Selcuk Journal of Agriculture and Food Sciences

https://dergipark.org.tr/tr/pub/selcukjafsci

Research Article

(2024) 38(3), 445-462 DOI:10.15316/SJAFS.2024.040 e-ISSN: 2458-8377 Selcuk J Agr Food Sci

Sunflower Crop Yield Prediction Using Machine Learning Methods

Seda Hatice Gökler[1,*]

[1]Kahramanmaraş Sütçü İmam University, Industrial Engineering, Kahramanmaras, Türkiye

HIGHLIGHTS

- Sunflower plant is affected by many factors
- Providing a timely and robust prediction for sunflower crop yields
- Increasing the effectiveness of the AI methods using Halving Grid Search method

Abstract

Sunflower, one of the most important crops, is produced in many countries to meet especially for edible oil demand. Since the sunflower plant is affected by many factors, such as the amount of rain and air temperature, the yield changes from year to year, which has adverse effects on the balance between demand and supply. Because of the product produced in many countries is not enough; it has to be imported. Turkey is one of the world's leading sunflower importers. The yield must be accurately estimated for the imported quantity to be correct. Importing in large quantities causes inventories, while small quantities cause the sunflower oil demand to not be met. It is used methods such as the direct method, simulation, and remote sensing to estimate sunflower yield. However, these methods have some shortcomings. In this article, machine learning methods, such as decision tree (DT), support vector machine (SVM) and random forest (RF), are used for production prediction. In order to increase the effectiveness of the methods, the values of the hyperparameters are determined by Halving Grid Search (HGS) method that is tuning method. The methods were implemented in Edirne, which is among the province with the highest sunflower yield in Turkey. The results were evaluated with ANOVA method and performance evaluation criteria, MAE, MAPE, RMSE, and R^2. The R^2 values obtained for the test data were determined as 0.92, 0.68 and 0.80 for the DT, SVM and RF methods, respectively. In addition, the number of combinations and execution times were compared using the grid search method and the HGS method for the DT method that gave the best results. While 644204 combinations were solved in 4608 seconds with grid search, 5324 combinations were solved in 23 seconds with HGS. Thus, DT method, providing the prediction with the lowest error, is determined a suitable method for sunflower yield prediction and then accurate buying decision making.

Keywords: Sunflower production; Machine learning; Decision tree; Halving grid search method.

1. Introduction

Agriculture plays an important role in the economic development of countries by increasing food security and social well-being and limiting the impact of climate change (Mok et al. 2014; Byerlee et al. 2009; Palatnik

Citation: Gökler SH (2024). Sunflower Crop Yield Prediction Using Machine Learning Methods. *Selcuk Journal of Agriculture and Food Sciences*, 38(3), 444-462. https://doi.org/10.15316/SJAFS.2024.040

Correspondence: sedahaticegokler@ksu.edu.tr

Received date: 03/04/2024

Accepted date: 03/10/2024

and Roson 2012). Accurate and timely crop yield prediction is extremely valuable for agricultural resource managers and crop producers to ensure food security and sustainability encountered in agricultural production and planning import and export. Food security is one of the critical issues facing many countries. Major fluctuations in crop yield from year to year have serious adverse effects on the balance between supply and demand (Abbott et al. 2011). If the precision of crop yield estimation is improved, the socioeconomic impact of crop loss can be minimized. However, crop yield prediction is extremely challenging due to numerous complex factors (Khaki and Wang, 2019).

As a crop type, the sunflower plant (*Helianthus annuus* L.) is grown in many countries to contribute to the economy. Sunflower is one of the most important annual crops in the world, and it is grown for edible oil (Putt 1977; Ceyhan et al. 2008). It has a high oil content (%36–%55) (Önder et al. 2001; Narin and Abdikan 2022). Sunflower oil ranks first in terms of edible oil quality. In addition, sunflower oil is one of the oils with high nutritional value. Although the sunflower plant is mostly planted to obtain oil, is also used as a snack, bird seed, industrial plant, and ornamental plant.

According to 2018 data, 46% of vegetable oil production in Turkey is met by sunflowers (USDA 2020). In sunflower cultivation, Turkey ranks 6th in world production and has a share of 4.12%. It is expected that the production amount will reach 2,6 million tons in 2023 (URL$_1$, 2023). Global sunflower production is estimated to be 50.7 million tons in the marketing year of 2022–23, with a decrease of 11.6% compared to the previous year. In Turkey, an increase in production is expected, but it is expected that this increase will not be able to cover the total losses. However, the need for edible oil in Turkey increases in parallel with the per capita consumption and population growth, the amount of production cannot meet the entire demand and efficient increases from year to year and has exceeds 500 thousand tons.

Because the increasingly significant oil deficit is met through seed and crude oil imports, making the country dependent on foreign sources for raw materials. Turkey's average proficiency level in the last 20 years has been 57 percent. During this period, 43% of the need was met by imports. Accordingly, total sunflower imports were 3.3 million tons in the 2019/20 season, while exports were only 1.94 million tons (Republic of Turkey Ministry of Agriculture and Forestry 2022). Turkey is one of the larger sunflower importer countries in the world. But, epidemics, natural disasters, and wars in the world and in importing countries cause major disruptions in supply. The yield must be accurately estimated for the imported quantity to be correct.

For sunflower yield prediction, simple methods such as farmers' long-term experience for specific fields, the average of several previous yields, or the last obtained yield can be used. However, these methods have some shortcomings. Nevertheless, crop yield varies from one year to another, with large deviations. The direct method, crop simulation, remote sensing and statistical methods are commonly known crop yield estimation methods. The direct method is based on ground measurements. Although these methods give reliable predictive results, they are not cost and time efficient, and therefore it is very difficult to apply the on large areas (Burke and Lobell 2017). Crop growth simulation models are also used for crop yield estimation, which includes ecophysiological processes to simulate crop growth, development, and yields according to soil characteristics, agricultural practices, and meteorological data (Leroux et al. 2019). In the statistical method, it is assumed knowing how input variables are related to the output. It is wrongly determined the relationship between input and output by user, it may result in inaccurate prediction model. Remote sensing (RS) technology is a method for better productivity and yield estimation because it allows for the evaluation of the widest fields and gives functional preliminary information about the growing crops (Narin and Abdikan 2022). However, most agricultural fields in developing countries are managed by farmers with small production areas. It is also true that such predictions are difficult to achieve in regions that lack extensive observational records.

Recently, machine learning (ML) methods have been used for yield prediction. In these methods, since the complex relationships of the variables can be defined by the learning model, higher accuracy prediction can be made compared to traditional methods (Kayad et al. 2019). The methods can be successfully used to identify factors that increase crop production under different environmental conditions and also to model and predict future yields (Mourtzinis et al. 2021). SVM, RF or ANN are some of the most popular ML methods used for the prediction of crop yield (Debaeke et al. 2023).

There are few studies using ML methods predicting yield in the literature. Gonzalez-Sanches et al. (2014) compared the predictive accuracy of ML methods such as multilayer neural networks, support vector regression, k-nearest neighbor, and linear regression techniques for crop yield prediction. Călin et al. (2022) aimed to predict sunflower and corn yields by using ML method, based on the plating date and region, with limited available data. Amankulova et al. (2023) used three ML-based regression analysis techniques, multiple linear regression (MLR), random forest regression (RFR), SVM, to predict crop yield. In the methods, accepted values were extracted from remote sensing (RS)-derived vegetation indices (VIs) were as explanatory variables while the predicted crop yield was the response variable.

