*,https://data.tuik.gov.tr/Kategori/GetKategori?p=tarim-111&dil=2

Yermal, L. & Balasubramanian, P. (2017). Application of Auto ARIMA Model for Forecasting Returns on Minute Wise Amalgamated Data in NSE, *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, India*, 2017, pp. 1-5, doi: 10.1109/ICCIC.2017.8524232.