УДК 004.85:004.056]:631.1

APPROACH AND MODEL FOR FORECASTING WINTER WHEAT YIELD USING MACHINE LEARNING

Anatolii Tryhuba, Dr.Sci.(Engin.), Alla Zheliezniak, PhD, Inna Tryhuba, PhD,
Andrii Tatomyr, PhD
Lviv National Environmental University
1, V.Velykoho str., Dubliany-Lviv, 80381, Ukraine
e-mail: trianamik@gmail.com, azheleznjak@ukr.net, rinle@ukr.net, andrew.tatomyr@gmail.com

https://doi.org/10.31734/agroengineering2024.28.182

Tryhuba A., Zheliezniak A., Tryhuba I., Tatomyr A. Approach and model for forecasting winter wheat yield using machine learning

An analysis of the relevant subject area and scientific literature on the use of intelligent approaches for forecasting and planning activities in agriculture has been conducted. This analysis highlights the feasibility of employing machine learning to predict processes in agriculture. The purpose of this article is to develop a model for predicting winter wheat yields using historical data and machine learning algorithms, while taking into account the specific characteristics of processes and resource use in agriculture. The proposed forecasting approach for winter wheat yields relies on historical data and machine learning algorithms that consider the unique aspects of agricultural processes and the resources involved. The selection of an effective model for predicting winter wheat yield is based on a developed algorithm, which involves a systematic implementation of seven stages.

To prepare the data, the authors utilized intelligent analysis algorithms that assess the relationships between various factors affecting winter wheat yield. With qualitatively prepared data, the research substantiates the model for predicting winter wheat yield by evaluating its accuracy indicators. Three algorithms were chosen for the study: least squares (OLS), gradient boosting (XGBoost), and linear regression with polynomial features. Separate models were created for each algorithm and compared based on quality indicators. The findings indicate that the best model is the gradient boosting (XGBoost) model, which demonstrated the lowest values across all quality metrics - MSE, RMSE, MAE, and R-squared. Future research should focus on the development of an intelligent information system for planning agricultural processes, which includes a module for forecasting winter wheat yields based on the validated model proposed in this study.

Keywords: forecasting, yield, winter wheat, XGBoost algorithm, model, machine learning.

Тригуба А., Желєзняк А., Тригуба І., Татомир А. Підхід і модель прогнозування врожайності озимої пшениці на основі машинного навчання

Проведено аналіз предметної галузі та наукової літератури щодо використання інтелектуальних підходів до прогнозування та планування діяльності в сільському господарстві. Обґрунтовано доцільність використання машинного навчання для прогнозування процесів у сільському господарстві. Обґрунтовано модель прогнозування врожайності озимої пшениці на основі використання історичних даних, алгоритмів машинного навчання та врахування особливостей процесів і використання ресурсів у сільському господарстві. Запропонований підхід до прогнозування врожайності озимої пшениці ґрунтується на використанні історичних даних та алгоритмів машинного навчання, які враховують особливості виконуваних процесів та ресурсів, задіяних у сільському господарстві. Вибір ефективної моделі прогнозу врожайності озимої пшениці базується на розробленому алгоритмі, який передбачає системне виконання семи етапів. Для підготовки даних використано інтелектуальні алгоритми аналізу, які забезпечують оцінку взаємозв’язків між факторами, що впливають на врожайність озимої пшениці.

На основі якісно підготовлених даних обґрунтовано модель прогнозу врожайності озимої пшениці, здійснивши оцінку точних показників. Для дослідження обрано три алгоритми (система найменших квадратів (OLS), посилення градієнта (XGBoost) і лінійна регресія з поліноміальними характеристиками. У результаті були створені окремі моделі, порівняні за показниками якості. На основі результатів виявлено, що найкращою моделлю є модель посилення градієнта (XGBoost). Він має найнижчі значення з усіх показників якості - MSE, RMSE, MAE і R-квадрат. Подальші дослідження необхідно проводити в напрямку створення інтелектуальної інформаційної системи планування процесів у сільському господарстві з модулем прогнозування врожайності озимої пшениці на основі обґрунтованої нами моделі.