In this article, it is aimed to structured ML models predicting sunflower yield. But the biggest disadvantage of machine learning methods is determining the hyperparameter values that give the best results. Inappropriate or wrong hyperparameter values used can reduce the prediction performance of the method. In this study the HGS method, is hyperparameter tuning method, was used to determine to reduce all possible combinations of hyperparameters in training phase. Thus, the model is trained on a small subset of data rather than the entire training data.

In the application, 5 input variables were determined as cultivated area, average humidity, average temperature, total sunshine time, and average precipitation, and production amount (ton) was determined as the output variable. 58 years of data for all variables were collected, and a 6x58 dataset was created. The created dataset was divided into 6x41 training dataset and 6x17 testing dataset.

Developed models of SVR, RF and DT methods applied to Edirne province. The prediction accuracy of ML methods was evaluated with ANOVA method and performance criteria such as mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE), and determination coefficient (R^2).

The following sections of this paper are organized as follows. Section 2 elaborates the methods and materials used in this study. In section 3, model's implementation is described. Results are presented in section 4. Section 5 gives a discussion on the results obtained.

2. Materials and Methods

2.1. Materials

Data for the years 1960–2021 were used as the basis. Sunflower cultivation area size and yield amount data were obtained from the Turkish Statistical Institute (TUIK) (URL2), and climate data were obtained from the General Directorate of Meteorology (URL3).

In ML models, cultivated area, average humidity, average temperature, total sunshine time, and average precipitation were used as input variables, and production amount was used as the output variable. In practice, firstly, the missing values in the input and output variables were removed and the normalized values of the remaining variables were calculated. "Min-Max Normalization" method was used to normalize the variables (Eq. 1).

$$X^{\iota} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where X^{ι} is the normalized data, X is the actual data, X_{min} is the minimum value of dataset, and X_{max} is the maximum value of dataset.

For these variables 58-year of data for Edirne province were provided. The 58-year data set was prepared divided into two groups, 70% of which was training (58×0.7=40.6 ≈ 41 years) and 30% was testing (58×0.3=17.4 ≈ 17 years). The first 3 and last 3 rows of the dataset are given due to space constraints (Table 1).

Table 1. Data set

Year	Production amount (ton)	Cultivated area (hectares)	Average humidity (%)	Average temperature (°C)	Total sunshine duration (hours/day)	Average precipitation (mm)
1960	21198	25080	65.22	18.77	261.30	64.05
1961	15582	20440	61.72	19.82	295.37	40.32
1962	8613	14600	57.90	20.50	316.17	33.32
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
2019	249569	95050	60.75	21.78	296.65	40.60
2020	240434	90916	61.15	21.83	259.62	39.37
2021	285286	107351	64.12	21.43	254.83	35.50

Descriptive statistic values of input and output variables are given in Table 2. As seen in Table 2, the descriptive statistics values showing the changes in input variables such as standard deviation, variance, and range are quite different. A similar situation is also valid for the output variable, production amount.

Table 2. Descriptive statistics of input and output variables

Input/Output Variables	Mean	Standard Deviation	Variance	Median	Minimum	Maximum	Range
Production amount (ton)	156330	75903.8	5761384658	162623	8613	332894	324281
Cultivated area (hectares)	97958.3	34967.0	1222692893	107658	14600	142665	128065
Average humidity (%)	61.6124	2.97277	8.83737	61.3083	53.2000	69.5333	16.3333
Average temperature (°C)	20.3480	1.00153	1.00306	20.1250	18.7667	22.9000	4.13333
Total sunshine duration (hours/day)	273.720	18.8482	355.256	271.366	232.464	316.167	83.7025
Average precipitation (mm)	40.3052	12.2119	149.130	37.9833	17.5167	73.6667	56.1500

2.2. Machine Learning Methods

In the literature, SVM, DT, and RF methods are used for modeling and performance evaluation in estimation problems. SVM finds the most appropriate hyperplane to classify data and generally works effectively with high-dimensional data. While DT uses a series of simple decision rules to classify data, RF increases the generalization ability by creating a collection of these trees. Hyperparameter optimization aims to find the best parameter values in the structure or operation of the method to increase the performance of the model. The halving grid search method allows training to be done on a smaller data set instead of the data set to be examined and to obtain more efficient results in a shorter time with fewer operations. Finally, statistical performance criteria such as RMSE and MAE are used to measure the success of the model, evaluating the efficiency and reliability of the models. Each of these methods plays an important role in solving estimation problems and helps to obtain the best results by complementing each other.

2.2.1. Decision Tree Method

The DT has a hierarchical tree structure consisting of nodes called root node branches, decision nodes, and leaf nodes (Rokach and Maimon, 2014). DT is a tree-structured classifier, where decision nodes, branches, and each leaf represent the features of a dataset, the decision rules, and the outcome, respectively. In a DT, for predicting the given dataset, the algorithm starts from the root node of the tree and passes through the decision node to the last leaf node, thus creating a branch (Figure 2).

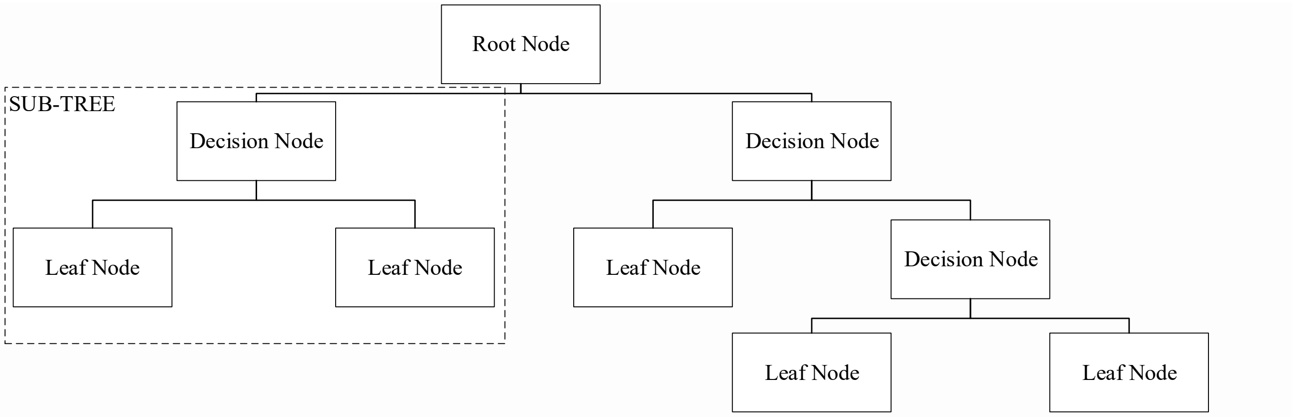


Figure 2. DT Structure

In the tree, classes are represented by leaves, and only one road goes to each leaf. While creating the tree, yes-or-no questions are asked. In this algorithm it is compared the values of the root attribute to the record (real dataset) attribute and, according to the comparison result, followed the branch and jumped to the next node. In the next node, it compares the attribute value to the other sub-nodes again and moves further. It continues the process until it reaches the leaf node of the tree.

2.2.2. *Support Vector Machine Method*

SVM is an ML algorithm used for linear or nonlinear classification and regression. SVM is an ML algorithm used for linear or nonlinear classification and regression. SVM is an easy, adaptable, and efficient method because it can manage high-dimensional data and nonlinear relationships (Cortes and Vapnik 1995). SVM regression is a regression method maintaining all the main features that characterize the SVM algorithm (maximal margin). SVM regression is a powerful tool to explain complex relationships between the input variables and the target variable. The method aims to find the hyperplane that passes through as many data points as possible within a certain distance, called the margin, instead of fitting a line to the data points (Figure 3). SVM handles non-linear relationships between input variables and the target variable using a kernel function and thus it reduces the prediction error.

