# Modeling of wheat yield in the steppe region of Ukraine using machine learning techniques

Petro Hrytsiuk[1], Tetiana Babych[1], Olena Hladka[1] and Maryna Nehrey[2]

[1] National University of Water and Environmental Engineering, Soborna str., 11, Rivne, 33000, Ukraine
[2] Collegium Helveticum, ETH Zurich, Zürich, Switzerland

## Abstract

The task of this study is to evaluate the climatic factors impact on the detrended values of wheat yield using machine learning techniques. The average decadal temperature values for April, May, June and monthly amounts of precipitation for this period for five regions of the steppe zone of Ukraine were selected for the study. The work uses an innovative approach, according to which the detrended yield values are divided into two groups, labeled as "low yield" and "high yield". In the role of classifiers, six machine learning models were used, which were fitted to the available data and demonstrated classification accuracy above 80% on test samples. The support vector method and the random forest method are the most effective classifiers and provide 85% classification accuracy (on test data).

## Keywords

Wheat yield, climatic factors, machine learning, classification

## 1. Introduction

Grain production is one of the most important branches of the economy of Ukraine, ensuring the food needs of the population and a stable inflow of currency. The average annual production of cereals in Ukraine for 2019-2021 reached the level of 75 million tons (in 2021, a record crop of 84 million tons was harvested in Ukraine), and the average annual export during this time was 50 million tons [1].

At the same time, it is necessary to note the significant instability of grain production in Ukraine, associated with the impact of changing climatic factors, which have undergone significant changes in the last 30 years. This led to a change in the assortment of cultivated grain crops and the geography of their location [2, 3]. There is an increase in the production of heat-loving crops, such as corn, soybeans, and sunflowers in the chernozem zone of Ukraine and in the Polissia zone. In recent years, against the backdrop of climate change, the wheat share in the total grain harvest has decreased from 50% to 40%, and the corn share has increased from 15% to 42% [1]. Warming, which is accompanied by a decrease in the amount of precipitation, causes a negative impact on the yield of grain crops. The steppe region of Ukraine is particularly sensitive to changes in climatic factors, where frequent droughts lead to a significant drop in grain yields. Therefore, this region is losing its leading position in the grain production, instead, the share of the central and western regions of Ukraine is increasing.

Domestic consumption of grain in recent years did not exceed 20 million tons. This is approximately 30% of all grain production, and 70% of grain is exported. Thus, grain production from the main food resource of the country, which it was in the 20th century, turned into the largest source of foreign exchange for Ukraine and the key of its economic development. In the last three years alone, revenues from grain exports amounted to approximately 30 billion US dollars.

The basis for planning a long-term grain export strategy is the yield forecasting. This is a complex task, the essence of which is determined by the random nature of many influencing factors. Therefore, to solve this problem, it is advisable to apply intelligent data analysis techniques with modern computer technologies using.

*ICST-2024: Information Control Systems & Technologies, September 23-25, 2023, Odesa, Ukraine.*
p.m.hrytsiuk@nuwm.edu.ua (P. Hrytsiuk); t.iu.babych@nuwm.edu.ua (T. Babych); o.m.hladka@nuwm.edu.ua (O. Hladka); mnehrey@ethz.ch (M. Nehrey)

0000-0002-3683-4766 (P. Hrytsiuk); 0000-0001-6927-7313 (T. Babych); 0000-0003-4728-0663 (O. Hladka); 0000-0001-9243-1534 (M. Nehrey)

At the current stage, the most modern concepts of mathematical modeling are used to build predictive models, among which the machine learning techniques occupy a leading place. In this research, there was used such a powerful machine learning tool as classification methods.

The number of works devoted to the research of climatic factors impact on the grain crops yield in Ukraine is limited [2, 4, 5]. Complicated access to agroclimatic data is one of the reasons for the insufficient number of publications. From the view point of the grain crops cultivation, the territory of Ukraine can be divided into several agro-climatic zones: the steppe region, the black soil zone of the forest-steppe region, the western region. For each of these zones, the nature of yield dependence on climatic factors will be different. The main purpose of this study is to analyze and model the impact of climatic factors on wheat yields fluctuations in the steppe region of Ukraine.

## 2. Literature Review

Wheat production is the basis of Ukrainian agriculture, but climate change threatens it at risk in some regions of Ukraine. In a comprehensive analytical review conducted within the framework of the German-Ukrainian Agricultural Policy Dialogue project [5], the impact of climate changes on winter wheat yields in the three agroecological zones of Ukraine, as previously mentioned, was assessed. According to the authors' conclusion, the main concern is the fertile steppe zone, where the climate is hotter and drier, and frequent droughts are also observed.

The increase in the droughts frequency in recent years is seen as a major threat to agriculture. The author of [6] investigated the impact of climate change on the level of major agricultural crops production, as well as on Hungary's GDP. The paper [7] shows that the machine learning models shows have stronger predictive power than standard econometric approaches.

Scientific and technical progress contributed to the arrival of large volumes of statistical data from various branches of agriculture. This greatly expanded the possibilities of using computer technologies for the analysis and modeling of climatic effects on the agricultural crops yield. In recent years, there have been publications describing the machine learning methods application to forecasting the agricultural crops yield.

When developing a crop yield forecasting model in India to determine whether a given climate factor would affect yield using machine learning, a logistic regression model was found to be the most accurate [8]. The paper [9] provides an overview of some of the existing supervised and unsupervised machine learning models related to crop yield. Analytical models such as decision trees, random forests, support vector machines, Bayesian networks, and artificial neural networks are used to analyze the key factors impact on yield. These methods make it possible to analyze soil, climate and water regimes that significantly affect crop growth and yield. The review [10] presents machine learning (ML) approaches from the point of view of an applied economist.