Ключові слова: прогнозування, врожайність, озима пшениця, алгоритм XGBoost, модель, машинне навчання.

Introduction. In recent years, iInformation technologies have fundamentally changed the concept of farming, making it more profitable, efficient, safe, and simple. Using smart technologies and precision farming systems. Farmers can build data-based knowledge in crop and livestock production. Smart technologies provide tools (sensors, drones, satellite images, etc.) to collect and integrate a variety of data, the study of which can provide better results for making management decisions. In 2022, the market value of smart agriculture in the world was 15.6

billion U.S. dollars. The expected growth of the smart agriculture market value in the world in 2027 is 33 billion U.S. dollars, which is almost twice as much as last year's figures [1]. Smart farming technology market research presented by Market & Market's Precision Farming Market by Technology also shows an expected global growth forecast through 2031 at a CAGR of 10.7% [2]. One of the most important topics in agriculture is the assessment and forecasting of crop yield and productivity improvement. Since productivity in agriculture depends on many factors, the use of machine learning models can provide more accurate predictions, allowing farmers to avoid unnecessary losses of resources, harvest, and optimize the technological process.

The field of crop production is characterized by the fact that a farmer or agronomist is forced to make decisions in conditions of incomplete and inaccurate input information (for example, regarding feeding and fertilizing plants in different areas of the field). This difficult task can be solved based on intelligent data analysis using mathematical models and intelligent systems. In animal husbandry, intelligent process planning information systems can help farmers better monitor the needs of individual animals, adjust feeding rations, prevent disease, and improve herd health. Using wireless IoT applications, farmers can monitor the location, well-being, and health of animals. Using the results of data and image processing based on machine and deep learning methods, it is possible to identify sick animals and separate them from the herd to prevent the spread of the disease.

Intelligent systems and smart technologies help to improve the sustainability of the agricultural enterprise, increase productivity, and ensure a reduced impact on the educational environment. The application of an intelligent information system using machine learning models for agriculture can be based on a combination of technologies such as machine learning, artificial intelligence, the Internet of Things, technologies and devices for collecting, processing, analyzing, and using data, an automated control system for individual processes (for example irrigation management), etc. Machine learning models are capable of detecting complex relationships between input and output data, as well as making predictions based on these relationships [4-7]. The use of machine learning for processing big data in the agricultural sector can generally improve the effectiveness of process management and decision-making.

Analysis of published data and problem setting. Solving process management problems with the use of intelligent information technologies is a fairly common solution in various applied fields [6; 9-11; 19], as it involves the search for new and improvement of existing approaches to the implementation of forecasting processes.

Many scientists in their research [8; 12; 13; 17] pay attention to intelligent data analysis, the application of machine learning methods, and the construction of predictive models, expert systems, and traditional models of statistical analysis for the search for dependencies and data processing. Many of these methods have their advantages and disadvantages, which primarily affect the choice of approach for building a predictive model.

Analysis of the latest research in the field of information technologies showed that some scientists [7; 14; 16; 20-23] were involved in the justification and development of intelligent information technologies for agriculture, including based on data processing based on machine learning models. Some authors in their works note that machine learning algorithms can be effectively used to study the relationships between production factors, which gives reason to consider the feasibility of their use for forecasting future values based on historical data [4-5; 18].

Therefore, the use of machine learning methods in the development of intelligent information systems for planning processes in agriculture will make it possible to increase the accuracy of forecasting the cultivation of crops (for example, winter wheat). At the same time, the solution of this scientific and applied problem depends on the need to collect high-quality and relevant data, their processing, and preparation for further training of the model. The implementation of the task is possible under the condition of the implementation of a smart approach and the principles of precision agriculture, the use of sensors, and other devices for collecting, accumulating, and transmitting data. The conducted research will make it possible to improve the quality of process management in agricultural enterprises and contribute to the achievement of expected productivity indicators in all branches of agriculture.

The purpose and objectives of the study. The purpose of the article is to substantiate the approach and model for forecasting winter wheat yields based on the use of historical data, and machine learning algorithms, and taking into account the peculiarities of processes and resource use in agriculture. To achieve this goal, the following tasks need to be solved:

1. to propose an approach and prepare data for training the model for predicting winter wheat yield;

2. to substantiate the model for predicting winter wheat yield based on the evaluation of quality indicators.