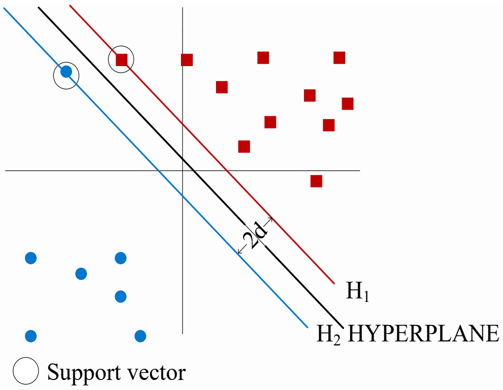


Figure 3. Support Vector and Hyperplane Structure

2.2.3. *Random Forest Method*

RF used to predict continuous outcomes, is one of the popular ML methods (Breiman 2001). RF is an ensemble technique with the use of multiple DTs. In other words, it is RF containing many trees constructed in a "random" way form. It is called Bootstrap and Aggregation because of it is combined multiple DTs in determining the final output instead of evaluating individual DTs. Each tree is formed from a different sample of rows and at each node of tree it is selected a different sample of features for splitting. Trees make individual predictions. These predictions are then averaged to produce a single result (Figure 4). RFs generally outperform DTs. However, RF method's performance can be affected by data characteristics.

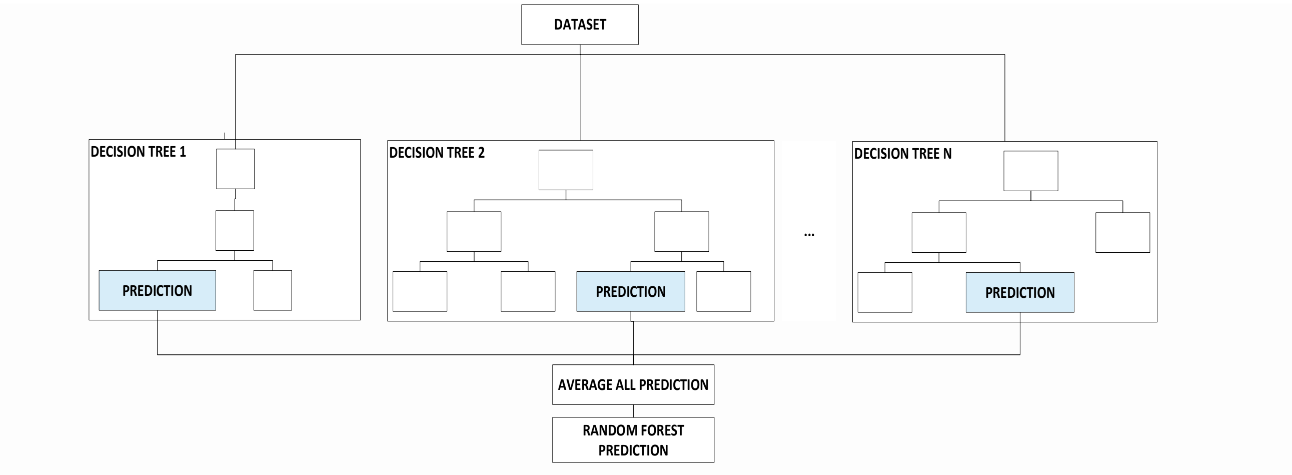


Figure 4. RF Structure

2.2.4. Hyperparameter Tuning and Halving Grid Search

ML methods contain a large number of hyperparameters that affect their performance. These hyperparameters have a wide range of values when it is determined by expert experience or by examining different studies in the literature. Therefore, the hyperparameter value may result in poor performance. This is because each ML method has its own best set of hyperparameter values that can vary according to different or updated input data (Salam and El Hibaoui 2021; Xu et al. 2021). Therefore, it is necessary to use hyperparameter tuning approaches that enable even inexperienced users to achieve good performance. In the literature, the grid search (GS) method is a frequently used method for hyperparameter tuning due to its ease of use (Hadjout et al. 2022; Aouad et al. 2022). The GS method aims to run ML methods for all combinations of hyperparameter values within the value range defined by the user and to obtain the combination that gives the best results. As the number of hyperparameters increases, the number of combinations to be processed increases, which is time-consuming and has high computational costs. In order to eliminate this disadvantage of the GS method, HGS method is used. In the method, all possible combinations of hyperparameters are trained on a small subset of data rather than the entire training data, fewer transactions are carried out in less time.

2.2.5. Performance Evaluation Criteria

The MAE, MAPE, RMSE, and R^2 are the most commonly used criteria when evaluating model performance (Chung et al. 2022; Cui 2022; Tang et al. 2024; Yan et al. 2024). In prediction problems, the prediction error between the predicted value and the actual value must be minimal. Therefore, statistical performance criteria such as MAE, MAPE, and RMSE were used. In addition, statistical performance criteria are used to determine prediction accuracy. Therefore, the R^2 value was calculated. R^2 is the coefficient of determination that provides information about the goodness of fit of a model and takes values between 0 and 1. While an R^2 value closer to 0 indicates that the prediction is incompatible with the data, and the value of 1 indicates that the regression predictor fits the data perfectly.

The performance evaluation criteria used are given in Eqs. 2-5, respectively.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|T_i - P_i| \quad (2)$$

$$MAPE \quad = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{T_i - \quad P_i}{T_i}\right| \tag{3}$$

$$RMSE \quad = \sqrt{1/N\sum_{i=1}^{N}(T_i - P_i)^2} \tag{4}$$

$$R^2 \quad = \left[\frac{\sum_{i=1}^{N} \; ((T_i - \overline{T})\,(P_i - \overline{P})\,)}{\sqrt{\sum_{i=1}^{N} \; (T_i - \overline{T})^2 \; \sum_{i=1}^{N} \; (P_i - \overline{P})^2}}\right]^2 \tag{5}$$

where T_i is i^th^ actual data, $\overline{T}$ is the mean of actual data, P_i is the i^th^ predicted output data and $\overline{P}$ is the mean of i^th^ predicted output data in the dataset.

2.3. Study Area

Edirne province is one of the most province in sunflower cultivation in Turkey. In 2021, 285,286 tons of sunflower production were realized from 1,073,508 decares of cultivated area in Edirne province. Edirne province's latitude and longitude are 26 E 34 and 41 N 40, respectively (Figure 5).

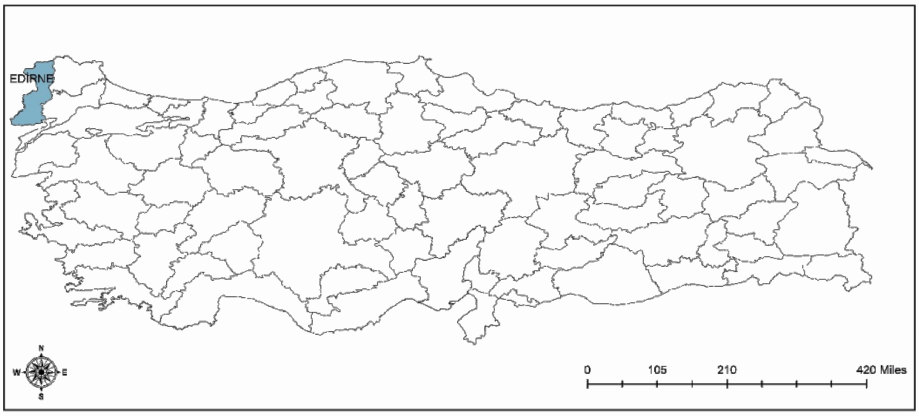


Figure 5. Edirne province in Turkey

Edirne is a transition region under the influence of both the Mediterranean climate and the continental climate peculiar to Central Europe. It has suitable conditions in terms of humidity, temperature, rain, and duration of sunshine. The annual average temperature is 13.4 °C, precipitation is 585.9 mm, and relative humidity is 70%.