The paper [11] examines the impact of extreme values of climatic factors on global agricultural yields. The paper [12] aims to identify the best yield prediction model that can help farmers decide which crop to grow based on climate conditions and nutrients present in the soil. In an analysis of yield prediction by three different supervised machine learning models, the authors concluded that the best accuracy was achieved with the Random Forest Classifier in both Entropy and Gini Criterion.

The study [13] proposed a machine learning-based forecasting system to forecast the yield of six agricultural crops at the countries in West Africa. Climatic and weather data and agricultural yields were combined to predict crop yields and build a decision support system for planning crop plantings. To build such a system, decision tree, multivariate logistic regression and k-model of nearest neighbors were used. It was found that the prediction results of the decision tree model and the K-Nearest Neighbor model are correlated to the expected data.

The structure of deep learning for forecasting yield using remote sensing data is presented in the paper [14]. An approach to dimensionality reduction based on histograms is proposed and the structure of a deep Gaussian process is demonstrated, with the help of which spatially correlated errors are eliminated and the accuracy of soybean yield forecasting (in a US county) is significantly increased.

One of the most powerful tools of machine learning is artificial neural networks. The paper [15] uses a semiparametric variant of a deep neural network, which can simultaneously account for complex nonlinear relationships in high-dimensional datasets. Using data on corn yield from the US Midwest, it

was shown that this approach outperforms both classical statistical methods and fully non-parametric neural networks in yield prediction.

In a previous study [4] Hrytsyuk et al. demonstrated that in terms of the influence of climate on wheat yield, all regions of Ukraine are divided into three agro-climatic zones. Annual changes in yields can be separated into a trend component and a deviation from the trend, explained by the influence of climate. Application of the binarization method to the yield trend deviation facilitated the development of machine learning classification models that can predict wheat yields with a prediction horizon of three months.

# 3. Methodology

In our model, the impact of climate on wheat yield is quantified through the cumulative effects of temperature and precipitation factors, each influencing distinct intervals of the growing season, as delineated in Table 1. Our research is divided into two main parts. In the first part, we use correlation and regression analyses to assess the effects of specific climatic factors - $t_1, t_2, \cdots, t_9, R_{10}, R_{20}, R_{30}$ on the deviations of yields $eps$ from their expected trend values. This analysis results in a regression model that is capable of predicting wheat yields for the current year.

In the second part, we perform binarization of these yield deviations $eps$. Each value of $eps$ is transformed into a binary factor $eps1$, which can be either 0 or 1. This binary classification enables us to treat the data for a specific area and year as a sample that belongs to one of two categories: high yield ($eps1$ = 0) or low yield ($eps$ = 1). This approach enables the application of machine learning techniques to develop classification-based predictive models for wheat yields.

## 3.1. Data Collection

The main food crop in Ukraine is wheat. The average annual production of wheat in Ukraine for 2019-2021 reached the level of 26.5 million tons. The weight share of wheat in grain exports during this time was 38%. This work is devoted to the study of the influence of climatic factors on fluctuations in wheat yield in the steppe region of Ukraine. Statistical climate data and wheat yield data for the period 2000-2021 for the Kherson, Mykolaiv, Odesa, Zaporizhzhya, Dnipro and Kirovohrad regions, which are located in the steppe region of Ukraine, were used for this research. Climatic characteristics were taken from [16], yield data were gotten from [1]. Successful wheat vegetation in the period from April to June has a decisive impact on the crop yield [4]. Average ten-day temperature values of April, May, and June and monthly amounts of precipitation for this period were used to assess the impact of climate on wheat yield (Table 1).

Table 1
Definition variables

| Variable | Definition | Period |
|---|---|---|
| t1, °C | Average temperature | from April 1st to April 10th |
| t2, °C | Average temperature | from April 11th to April 20th |
| t3, °C | Average temperature | from April 21st to April 30th |
| t4, °C | Average temperature | from May 1st to May 10th |
| t5, °C | Average temperature | from May 11th to May 20th |
| t6, °C | Average temperature | from May 21st to May 31th |
| t7, °C | Average temperature | from June 1st to June 10th |
| t8, °C | Average temperature | from June 11th to June 20th |
| t9, °C | Average temperature | from June 21st to June 30th |
| R10, mm | Amount of precipitation in | April |
| R20, mm | Amount of precipitation in | May |
| R30, mm | Amount of precipitation in | June |
| eps, c/ha | Detrended wheat yield | Year |

Ten-day temperature values make it possible to more accurately take into account the impact of external temperature at different stages of plant vegetation. Monthly precipitation amounts are used because many ten-day precipitation amounts in the steppe zone are close to zero. Statistical parameters of climatic factors and yield are given in Table 2. The parameter *eps* represents the deviation of yield from the trend value. Its magnitude and sign are determined by the impact of climatic factors on wheat yield in the current year.