Research results. Building a predictive model for agriculture based on machine learning methods opens up new perspectives for resource management, yield planning, overcoming production risks, etc. This approach will allow the farmer to more accurately forecast various factors and indicators for the future. To substantiate the method of machine learning of the intelligent planning information system in the course of fulfilling the set tasks of the research, the task of planning the cultivation of winter wheat was chosen based on the data of one of the agricultural enterprises of the region, taking into account the selection of specific parameters, based on which the model for the intelligent information system will be built.

To choose a machine learning algorithm to solve the given task, such factors as the size, quality, and nature of the data, the purpose, and set goals of the research were taken into account, and further use of predictive data. The choice of a machine learning algorithm for prediction depends on the characteristics of the data set. For example, for processing structured data and forecasting time series, it can be effective to use algorithms such as Random Forest or Long Short-Term Memory (LSTM).

When working with a large amount of data and extremely complex relationships, deep learning algorithms, in particular neural networks, can be a powerful tool for accurate forecasts. Data preparation, commonly known as data preprocessing, entails handling raw data that has been gathered and readying it for use in machine learning algorithms. This process ensures that the input data for training the model is of high quality, thereby enhancing the effectiveness of the machine learning model in making accurate predictions.

To solve the task set during the research, namely the substantiation and construction of a machine learning model of an intelligent information system for planning processes in agriculture, an analysis and data collection was carried out on the example of winter wheat cultivation planning (Table). The data were taken from the Department of Agricultural Development of the Lviv Regional State Administration. In particular, data on nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, soil acidity, precipitation, and winter wheat yield were collected from a survey of agricultural enterprises in Lviv region in 2021. This made it possible to form a suitable dataset for machine learning.

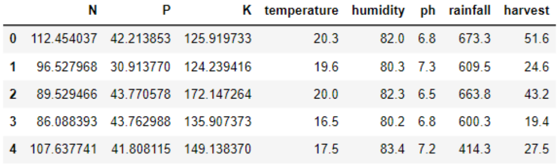
The specified data set involves using seven input variables (X1...X7) and, accordingly, one output variable (Y1). Based on data collection in specific farms, 1,516 instances of data on factors affecting the cultivation of winter wheat were obtained, which were distributed by attributes.

The interactive Jupyter Notebook environment was chosen to fulfill the research task, allowing to visualize data, conduct experiments, and debug machine learning models. The interactive Jupyter Notebook environment can be used for a wide range of machine learning tasks, including cleaning data and preparing it for machine learning; data analysis and understanding of their characteristics; development and training of machine learning models; evaluation of the quality of machine learning models; deployment of machine learning models in real systems.

Table Main characteristics of the dataset for building the model
Таблиця Основні характеристики набору даних для побудови моделі

The name of the data	Marking	A description or rationale for the choice
Nitrogen	N	Nitrogen is largely responsible for the growth of leaves on the plant
Phosphorus	P	Phosphorus is largely responsible for the growth of roots and the development of flowers and fruits
Potassium	K	Potassium is a nutrient that contributes to the proper performance of general plant functions
Temperature	temperature	Temperature in degrees Celsius
Humidity	humidity	Relative humidity in percent
Acidity	ph	Soil ph value
Amount of precipitation	rainfall	Amount of precipitation in mm
Crop capacity	harvest	Wheat yield, t/ha

After importing the necessary libraries into the interactive Jupyter environment, the loading and output of the initial data array were implemented. The data was imported from a CSV file into a DataFrame and saved in tabular format. The reason for using DataFrame is that it is the main data structure in the Pandas library that is used for convenient data processing and analysis. Figure 1 shows a fragment of imported data for further implementation of the given task.