3. Implementation

3.1. Definition of Model

In this study, DT (Iniyan et al. 2023; Singh and Singh 2017; Kalichkin et al. 2021), SVM (Priyadharshini et al. 2022; Gandhi et al. 2016; Gonzalez-Sanchez et al. 2014), and RF (Everingham et al. 2016; Fukuda et al. 2013), which are the most used ML methods for crop yield prediction in the literature (Debaeke et al. 2023), were selected to predict annual yield in Edirne. Furthermore, Benos et al. (2021) examined the studies in which ML methods were used for yield prediction in the literature and determined that the ones that gave the best output were ANN, SVM and DT, respectively.

As a result of the examination of the related studies in the literature, five variables that are thought to affect the sunflower yield were determined. They are cultivated area (Gonzalez-Sanches et al. 2014.; Gandhi et al. 2016), average humidity (Laxmi and Kumar 2011; Dahikar and Rode 2014), average temperature (Jiang et al. 2004; Kaul et al. 2005; Thonhboonnak et al. 2011; Laxmi and Kumar 2011; Dahikar and Rode 2014), total sunshine duration (Jiang et al. 2004; Thonhboonnak et al. 2011), and average precipitation (Kaul et al. 2005; Liu et al. 2001; Laxmi and Kumar 2011; Thonhboonnak et al. 2011; Dahikar and Rode 2014). While the cultivated area and yield of sunflowers are considered on an annual basis, the climate elements (temperature, humidity, precipitation, sunshine duration) are taken as the average of the 6 months between April and September, which is the production period of sunflower (Mishra et al. 2016; Jain et al. 2017; Paudel et al. 2021). Yield is expressed in tons of crop grown per hectare or decare in arable regions. The cultivated area is agricultural lands where the sunflower plant is grown for one year. The average humidity is the average concentration of water vapor present in the air. The average temperature of the air as indicated by a properly exposed thermometer during a given time period, such as a day, a month, or a year. The average sunshine duration is the average length of time that the ground surface is irradiated by direct solar radiation. The average precipitation is the average amount of weather events such as rain and snow that occur as a result of the condensation of water vapor in the atmosphere.

3.2. Implementation of HGS Method

The hyperparameter combinations determined using HGS for ML methods in the study are shown in Table 3.

Table 3. Hyperparameters of methods

	SVM	RF	DT
Hyperparameters	Gamma parameter	Number of trees	Confidence parameter
	C parameter	Maximum depth	Minimum number of leaves
	Max number of iterations	Prepruning	Prepruning
	Kernel cache		Pruning
	Kernel type		Maximum depth

All hyperparameter combinations created in accordance with the given lower and upper limits were solved by the grid search method, and since all hyperparameter combinations were tried, the hyperparameter combination with the best performance value was obtained.

In the SVM method, it has been determined by the HGS method that there are five hyperparameters that affect the prediction result. For the determined hyperparameters, 58564 combinations were created in the grid search method, and the minimum and maximum values used are summarized in Table 4. Also, "dot", "radial", "neural" and "anova" are used as kernel type.

Table 4. Hyperparameters values for the SVM method

Hyperparameters	Min	Max	Number of steps
Gamma parameter	0	100	10
C parameter	0	100	10
Max number of iterations	1	100	10
Kernel cache	0	100	10

The hyperparameters of SVM with the best RMSE value are radial kernel type, 80 kernel cache, 80 max iterations, 20 C, and 30 Gamma. Test data values of predicted and actual production amounts using the SVM method are shown in Figure 6.

A total of 58 support vectors were used in the SVM method and a bias of 0.457 was obtained. The weights calculated with the SVM method are seen in Table 5.

Table 5. Input variable weights obtained with SVM

Cultivated area (hectares)	Average humidity (%)	Average temperature (°C)	Total sunshine duration (hours/day)	Average precipitation (mm)
1.928	0.661	1.281	0.239	0.271

Since the values in the data set are numerical in the RF method, the 'least-square' separation criterion was used. By using the HGS method, it was determined that three important parameters affect the prediction performance. Therefore, the grid search method was applied for these three hyperparameters (Table 6).

Table 6. Hyperparameters and values used for the RF method

Hyperparameter	Min	Max	Number of steps
Number of trees	1	100	100
Maximum depth	0	100	100
Prepruning	√	X	

Applied: √ Not applied: χ

After determining the hyperparameters suitable for the RF model, prediction values were calculated with the proposed algorithm. The hyperparameters of the RF with the best RMSE values are the combination of 3 tree count, 55 maximum depth, and no preprunning.

Test data values of predicted and actual production amounts using the RF method are shown in Figure 7.