Table 2
Summary statistics of numerical features

| | t1 | t2 | t3 | t4 | t5 | t6 | t7 | t8 | t9 | R10 | R20 | R30 | eps |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Minimum | 2.54 | 6.65 | 8.51 | 10.49 | 11.50 | 12.82 | 14.80 | 16.86 | 16.79 | 0.00 | 0.30 | 8.00 | -21.02 |
| Median | 8.82 | 10.79 | 12.34 | 14.31 | 16.40 | 19.05 | 20.26 | 21.33 | 22.73 | 24.50 | 40.90 | 53.00 | 0.96 |
| Mean | 8.59 | 11.03 | 12.94 | 15.27 | 16.91 | 19.24 | 19.90 | 21.52 | 22.20 | 30.81 | 48.73 | 63.37 | 0.00 |
| Maximum | 15.20 | 16.95 | 21.30 | 24.90 | 24.50 | 28.68 | 25.95 | 28.20 | 29.05 | 102.0 | 156.0 | 329.0 | 16.64 |
| Standard Deviation | 2.21 | 2.17 | 2.38 | 3.03 | 2.62 | 2.90 | 2.49 | 2.42 | 2.69 | 25.19 | 32.92 | 44.18 | 7.22 |

## **3.2.** Analysis of wheat yield dynamics

An analysis of wheat yield dynamics in the regions of Ukraine over the past 22 years shows that the yield is increasing [1,4]. The yield increase was the result of investment attractiveness increase of the grain industry and the significant investment that has flowed into the industry. As a result, the seed base has improved, agrotechnical culture has increased and the logistics network (elevators, grain wagons, ports) has developed. In 2021 a record cereals and legumes crop was harvested in Ukraine — 84 million tons. However, the tendency to increase grain yield is accompanied by significant yield fluctuations, the cause of which is mostly the weather and climate factors impact. The wheat yield dynamic in the Kherson region can serve as an illustration (Figure 1). The magnitude of deviations from the trend (detrended yield) directly depends on the impact of climatic factors, the main of which are droughts (2003 and 2012). To modeling of yield dynamics a linear trend model we used

$$trend_t = a_0 + t \cdot a_1 \quad (1)$$

Here $a_0$, $a_1$ - the trend coefficients, determined by statistical data using the least squares method [17]. An interval forecast is built on the linear trend basis, and for him the forecasting reliability level can be established. To construct an interval forecast of yield, it is necessary to check the hypothesis about a normal distribution of detrended yield eps

$$eps_t = y_t - trend_t. \quad (2)$$

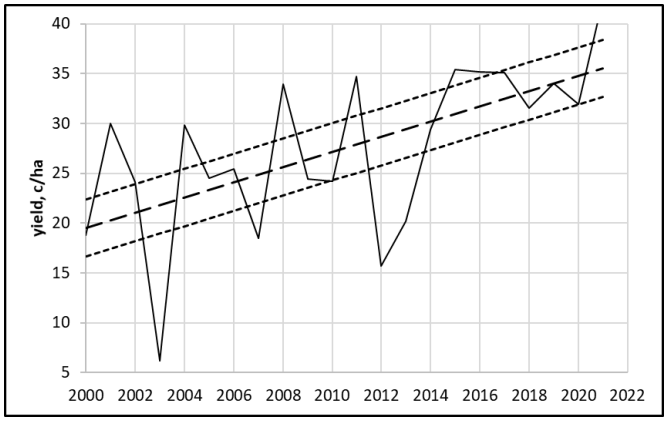


Figure 1: Wheat yield dynamics in the Kherson region. The dashed line is a linear trend. Dotted lines are high and low yield boundaries. Author's calculations according to [1]

To test the hypothesis of a normal distribution of detrended yields, a combined sample of detrended yields for six regions of the steppe zone of Ukraine (132 observations) was used. Statistical data on climate and wheat yield for the Kherson, Mykolaiv, Odesa, Zaporizhzhia, Dnipro and Kirovohrad regions were used. Similar weather and climate conditions and soil type allow these

regions to be united into one homogeneous region. The hypothesis of a normal distribution of detrended yields was confirmed by the Kolmogorov-Smirnov test.

### 3.3. Binarization of detrended yield

To solve many problems when planning an agrarian business, it is not necessary to have an accurate yield forecast. For example, to make a decision about investing in a specific project, it is enough to have an estimate of the future yield in terms of "high yield" – "low yield". The term "low yield" means a detrended yield value that is significantly lower than the average detrended yield value. All other cases are interpreted as "high yield". This approach enables the use of classification methods in yield forecasting.

We used the hypothesis of a normal distribution of detrended yield for the binary classification of detrended yield values in the categories "high yield" and "low yield". The main task of this study is to forecast low wheat yield values. To the "low yield" group, includes those yield values that with a probability of $p < 0.33$ are located on the integral curve of the normal distribution of detrended yields, that is, those for which the condition is fulfilled

$$F(eps) < 0.33. \quad (3)$$

Yield values for which condition (3) is not fulfilled will be assigned to the "high yield" group. To implement a classification approach to yield prediction, a binary variable eps1 is introduced, which has only two values: 1 ("low yield") and 0 ("high yield"). By the same time, the value of the eps1 variable is determined by the rule

$$eps1 = \begin{cases} 1, if\ F(eps) < 0.33; \\ 0, if\ F(eps) \geq 0.33. \end{cases} \quad (4)$$

According to the classification results, it was found that the number of cases classified as "low yield" represents 25.5%, the number of cases "high yield" values is 74.5%. Note that the normal distribution of detrended yields is not a necessary condition for their classification. This hypothesis only simplifies the classification procedure. The number of cases classified as 'low yield' represents 25%.

### 3.4. Analysis of climatic factors impact on the wheat yield

Such climatic factors as the average 10-day temperature and monthly precipitation cause fluctuations in wheat yield relative to the trend. Therefore, assessing the climatic factors impact on grain yield is an important tool when planning the placement of future crops and when planning future investments in the agricultural sector [18].

Table 3
The linear correlation coefficients between climatic factors and wheat yield for Kherson region (authors' calculations (according to [1, 16] )

| | t1 | t2 | t3 | t4 | t5 | t6 | t7 | t8 | t9 | R10 | R20 | R30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Eps | 0.05 | 0.04 | -0.29 | -0.52 | -0.80 | -0.40 | -0.46 | -0.49 | -0.09 | 0.53 | 0.27 | 0.17 |

To assess of climate impact on the wheat yield the mean (average) ten-day temperature and total monthly precipitation was used. Correlation coefficients reflecting the climatic factors' impact on wheat yield in Kherson region are shown in Table 3 As can be seen, the most noticeable impact on the wheat yield is caused by the mean ten-day temperature in May and June and the total monthly precipitation in April.