	N	P	K	temperature	humidity	ph	rainfall	harvest
0	112.454037	42.213853	125.919733	20.3	82.0	6.8	673.3	51.6
1	96.527968	30.913770	124.239416	19.6	80.3	7.3	609.5	24.6
2	89.529466	43.770578	172.147264	20.0	82.3	6.5	663.8	43.2
3	86.088393	43.762988	135.907373	16.5	80.2	6.8	600.3	19.4
4	107.637741	41.808115	149.138370	17.5	83.4	7.2	414.3	27.5

Fig. 1. A fragment of the database on the cultivation of winter wheat

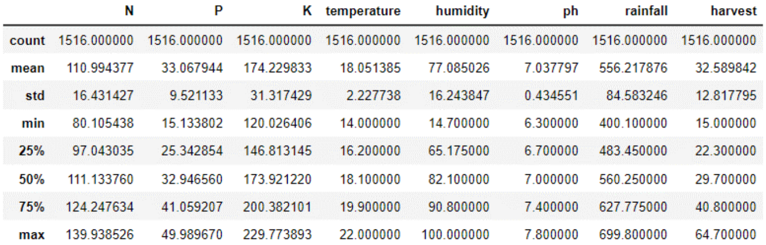
Рис. 1. Фрагмент бази даних щодо вирощування озимої пшениці

The next stage of working with the data loaded into the Jupyter environment involved the following three consecutive operations:

1) calculation of unique values by attributes and derivation of total number of them to obtain the distribution of values by this indicator and understand the characteristics of the data.

2) definition of the DataFrame dimension.

3) call the apply function for each column in the DataFrame (using an anonymous lambda function) to count the number of missing values in each column. This allowed to determine the number of missing values in each column.



	N	P	K	temperature	humidity	ph	rainfall	harvest
count	1516.000000	1516.000000	1516.000000	1516.000000	1516.000000	1516.000000	1516.000000	1516.000000
mean	110.994377	33.067944	174.229833	18.051385	77.085026	7.037797	556.217876	32.589842
std	16.431427	9.521133	31.317429	2.227738	16.243847	0.434551	84.583246	12.817795
min	80.105438	15.133802	120.026406	14.000000	14.700000	6.300000	400.100000	15.000000
25%	97.043035	25.342854	146.813145	16.200000	65.175000	6.700000	483.450000	22.300000
50%	111.133760	32.946560	173.921220	18.100000	82.100000	7.000000	560.250000	29.700000
75%	124.247634	41.059207	200.382101	19.900000	90.800000	7.400000	627.775000	40.800000
max	139.938526	49.989670	229.773893	22.000000	100.000000	7.800000	699.800000	64.700000

Fig. 2. Description of the study data

Рис. 2. Опис даних дослідження

The input data for training the model analyzed in this way (for example, for anomalies or missing data) can provide insight into how effective the machine learning model can be in further predictions.

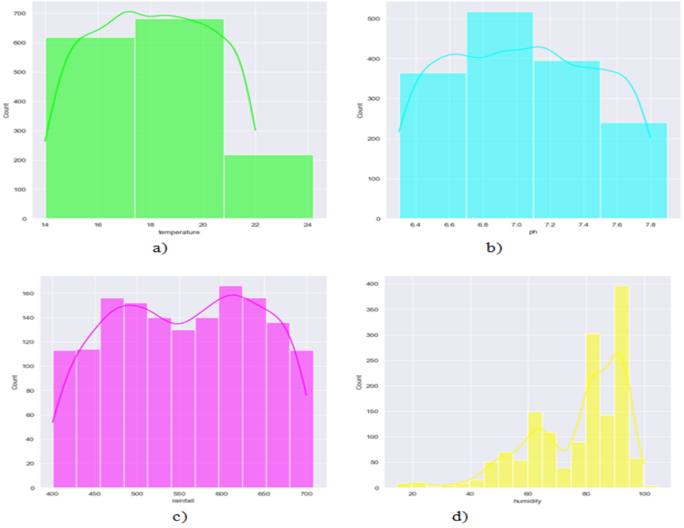


Fig. 3. Distribution of data values by such attributes as temperature (a), soil acidity (b), rainfall (c), and humidity (d)

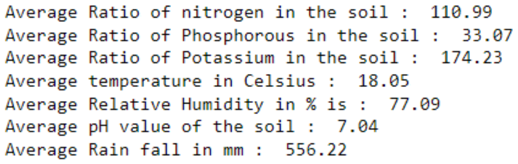
Рис. 3. Розподіл значень даних за такими атрибутами, як температура (а), кислотність ґрунту (б), кількість опадів (в), вологість (г)