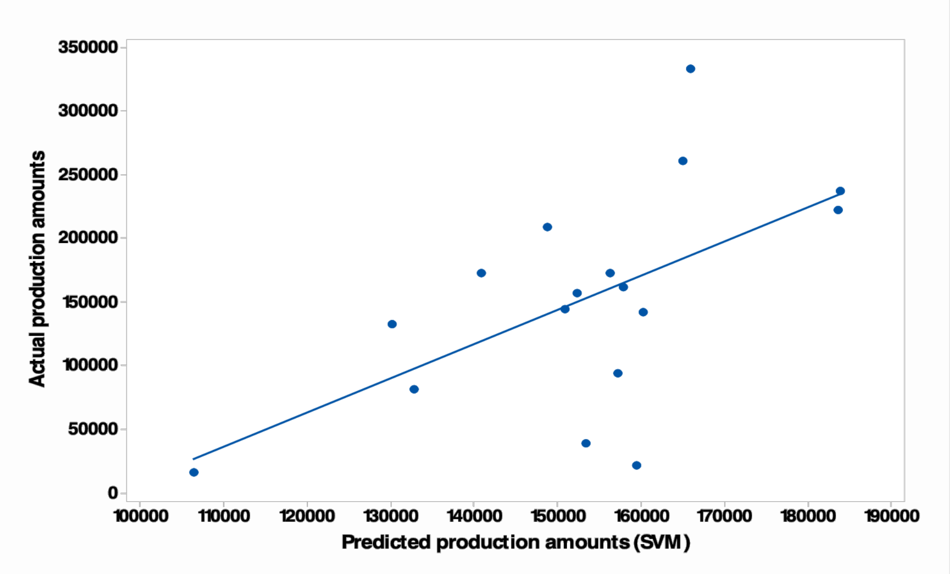


Figure 6. Actual and predicted production amounts using SVM method

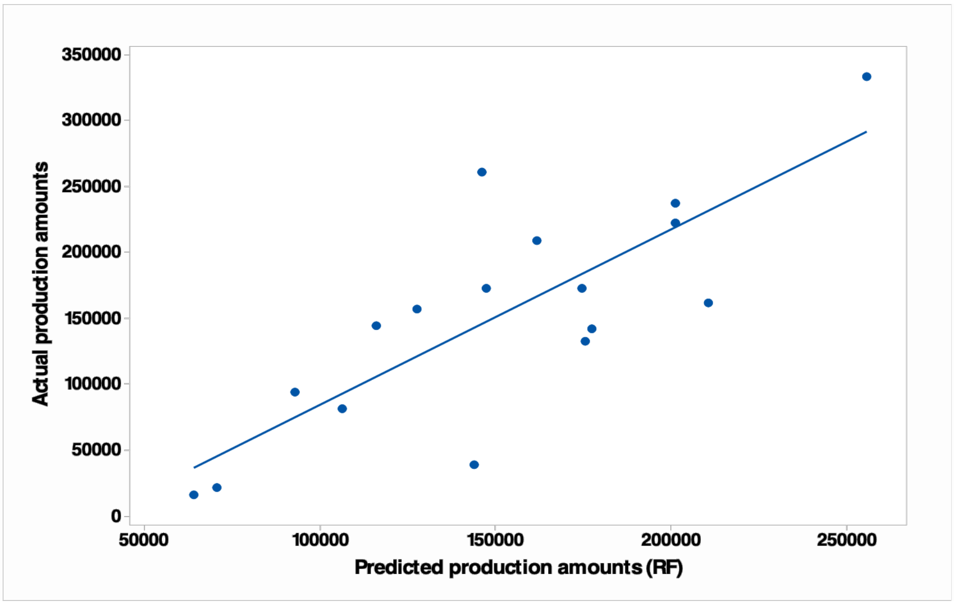
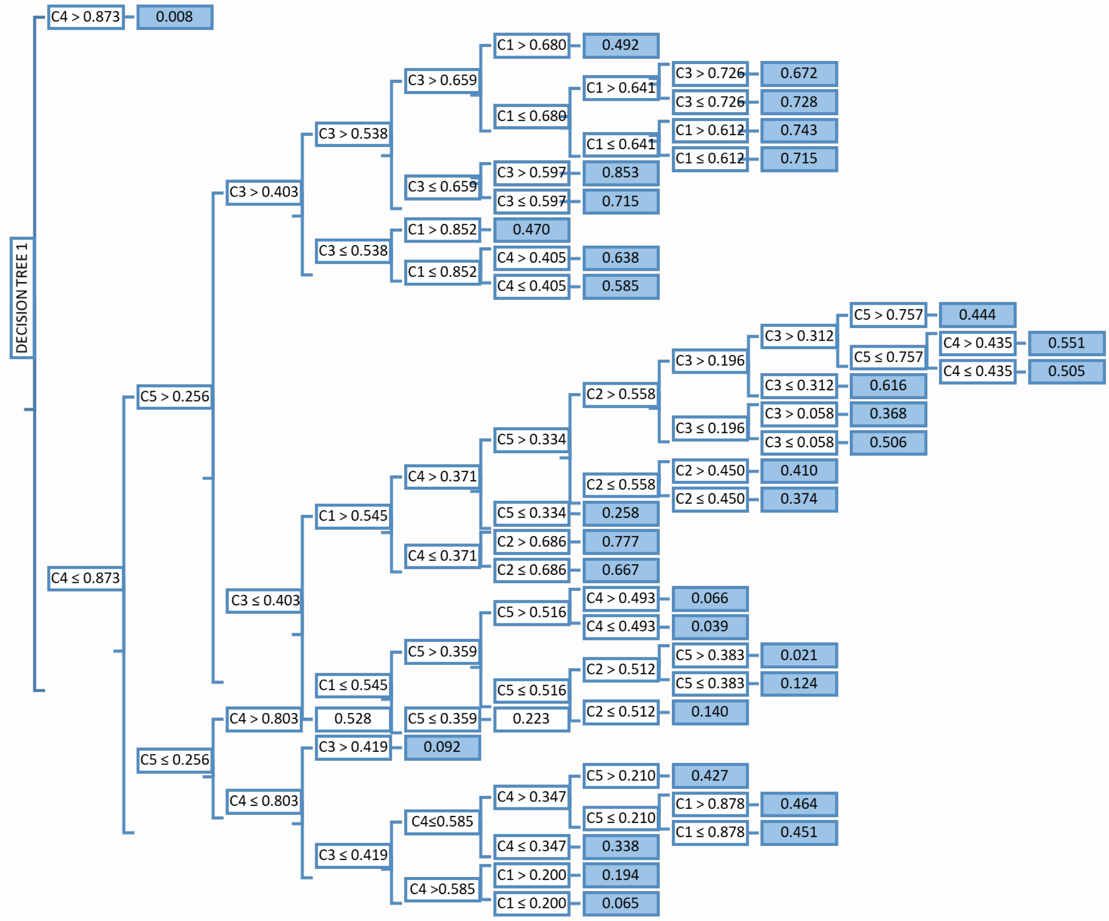


Figure 7. Actual and predicted production amounts using RF method



C1: Cultivated area C2: Average humidity C3: Average temperature C4: Total sunshine duration C5: Average precipitation

Figure 8. The first created DT for the RF method

Due to space constraints, one of the three created DTs for the RF method is shown (Figure 8). Final prediction values were obtained by taking the arithmetic average of the values obtained using three DTs.

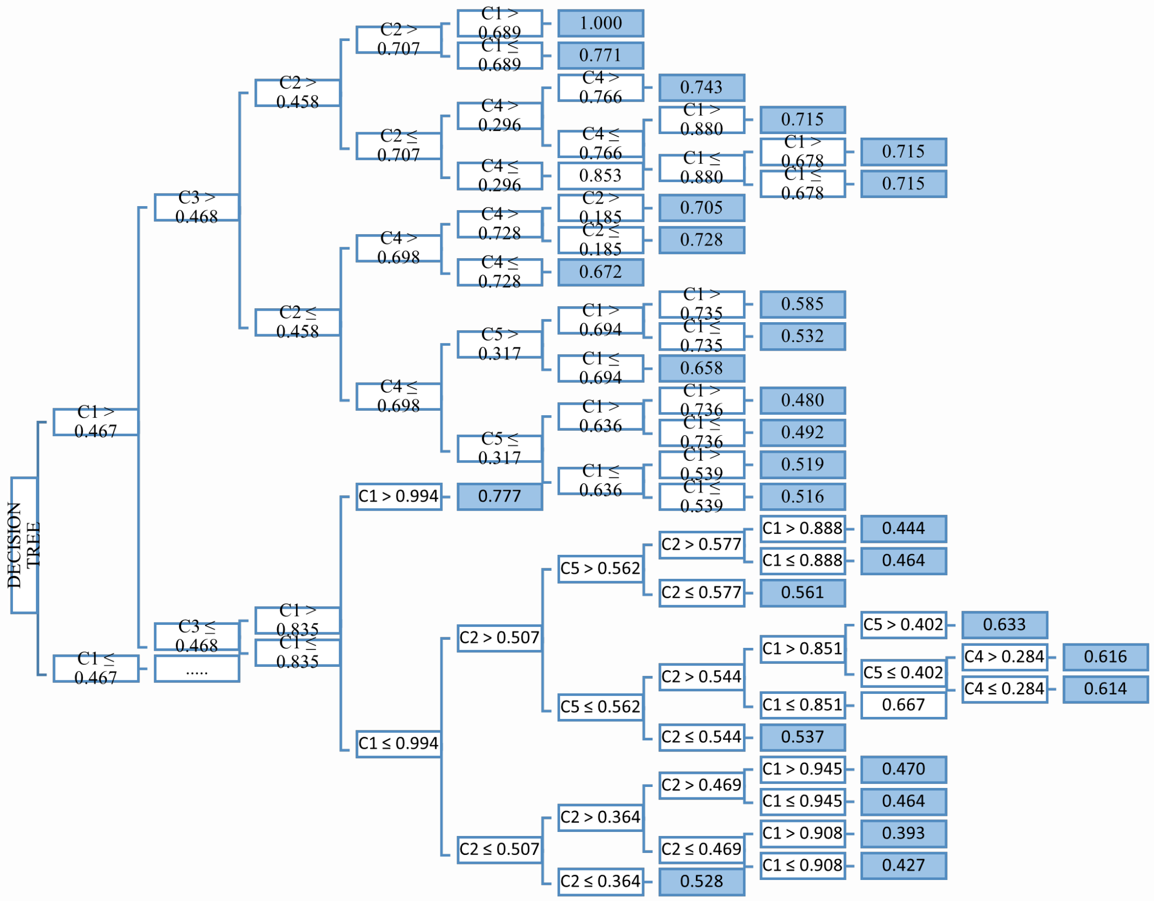
In the DT method, least squares method was used as the splitting criterion. Hyperparameters used are also shown in Table 7.