### 3.5. The multiple linear regression model. Features selection

According to the formulated assumptions, the wheat yield is formed under the impact of 12 climatic factors (9 temperature and 3 related to precipitation). To build a model of such a relationship, we will use the methods of multivariate correlation-regression analysis [17]. At the same time, the response eps is connected through the multiple regression equation with the factor features $t_1$, $t_2$, $t_3$, $t_4$, $t_5$, $t_6$, $t_7$, $t_8$, $t_9$, $R_{10}$, $R_{20}$, $R_{30}$. In the study, it is considered that climatic factors affect not the average yield, but the

deviation of the yield from trend value (detrended yield). Therefore, the detrended yield *eps* will be used as response

$$eps_t = y_t - tr_t. \quad (5)$$

A linear multiple regression equation of the following form will be used to model the dependence:

$$eps = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3 + \beta_4 t_4 + \beta_5 t_5 + \beta_6 t_6 + \beta_7 t_7 + \beta_8 t_8 + \beta_9 t_9 + \beta_{10} R_{10} + \beta_{20} R_{20} + \beta_{30} R_{30} + \varepsilon \quad (6)$$

Here $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_{10}, \beta_{20}, \beta_{30}$ are model parameters; $t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, R_{10}, R_{20}, R_{30}$ are model factors; eps is response; ε is model residual. The least squares method is usually used to determine model parameters.

To build regression models with a large number of parameters, it is necessary to have large data samples. For further research, it will be used the steppe region of Ukraine, which includes the Kherson, Mykolaiv, Odesa, Zaporizhzhia, Dnipro and Kirovohrad regions. The corresponding data sample contains 132 observations, each containing detrended yield and 12 climate factors. To process such large data sets, it is advisable to use specialized software.

We used the Python software environment and machine learning tools for data processing [19]. When studying statistic dependencies and developing a statistical model of a phenomenon, the problem lies in choosing the algorithm that is optimal for a specific case. In recent decades, the introduction of machine learning methods to solve the problems of classification and regression (quantitative response prediction) has begun. These methods include: multiple regression method, logistic regression method, linear discriminant analysis, random forest method, support vector machines, artificial neural networks.

## **3.6.** Machine Learning Algorithms

Recently machine learning-based systems are growing in popularity in research applications. In particular, the classification is an essential form of data analysis that formulates models while describing significant data classes [20]. In this work the several of classification algorithms for categorical predicting of wheat yield were used.

Logistic regression model. As noted above, to solve many problems when planning an agrarian business, it is enough to have an estimate of the future yield in terms of "high yield" – "low yield". This approach enables the use of classification methods in yield forecasting. The detrended yield was binarized according to the rule (4). As a result, a new data set, which differs from the one described in section 3.1 by replacing the numerical factor eps with the categorical factor eps1 was gotten. Each of the 132 observations of the new data set is characterized by a 12-dimensional feature vector.

The logistic regression model looks like this

$$P = F\left(X\beta'\right). \quad (7)$$

Here, F is a function whose values fall within the [0, 1] range and determine the probability P of a "low yield" occurrence. To implement the function F, a logistic distribution function is usually used:

$$F\left(z\right) = \frac{e^z}{1+e^z}. \quad (8)$$

Here, the parameter z is calculated from the ratio

$$z = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3 + \beta_4 t_4 + \beta_5 t_5 + \beta_6 t_6 + \beta_7 t_7 + \beta_8 t_8 + \beta_9 t_9 + \beta_{10} R_{10} + \beta_{20} R_{20} + \beta_{30} R_{30}. \quad (9)$$

To choose the best model, it is necessary to estimate the value of the coefficients $\beta i$ of the logistic regression model. Usually, the maximum likelihood method [17] is used for this.

The logistic regression model allows you to classify the samples according to the rule

$$p = \begin{cases} 1, if\ P > 0.5; \\ 0, if\ P \le 0.5. \end{cases} \quad (10)$$

The value p = 1 corresponds to the case of "low yield", the value p = 0 corresponds to the case of "high yield". In (10) "p" denotes a binary variable, while "P" represents the probability expressed by formula (7).

Evaluation of classifiers. The following indicators are used for evaluating the performance of the classifiers: matrix of errors (Confusion matrix), overall accuracy of classification (Accuracy), sensitivity of classification (Sensitivity), specificity of classification (Specificity) and the area under the ROC curve [21]. The Confusion matrix is built based on the results of classification by the model and the actual belonging of observations to classes [19]. Four cases are distinguished in the matrix:

- *TP (True Positives)* – the model correctly detected a low yield value;
- *FP (False Positives)* – the model wrongly recognized a high yield as a low yield;
- *FN (False Negatives)* – the model wrongly recognized a low yield as a high yield;
- *TN (True Negatives)* – the model correctly identified a case of high yield.

In the general case, the Confusion matrix has the following form (table 4):

Table 4
Confusion matrix

| Actual data | Test results | |
|---|---|---|
| | High yield | Low yield |
| High yield | TN | FP |
| Low yield | FN | TP |

Using the values of the elements of the error matrix, the following performance indicators of the binary classifier can be determined:

- *Sensitivity SE = TP / (TP + FP)*;
- *Specificity SP = TN / (TN + FN);*
- *Accuracy AC = (TP + TN) / (TP + FP + FN + TN).*

The area under the ROC curve AUC is a universal criterion for evaluating a classifier.