Table 7. Hyperparameters and values used for the DT method

Hyperparameter	Min	Max	Number of steps
Confidence parameter	0	1	10
Minimum number of leaves	1	100	10
Prepruning	√	χ	
Pruning	√	χ	
Maximum depth	0	100	10

Applied: √ Not applied: χ

In the method, 5324 combinations were created for 5 hyperparameters, and RMSE values were obtained. The iteration that gives the best prediction result has 20 maximum depth, 0.8 confidence parameter, 51 minimum number of leaves, no pruning and no prepruning hyperparameters. Test data values of predicted and actual production amounts using the DT method are shown in Figure 9. Additionally, the created DT is shown in Figure 10.



C1: Cultivated area C2: Average humidity C3: Average temperature C4: Total sunshine duration C5: Average precipitationa

Figure 10. The created DT

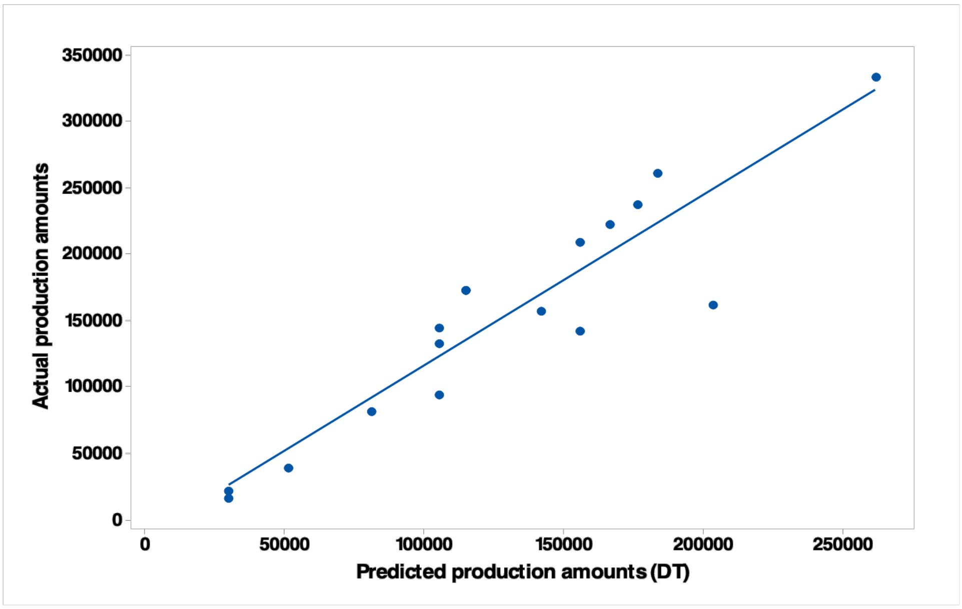


Figure 9. Actual and predicted production amounts using DT

In Figure 11, the closest predicted values to the actual production amount are determined by the DT method with hyperparameters determined with HGS.

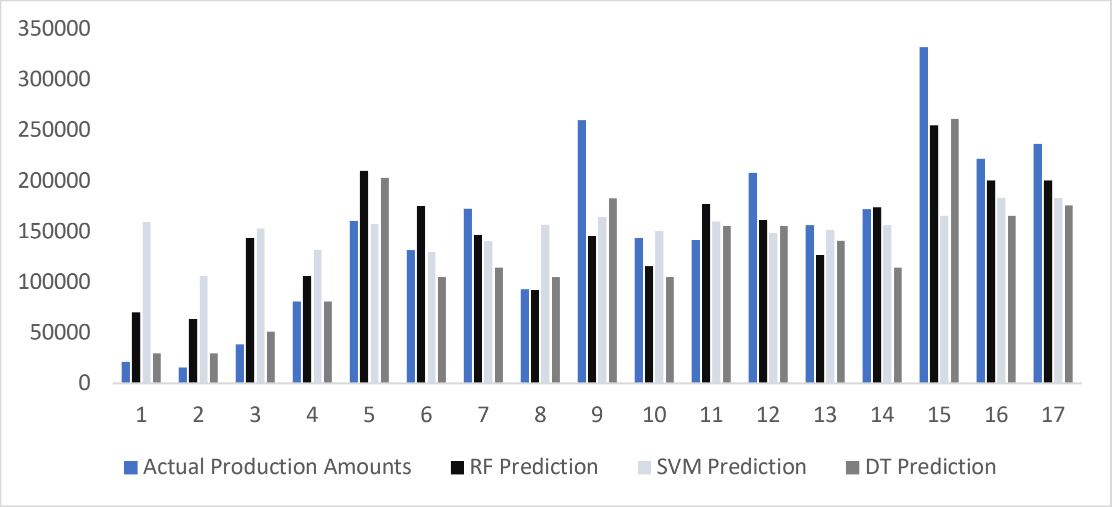


Figure 11. Actual and predicted production amounts using ML methods

4. Results

4.1. Results for comparison of the machine learning methods

The prediction result obtained using the DT method and the actual sunflower yield are 95 % similar. At the end of the application, the DT method was compared with other ML methods such as SVM, and RF. Rapidminer Studio program was used for all ML applications. In Table 8, the performance evaluation criteria for the training and testing data of the prediction methods are separately shown.

When the MAE values are examined, the DT model provides the lowest error in both training data (0.112) and test data (36352.03). While the RF model has higher error rates, with 0.138 in training data and 43436.89 in test data, the SVM model shows the highest error, with 0.268 in training and 58919.2 in test data.

Similarly, when the MAPE statistical performance criterion is evaluated, DT provides the best results in both training (0.263) and testing (0.274). The highest error is obtained in the SVM method.

When examined in terms of RMSE, DT again has the lowest error values, with 0.164 in training and 43618.12 in test. In test data, RF (52865.08) and SVM (65986.54) methods perform weaker than DT.

R^2 values show the success of the model. DT shows the best fit with 0.954 in training and 0.920 in testing. RF shows a strong performance in training (0.939) but slightly drops in test data (0.800). SVM's R^2 values in training (0.720) and test (0.680) data show lower performance compared to other models.

Finally, when training time is taken into account, DT is the fastest model, training in just 23 seconds. RF and SVM models were trained in longer times, 945 and 884 seconds, respectively.

Table 8. Comparison of the machine learning methods

Performance Criteria	Methods					
	DT		RF		SVM	
	Train	Test	Train	Test	Train	Test
MAE	0.112	36352.03	0.138	43436.89	0.268	58919.2
MAPE	0.263	0.274000	0.378	0.651	0.554	1.10000
RMSE	0.164	43618.12	0.179	52865.08	0.242	65986.54
R^2	0.954	0.920	0.939	0.800	0.720	0.680
Training Time (second)*	23		945		884	

* These are the total training times used to solve selected optimum hyperparameter combinations with HGS

In order to see the advantage of the HGS method, the DT method is solved with max depth (P_1), pruning (P_2), confidence parameter (P_3), pre-pruning (P_4), minimum number of leaves (P_5), minimum division (P_6), and pre-pruning (P_7) parameters without using HGS. Table 9 includes the criteria used, the number of combinations, performance criteria, and solution times.