Linear discriminant analysis. Linear discriminant analysis (LDA) is a method of multivariate analysis that allows to evaluate the differences between two or more groups of objects according to several variables at the same time [22]. Discriminant analysis is based on the assumption that descriptions of objects of each k-th class are realizations of a multidimensional random variable distributed according to a normal law with mean $\mu_k$ and covariance matrix $C_k$. The task of discriminant analysis is to draw an additional axis $z(x)$

$$z(x) = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \cdots + \beta_m x_m , \quad (11)$$

which passes through the point cloud in such a way that projections onto it provide the best resolution into two classes.

Decision tree model. Decision trees used in data mining are of two main types: a classification tree and a regression tree (the predicted result is a real number) [23]. Decision trees split the space of objects according to some set of splitting rules. These rules make it possible to implement sequential dichotomous data segmentation. At each partitioning step, the amount of information about the variable under study (response) increases. When building a tree, it is important to set the optimal branching level.

The disadvantage of the decision tree method is instability: two trees built on the same training sample can give completely different resulting classes. This shortcoming can be eliminated by building ensembles of decision trees — a "Random Forest". The Random Forest classifier is based on bagging [24]. At the same time, several decision trees are built, repeatedly interpolating the data with replacement (bootstrap), and as a consensus answer, it gives the result of the voting of the trees (their average forecast). Boosting is another method for constructing a Random Forest [25]. Method of support vectors. The basic idea of a support vectors classifier is to build a separating surface using only a small subset of points that lie in the zone critical for separation, while other correctly classified points of the training sample outside this zone are ignored by the algorithm [26]. Since there can be many separating hyperplanes, the hyperplane that is the most distant from the training points is selected from among them. Method of cross-validation. Even with a large data set and random sampling has been applied to the training sample, the resulting model may be statistically unreliable. After all, another set of samples can lead to another model, which is significantly different from the first one. This shortcoming can be eliminated by cross-validation method [27].

1. First, you need to arbitrarily divide the initial data set into k groups ("folds") of approximately the same size.

2. One of the folds is selected as a data set for testing the model (testing set). The model is built based on the data of the remaining k-1 folds that form the training set. The MSE test error based on the observations of the testing set was calculated

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i\right)^2. \quad (12)$$

3. The process described above is repeated k times, each time using a different set as a testing set.

4. The total test MSE was calculated as the average of $k$ test MSEs. Similarly, other parameters of the model were averaged.

# 4. Results

## 4.1. Linear regression model

Linear regression model. We will build a linear regression model that will allow us to estimate the influence of climatic factors on wheat yield fluctuations. The values of the linear regression model estimates are shown in Table 5. The LM1 model is generally adequate (F-statistic=10.46; Prob (F-statistic) = 7.44e-14), but many factors in this model are insignificant (t1, t8, t9, P30). As can be seen from the table, factors t3, t5, t7, R10 have the greatest influence on yield.

Table 5.
Values of estimates of the linear regression model
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | eps | R-squared: | 0.513 |
| Model: | OLS | Adj. R-squared: | 0.464 |
| Method: | Least Squares | F-statistic: | 10.46 |
| Date: | Fri, 28 Jun 2024 | Prob (F-statistic): | 7.44e-14 |
| Time: | 11:06:52 | Log-Likelihood: | -400.20 |
| No. Observations: | 132 | AIC: | 826.4 |
| Df Residuals: | 119 | BIC: | 863.9 |
| Df Model: | 12 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 17.1230 | 7.664 | 2.234 | 0.027 | 1.948 | 32.298 |
| t1 | 0.2156 | 0.326 | 0.662 | 0.509 | -0.429 | 0.860 |
| t2 | -0.6916 | 0.326 | -2.123 | 0.036 | -1.337 | -0.047 |
| t3 | 1.1495 | 0.336 | 3.421 | 0.001 | 0.484 | 1.815 |
| t4 | -0.4280 | 0.229 | -1.869 | 0.064 | -0.881 | 0.025 |
| t5 | -1.2432 | 0.388 | -3.204 | 0.002 | -2.011 | -0.475 |
| t6 | 0.5951 | 0.302 | 1.971 | 0.051 | -0.003 | 1.193 |
| t7 | -1.3222 | 0.315 | -4.198 | 0.000 | -1.946 | -0.699 |
| t8 | 0.1193 | 0.334 | 0.358 | 0.721 | -0.541 | 0.780 |
| t9 | 0.3831 | 0.230 | 1.667 | 0.098 | -0.072 | 0.838 |
| R10 | 0.0782 | 0.019 | 4.020 | 0.000 | 0.040 | 0.117 |
| R20 | 0.0396 | 0.016 | 2.455 | 0.016 | 0.008 | 0.072 |
| R30 | 0.0125 | 0.011 | 1.119 | 0.265 | -0.010 | 0.035 |

## 4.2. Logistic regression model

As described above, the excess yield over the trend eps can be translated into the binary form eps1 according to rule (3). This makes it possible to build classification models of yield forecasting. First, let's build a GLM logistic regression model for a data set that describes 6 regions of the steppe region of Ukraine. To increase the statistical significance of the model during its construction, the basic principles of statistical modeling should be followed [28]. All data should be divided into two parts: the training sample (most of the original data used to build the model) and the control sample (the rest of the data that did not make it into the training sample). The control sample data are new

(unknown) to the built model, so they are used to evaluate the quality of the built model. In this work, we used the ratio of the amount of data in the training and control sample as 75% to 25%. Based on the GLM model, a forecast is built on the test sample. The accuracy of the predictive model presented in the table is 0.727 (24 out of 33 results matched).

### 4.3. A random forest model

A fragment of the decision tree of the problem is presented in Figure 2. At the first step, the algorithm determines the most significant factor and builds a dichotomy rule for it. Such a rule is the logical expression "t5 < 19.785". This means that the average temperature of the second decade of May is the main factor affecting wheat yield. If its value exceeds 19.785°C, the yield is likely to be low. In the next step, the obtained classes are again divided into subclasses according to another rule. This makes it possible to clarify the general rule of classification. At the next stage, a group of trees is combined into a random forest.

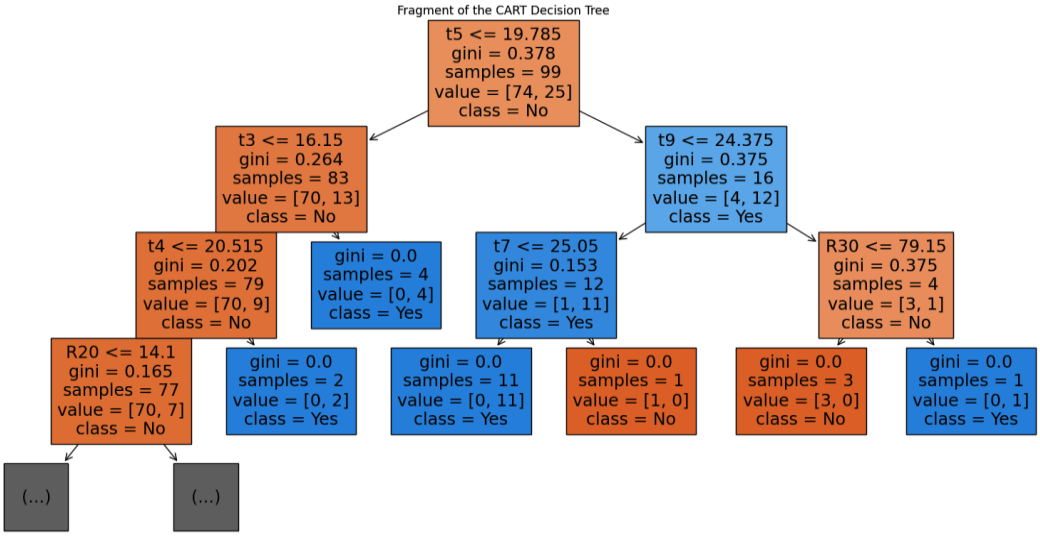


Figure 2: Classification of samples on one of the decision tree branches

Based on one of the random datasets, a model of a random forest of regression type is built using all influencing factors. As can be seen from Figure 3, in order to achieve high accuracy in classifying our data, it is necessary to use between 40 and 180 trees. This configuration of the model provides its best parameters: classification accuracy 0.88, average classification error MSE = 0.121. The importance of various traits for classification by the random forest method is illustrated in Figure 4.

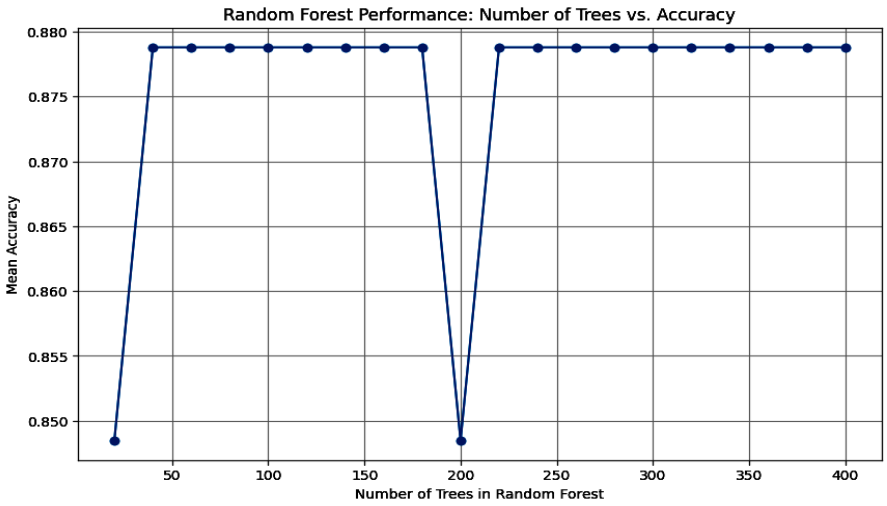


Figure 3: Dependence of classification accuracy on the number of trees

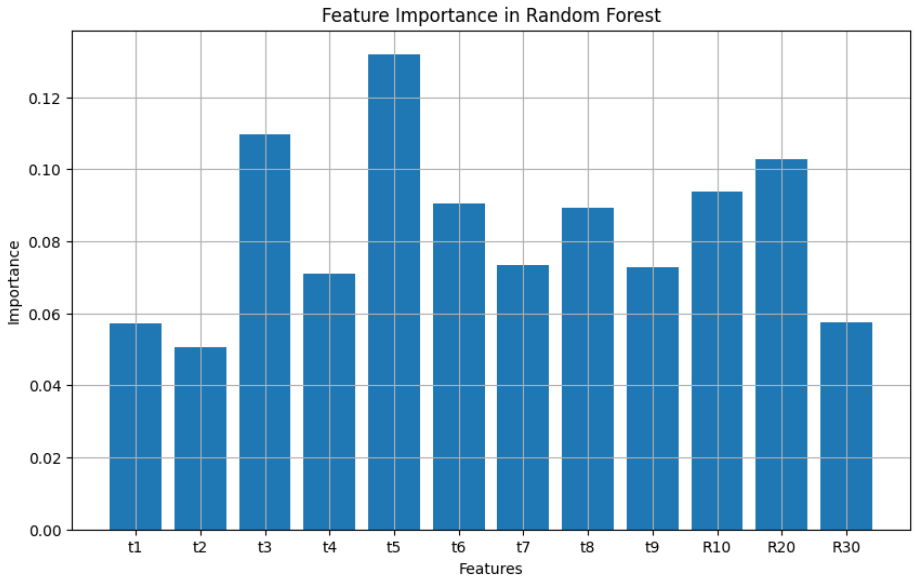


Figure 4: Estimation of the relative importance of object features by the random forest method