Table 9. Solution of DT method with different hyperparameters

Parameter							Number of parameter	Number of combinations	RMSE	AE	RRSE	R^2	Time (second)
P_1	P_2	P_3	P_4	P_5	P_6	P_7							
√	√	√	√	√	√	√	7 parameter	644204	0.078	0.063	0.332	0.89	4608
	√	√	√	√	√	√	6 parameter	58564	0.091	0.074	0.434	0.84	432
√		√	√	√	√	√	6 parameter	322102	0.086	0.066	0.336	0.88	1759
√	√		√	√	√	√	6 parameter	58564	0.091	0.074	0.434	0.84	223
√	√	√		√	√	√	6 parameter	322102	0.086	0.660	0.366	0.88	3056
√	√	√	√		√	√	6 parameter	58564	0.054	0.042	0.213	0.95	433
√	√	√	√	√		√	6 parameter	58564	0.054	0.042	0.213	0.95	894
√	√	√	√	√	√		6 parameter	58564	0.054	0.042	0.213	0.95	894
√	√	√	√	√			5 parameter *	5324	0.054	0.042	0.213	0.95	23

* Hyperparameters determined by HGS method

The RMSE, AE, and RRSE error values obtained with 644204 combinations using 7 parameters are 0.078, 0.063, and 0.332, respectively. Although the model also provides reasonable results in terms of R^2 value (0.89), it was the sample trained in the longest time with a training time of 4608 seconds.

Among the other combinations made with 6 parameters, the sample without the P_5 parameter draws attention. This sample has the best results in terms of RMSE (0.054), AE (0.042), RRSE (0.213), and R^2 (0.95), and its training time is 433 seconds.

Finally, the sample determined by the HGS method with 5 parameters gave the lowest RMSE (0.054), AE (0.042), RRSE (0.213), and R^2 (0.95) results, while it was the fastest model with a training time of only 23 seconds. This combination offers the most ideal solution in terms of both accuracy and speed. In summary, while increasing the number of parameters increases the complexity and training time of the model, the combinations with the best performance were 5 and 6 parameter models. In particular, the model determined with the HGS method stands out with the lowest error rate and the shortest training time.

At the end of the solutions obtained, it is seen that the performance criteria of the HGS method get better values, and the solution time is shortened.

4.2. Statistical method

In the study, MINITAB 18 program was used for statistical method solutions.

An ANOVA test was applied using the absolute values of the difference between the predicted and actual production amounts. The fact that the p value is less than 0.05 as a result of the test shows that the accuracy results between DT, RF, and SVM are statistically significant (Table 10). This low p-value indicates that the factors have an effect on the model, and the null hypothesis (that the factors have no effect) can be rejected. Accordingly, the p-value is less than 5 percent, and the H_0 hypothesis is rejected. In other words, there is a difference between the groups.

Table 10. Result of ANOVA

Source	DF	Adj SS	Adj MS	F	p
Factor	2	1.71848E+11	57282528009	9.00	0.000
Error	64	4.07500E+11	6367192115		
Total	66	5.79348E+11			

In addition, standard deviations of the absolute value of the difference between the predicted and actual yield for each method were calculated. It was seen that the DT method gave the smallest standard deviation (Table 11).

Table 11. Mean and standard deviation of method

Factor	N	Mean	Standard Deviation
RF	17	43437	31060
SVM	17	149569	110780
DT	17	36352	24847

The standard deviation value for RF is 31060, indicating that there is a certain degree of deviation in RF's prediction.

The standard deviation value of the SVM method is also quite high, indicating that SVM predictions show a large variability. SVM performs the worst in terms of both mean and deviation.

The DT method, on the other hand, has the lowest value in terms of both mean and standard deviation. This shows that DT gives more consistent and reliable results compared to other methods.

As a result, the DT method shows the best performance with both a low mean and a low standard deviation. Although RF exhibits slightly higher error rates compared to DT, it still shows reasonable performance. SVM stands out as the method that gives the worst results in terms of both mean and standard deviation.

5. Conclusion

The production of sunflower, one of the most important crops, varies due to fluctuations that occur every year. In the study, sunflower production was predicted with ML methods, that are frequently used in environments where such variability is high. The biggest challenge in machine learning methods is determining the hyperparameter combinations and values that affect prediction performance. For this reason, the combination of hyperparameters to be used in the models was determined using the HGS method. At the end of the application, the DT hyperparameter combination that gave good results was determined as the maximum depth, pruning, confidence parameter, pre-pruning, and minimum number of leaves. In addition, by using HGS in DT, results were obtained with 95% accuracy level in a short time of 23 seconds. RF and SVM methods, with 945 and 884 seconds, respectively, have been less successful than DT. Finally, with the ANOVA test, it was verified that ML methods used have different prediction capabilities.

The best prediction result was obtained with DT method based on the HGS method. Thus, with the DT model, the production amount for the coming years can be estimated for all sunflower-growing provinces according to the changing amounts of inputs. The amount to be imported, which cannot be met from the total predicted product amount, can be determined. In this way, the problems of keeping stock in case the imported quantity is larger than necessary and not being able to meet the demand if it is less will be prevented.

6. Discussions

Machine learning methods are used to obtain accurate results in prediction problems. Each of the machine learning methods has advantages and disadvantages. For example, SVM provides high accuracy, but it has high computational cost and kernel selection difficulties in large data sets. DT, although highly interpretable, has the risk of overfitting and may be inadequate in modeling complex data. RF increases the generalization ability of the model but may bring large computational and memory costs.

One of the biggest disadvantages of machine learning methods is the need to determine many hyperparameters, both structurally and operationally. This study aims to determine the hyperparameters of machine learning methods using HGS in a shorter time and with less workload. In practice, the HGS method has been applied to machine learning methods such as DT, RF, and SVM. The method that provides the best performance value criteria is DT. Additionally, by performing an ANOVA test, it was confirmed that the method with the lowest standard deviation was the DT method.

In future studies, it is aimed to determine the hyperparameters of the machine learning methods used by using different optimization methods. In addition, it is aimed to examine the accuracy of the model by using the developed model in different regions where sunflower production is carried out.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

Abbott P, Hurt C, Tyner E (2011). What's driving food prices in 2011. Farm Foundation. Oak Brook, IL, USA.

Amankulova K, Farmonov N, Mukhtorov U, Mucsi L (2023). Sunflower crop yield prediction by advanced statistical modeling using satellite-derived vegetation indices and crop phenology. *Geocarto International* 38(1):2197509.

Aouad M, Hajj H, Shaban K, Jabr RA, El-Hajj W (2022). A CNN-Sequence-to-sequence network with attention for residential short-term load forecasting. *Electr. Power Syst. Res* 211:108152.

Benos L, Tagarakis AC, Dolias G, Berruto R, Kateris D, Bochtis D (2021). Machine learning in agriculture: A comprehensive updated review. *Sensors* 21(11): 3758. https://doi.org/10.3390/s21113758.

Breiman L (2001). Random Forests. *Machine Learning* 45: 5–32.

Burke M, Lobel D (2017). Satellite-based assessment of yield and its determinants in smallholder african systems. *Proceedings of the National Academy of Sciences* 114(9):2189-2194.

Byerlee D, de Janvy A, Sadoulet E (2009). Agriculture for development: toward a new paradigm. *Annual Review of Resource Economics* 1:15-31.

Călin AD, Mureşan H-B, Coroiu AM (2022). Feasibility of using machine learning algorithms for yield prediction of corn and sunflower crops based on seeding date. *Studia Univ. Babes–Bolyai, Informatica* LXVII(2) https://doi.org/10.24193/subbi.2022.2.02.