## 4.4. Comparison of the classification models effectiveness

The study used six methods to build binary classification models: linear discriminant analysis (LDA), support vector method with linear kernel function (SVML), support vector
method with radial kernel function (SVMR), decision tree method (CART), random forest method (RF), logistic regression method (GLM). The Python software environment was used to develop all models. The data were not standardized due to the same scale of indicators. In the SVMR model, the type of kernel function was taken as default - a Gaussian kernel with a radial basis function (RBF). The following parameters of the model were used: sigma = 0.4, C = 2. Here C means "the box constraint level", sigma — "kernel scale mode". Let's compare this models using the method described in [29]. The stages of this technique
are as follows:

- Data Division: The initial dataset is split into two parts: 75% is designated for constructing the training sample, and 25% is reserved for the control (test) sample.
- Model Training and Testing: The models are trained on the training sample and subsequently used to classify the control sample. Among the tested machine learning methods, the random forest method and the support vector method showed the best accuracy (Table 6).
- Cross-Validation Procedure: This process involves partitioning the initial dataset into several equal groups, with one acting as the control group at a time. Each group serves as the control group in rotation. During each cycle, the model is trained on the remaining data and tested on the control group. At the end of the process, the average performance metrics for the models are compiled. These metrics include accuracy (AC), sensitivity (SE), specificity (SP), and the area under the ROC curve (AUC).

Table 6
Quality criteria of machine learning methods

| | Accuracy | Sensitivity | Sensitivity | AUC |
|---|---|---|---|---|
| LDA | 0.789 | 0.560 | 0.865 | 0.70 |
| SVL | 0.859 | 0.680 | 0.919 | 0.76 |
| CART | 0.778 | 0.520 | 0.865 | 0.69 |
| RF | 0.838 | 0.560 | 0.932 | 0.94 |
| GLM | 0.839 | 0.680 | 0.892 | 0.74 |

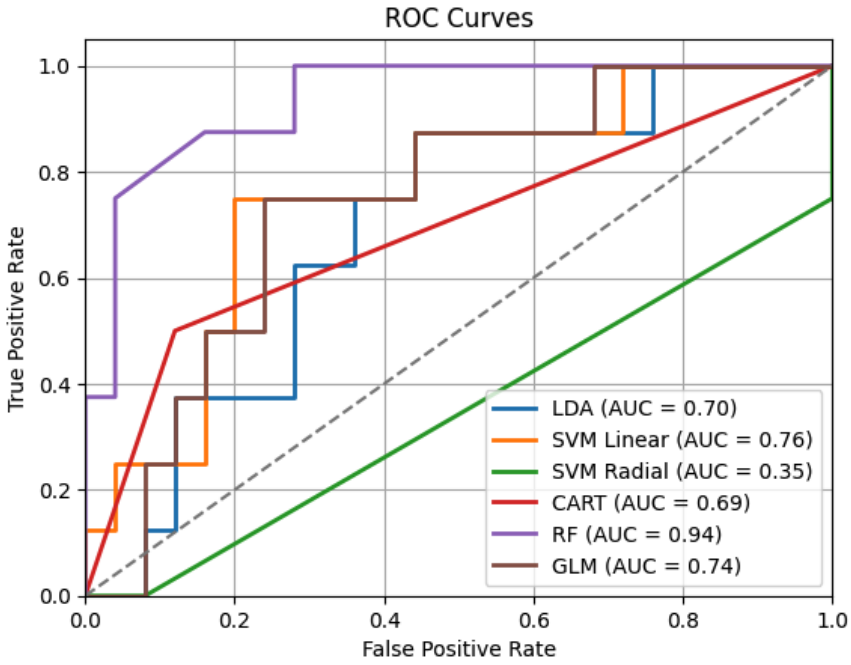


Figure 5: ROC curves for different classification models

A universal method for comparing the classifiers accuracy is ROC analysis [28]. ROC curves were constructed for the three used models (SVL, SVR, and RF) based on the complete table of initial data, which includes training and test samples (Figure 5). The area under the ROC curve AUC is a criterion for evaluating the classifier. For an ideal classifier, the ROC curve has the shape of a right angle. When evaluating the models by the AUC criterion, the best classifiers are the random forest method.

## 5. Conclusions

The importance of this research is determined by the fact that today there is an insufficient number of publications devoted to the impact of climate on the crop yield in Ukraine. There are even fewer publications that investigate this problem using machine learning techniques. The task of this work was to evaluate the climatic factors impact on detrended yield fluctuations using machine learning methods. This approach requires a large amount of data. To solve this problem, the data of six regions of the steppe zone, the climatic characteristics of which are similar, were combined.

We have shown that for assessing the weather factors impact on yield, it is sufficient to use the average ten-day values of temperature and monthly amounts of precipitation for the period from April to June. Models that reflect the impact of climatic factors on detrended wheat yield were built. It is shown that the temperature indicators in mid-May and early June and the amount of precipitation in April commit the greatest influence on the yield.