Ceyhan E, Önder M, Öztürk Ö, Harmankaya M, Hamurcu M, Gezgin S (2008). Effects of application boron on yields, yield component and oil content of sunflower in boron-deficient calcareous soils. *African Journal of Biotechnology* 7(16): 2854-2861.

Chung, W. H., Gu, Y. H., & Yoo, S. J. (2022). District heater load forecasting based on machine learning and parallel CNN-LSTM attention. *Energy* 246, 123350.

Cortes C, Vapnik V (1995). Support vector networks. *Machine Learning* 20:273-297.

Cui, M. (2022). District heating load prediction algorithm based on bidirectional long short-term memory network model. *Energy* 254: 124283.

Dahikar SS, Rode S (2014). Agricultural crop yield prediction using artificial neural network approach. *International Journal of Innovative Research In Electrical, Electronics, Instrumentation And Control Engineering* 2(1): 683-686

Debaeke P, Attia F, Champolivier L, Dejoux J-F, Micheneau A, Al Bitar A, Tr´epos R (2023). Forecasting sunflower grain yield using remote sensing data and statistical models. *European Journal of Agronomy* 142 (2023): 126677.

Everingham Y, Sexton J, Skocaj D, Inman-Bamber G (2016). Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development* 36(2): 27.

Fukuda S, Spreer W, Yasunaga E, Yuge K, Sardsud V, Müller J (2013). Random forests modelling for the estimation of mango (*Mangifera indica* L. cv. Chok Anan) fruit yields under different irrigation regimes. *Agric Water Manag* 116: 142-150. https://doi.org/10.1016/j.agwat.2012.07.003

Gandhi N, Armstrong LJ, Petkar O, Tripathy AK (2016). Rice crop yield prediction in India using support vector machines. *IEEE Xplorer, 13th International Joint Conference on Computer Science and Software Engineering (JCSSE).*

Gonzalez-Sanchez A, Frausto-Solis J, Ojeda-Bustamante W (2014). Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research* 12(2): 313-328.

Hadjout D, Torres J, Troncoso A, Sebaa A, Martínez-Álvarez F (2022). Electricity consumption forecasting based on ensemble deep learning with application to the Algerian market. *Energy* 243: 123060.

Iniyan S, Varmaa VA, Naidu CT (2023). Crop yield prediction using machine learning techniques. *Advances in Engineering Software* 175: 103326.

Jain N, Kumar A, Garud S, Pradhan V, Kulkarni P (2017). Crop selection method based on various environmental factors using machine learning. *International Research Journal of Engineering and Technology* 4(2): 56-72.

Jiang D, Yang X, Clinton N, Wang N (2004). An artificial neural network model for estimating crop yield using remotely sensed information. *Int. J. Remote Sensing* 25: 1723-1732.

Kalichkin VK, Alsova OK, Maksimovich KY (2021). Application of the decision tree method for predicting the yield of spring wheat. *AGRITECH-V-2021 IOP Conf. Series: Earth and Environmental Science* 839: 032042 https://doi.org/10.1088/1755-1315/839/3/032042

Kaul M, Hill RL, Walthall C (2005). Artificial neural networks for corn and soybean yield prediction. *Agric. Syst* 85: 1–18.

Kayad A, Sozzi M, Gatto S, Marinello F, Pirotti F (2019). Monitoring within-field variability of corn yield using Sentinel-2 and machine learning techniques. *Remote Sensing* 11(23): 2873.

Khaki S, Wang L (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science* 10: 452963.

Laxmi RR, Kumar A (2011). Weather based forecasting model for crops yield using neural network approach. *Statistics and Applications* 9(1), 55-69.

Leroux L, Castets M, Baron C, Escorihuela MJ, Bégué A, Seen DL (2019). Maize yield estimation in West Africa from crop process-induced combinations of multi-domain remote sensing indices. *European Journal of Agronomy* 108: 11-26.

Liu J, Goering CE, Tian L (2001). Neural network for setting target corn yields. *T ASAE* 44(3): 705-713.

Mishra S, Mishra D, Santra G (2016). Applications of machine learning techniques in agricultural crop production: A review paper. *Indian Journal of Science and Technology* 9(38).

Mok H-F, Dassanayake KB, Hepworth G, Hamilton AJ (2014). Field comparison and crop production modeling of sweet corn and silage maize (*Zea mays* L.) with treated urban wastewater and freshwater. *Irrigation Science* 32(5): 351–368. https://doi.org/10.1007/s00271-014-0434-4.

Mourtzinis S, Esker PD, Specht JE, Conley SP (2021). Advancing agricultural research using machine learning algorithms. *Scientific Reports* 11(1): 1–7.

Narin OG, Abdikan S (2022). Monitoring of phenological stage and yield estimation of sunflower plant using Sentinel-2 satellite images, *Geocarto International* 37(5): 1378-1392, https://doi.org/10.1080/10106049.2020.1765886.

Önder M, Öztürk Ö, Ceyhan E (2001). Yağlık ayçiçeği çeşitlerinin verim ve bazı verim unsurlarının belirlenmesi. *S.Ü. Ziraat Fakültesi Dergisi* 15 (28): 136-146.

Palatnik RR, Roson R (2012). Climate change and agriculture in computable general equilibrium models: alternative modeling strategies and data needs. *Climatic Change* 112: 1085–1100.

Paudel D, Boogaard H, de Wit A, Janssen S, Osinga S, Pylianidis C, Athanasiadis I (2021). Machine learning for large-scale crop yield forecasting. *Agricultural Systems* 187(1).

Priyadharshini K, Prabavathi R, Devi VB, Subha P, Saranya SM, Kiruthika K (2022). An enhanced approach for crop yield prediction system using linear support vector machine model. *In 2022 IEEE International Conference on Communication, Computing, and Internet of Things (IC3IoT)* :1-5.

Putt ED (1977). Early history of sunflowers. In: A.A. Schneiter (ed). Sunflower Technology and Production. ASACSSA and SSSA Madison, WI. p. 1-19.

Republic of Turkey Ministry of Agriculture and Forestry, Sunflower Bulletin, 20 May, 2022

Rokach L, Maimon O (2014). Data mining with decision tree; series in machine perception and artificial intelligence. *World Scientific* 81: 61-62.

Salam A, El Hibaoui A. (2021) Energy consumption prediction model with deep inception residual network inspiration and LSTM. *Math. Comput. Simul* 190: 97–109.

Singh R, Singh G (2017). Wheat crop yield assessment using decision tree algorithms. *International Journal of Advanced Research in Computer Science* 8(5):1809-1817.

Tang XY, Yang WW, Liu Z, Li JC, Ma X (2024). Deep learning performance prediction for solar-thermal-driven hydrogen production membrane reactor via bayesian optimized LSTM. *International Journal of Hydrogen Energy* 82, 1402-1412.

URL1 https://biruni.tuik.gov.tr (access date: 06.06.2023).

URL2 https://data.tuik.gov.tr (access date: 04.08.2023).

URL3 Turkish State Meteorological Service, https://www.mgm.gov.tr (access date: 23.09.2023).

USDA (2020). U.S. Department of Agriculture, Oil crops yearbook, https://www.ers. usda.gov/data-products/oil-crops-yearbook/oil-Crops-Yearbook/ (access date:02.10.2020)

Xu L, Hou L, Zhu Z, Li Y, Liu J, Lei T, Wu X (2021). Mid-term prediction of electrical energy consumption for crude oil pipelines using a hybrid algorithm of support vector machine and genetic algorithm. *Energy* 222: 119955.

Yan, X., Ji, X., Meng, Q., Sun, H., & Lei, Y. (2024). A hybrid prediction model of improved bidirectional long short-term memory network for cooling load based on PCANet and attention mechanism. *Energy* 292: 130388.