In this study, an approach was used, according to which trend deviations were divided into two groups, labeled as "low yield" and "high yield". Thus, the problem of yield forecasting was reduced to a classification problem. This approach, on the one hand, simplifies yield modeling, and on the other hand, allows for high accuracy in the classification of yield values. Six machine learning models were used as classifiers: discriminant analysis model, support vector models with linear and radial kernel functions, decision tree model, random forest model, and logistic regression model. To increase statistical significance, the cross-validation procedure with subsequent averaging of model parameters was used. All models were fitted to the available data set and demonstrated classification accuracy above 80% on test samples. The support vector method and logistic regression model showed better accuracy in classifying real data than other methods and provide a forecasting accuracy of 85% (on test data). This predicting accuracy is very good for complex natural processes.

The classification models built in this work make it possible to estimate in advance (in 3 months) the future wheat yield in terms of "high yield" — "low yield", creating a basis for making informed investment and marketing decisions. Since the algorithms used in this study are entirely accessible in terms of implementation, grain producers can use them for short-term yield forecasting. The

proposed method can be used to study the climate impact on the agricultural yield in other regions and countries.

Our research is a contribution to solving the problem of ensuring the sustainability of grain production in Ukraine. The obtained results can be used to stabilize the economic development of Ukraine and solve the food problem in the world.

## References


[1] State Statistics Service of Ukraine. URL: http://www.ukrstat.gov.ua.

[2] T. Adamenko, Climate change and agriculture in Ukraine: what farmers should know. German-Ukrainian agropolitical dialogue, Zapovit, Kyiv, 2019.

[3] Mitigation of Climate Change, Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge-New York, 2022.

[4] P. Hrytsiuk, T. Babych, S. Baranovsky, M. Havryliuk, Assessing of Climate Impact on Wheat Yield using Machine Learning Techniques. CEUR Workshop Proceedings, 3513, 2023, pp. 314–329.

[5] D. Muller, A. Jungandreas, F. Koch, F. Shirhorn, The impact of climate change on wheat production in Ukraine. Report on agricultural policy (APD), 2016.

[6] H.M. Zemankovics, Mitigation and adaptation to Climate Change in Hungary. In: J. Central Eur. Agric. Vol. 13(1), 2012, pp. 58-72.

[7] J. Ifft, R. Kuhns, K. Patrick, Can machine learning improve prediction – an application with farm survey data. Int. Food and Agribus. Manag. Rev 21 8 (2018) 1083–1098.

[8] B. Sharma, J. Yadav, S. Yadav, Predict Crop Production in India Using Machine Learning Technique: A Survey, in: 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), Noida, India, 2020, pp. 993–997.

[9] D. Elavarasan, D.R. Vincent, V. Sharma, A.Y. Zomaya, K. Srinivasan, Forecasting yield by integrating agrarian factors and machine learning models: a survey. Comput. Electron. Agric, 155, 2018, pp. 257–282.

[10] H. Storm, K. Baylis, T. Heckelei, Machine learning in agricultural and applied economics. Eur. Rev. Agric. Econ, 47 3 (2020) 849–892.

[11] E. Vogel, M.G. Donat, L.V. Alexander, M. Meinshausen, D.K. Ray, D. Karoly, N. Meinshausen, K. Frieler, The effects of climate extremes on global agricultural yields. Environ. Res. Let., 14 5 (2019).

[12] M. Kalimuthu, P. Vaishnavi, M. Kishore, Crop Prediction using Machine Learning, in: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). Tirunelveli, India, 2020, pp. 926–932.

[13] L.S. Cedric, W.Y. Hamilton, R. Aworkaa, J.T. Zoueucd, F.K. Mutomboa, M. Krichen, Crops yield prediction based on machine learning models: Case of West African countries. Sm. Agri. Tech, 2 (2022) 1–14.

[14] J. You, X. Li, M. Low, D. Lobell, S. Ermon, Deep gaussian process for crop yield prediction based on remote sensing data, in: the Proceedings of the AAAI Conference on Artificial Intelligence, 31 1 (2017) 4559-4565.

[15] A. Crane-Droesch, Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environ. Res. Lett., 13 11 (2018) 1–12.

[16] Meteorological data archive. URL: https://meteopost.com/weather/archive/

[17] N.R. Draper, H. Smith, Applied Regression Analysis. 3th Edition, Wiley, New York, 1998.

[18] P. Hrytsiuk, T. Babych, O. Mandziuk, Region sown areas portfolio optimization taking into account crop production economic risk. Global Journal Environemental Science Management, 5 (2019) 140-150.

[19] W. McKinney, Python for Data Analysis. O'Reilly Media, 2018.

[20] S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, in: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000, pp. 427–438.

[21] T. Fawcett, An Introduction to ROC Analysis. Pattern Recognit. Lett., 27 8 (2006) 861–874.

[22] A. Afifi, S. Azen, Statistical Analysis, Second Edition: A Computer Oriented Approach. Academic Press, New York, 1979.
[23] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and regression trees. Brooks/Cole Publishing, Monterey, 1984.
[24] L. Breiman, Bagging Predictors. Mach. Learn, 24, 1996, pp. 123–140.
[25] J. Friedman, Stochastic Gradient Boosting. Computational Statistics and Data analysis, 38 4 (2002) 367-378.
[26] T. Hastie, R. Tibshirani, J. Friedman, Model Assessment and Selection. The Elements of Statistical Learning, Springer Series in Statistics, 2009, pp. 219-259.
[27] D. Berrar, Cross-Validation. The Encyclopedia of Bioinformatics and Computational Biology, Academic Press, 2019.
[28] J. Gareth, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning. Springer, New York, 2013.
[29] M. Kuhn, K. Johnson, Applied Predictive Modeling. Springer, New York, 2013.