Figure 3 shows how the distributions of the data for such attributes as temperature, soil acidity, precipitation, and humidity change. As for temperature, its values are described by a Weibull distribution. There is a distinct peak (mode) that indicates the most extended temperature during winter wheat cultivation. This is probably the optimal temperature for growth (17-21°C). The distribution shows the variability of the temperature. The highest peak may indicate optimal conditions for growth, while low or high temperatures hurt wheat yields, as evidenced by the falling tail on the graph. The distribution of soil acidity values shows a concentration of data within neutral or slightly acidic values, which is optimal for growing winter wheat (approximately 6.0-7.5). The distribution has a pronounced peak in this part, indicating that most soils in the region are suitable for wheat cultivation. The pH values that deviate from the optimum values impair nutrient absorption, which negatively affects yields. The distribution of precipitation is normal, with two peaks corresponding to medium and high levels of precipitation. Optimal values for growing winter wheat are 300-500 mm. The presence of tails at both ends of the distribution indicates periods of drought and high precipitation, which is also reflected in winter wheat yields. Moderate precipitation contributes to good wheat growth, while heavy or insufficient precipitation leads to a decrease in yield. The distribution of humidity values has a peak in the range of medium percentage humidity (60-70%), which is optimal for wheat growth, and high humidity (80-95%), which reduces winter wheat yields. Humidity is described by an adjusted polynomial distribution with larger deviations toward lower humidity. Relative humidity in the optimal zone promotes healthy plant development. Too low humidity leads to drying out of plants, and too high humidity leads to fungal diseases.

Each of the analyzed attributes (temperature, soil acidity, precipitation, humidity) affects the yield of winter wheat. The graphs of the distribution of values make it possible to determine within which limits the values of these factors are and which are the most frequent and optimal for growing winter wheat. Deviations from the optimal conditions, which are reflected in the extreme values of the distributions, lead to a decrease in the yield of winter wheat.

At the next stage of working with the data, the following operations were performed: the average value of nitrogen, phosphorus, and potassium concentration in the soil was determined and derived; the average temperature in degrees Celsius; the average value of relative humidity in percent is derived; the average value of pH in the soil; average rainfall in millimeters.

Figure 4 presents the results of the evaluation of statistical characteristics based on DataFrame df.



```
Average Ratio of nitrogen in the soil :  110.99
Average Ratio of Phosphorous in the soil :  33.07
Average Ratio of Potassium in the soil :  174.23
Average temperature in Celsius :  18.05
Average Relative Humidity in % is :  77.09
Average pH value of the soil :  7.04
Average Rain fall in mm :  556.22
```

Fig. 4. Estimation of statistical characteristics for dataset attributes
Рис. 4. Оцінка статистичних характеристик для набору атрибутів даних

Since the temperature regime in the summer period is of great importance for achieving wheat productivity, the analysis of the distribution of winter wheat productivity indicators was carried out on days when the temperature was in the regime from 15 to 28 degrees Celsius.

The results of the study are presented in Figure 5.

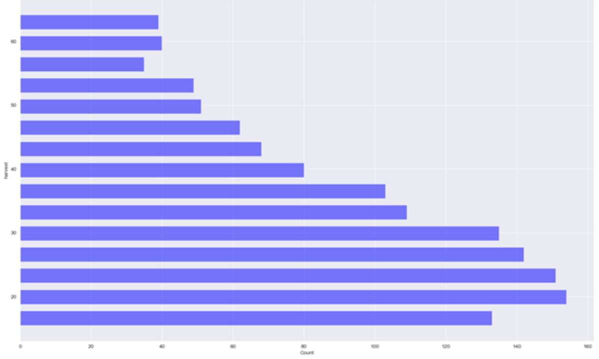


Fig. 5. Distribution of winter wheat yields by days with temperatures ranging from 15 to 28 degrees Celsius
Рис. 5. Розподіл врожайності озимої пшениці за днями з температурним режимом в діапазоні від 15 до 28 градусів Цельсія

The histogram (Fig. 5) shows the distribution of winter wheat cultivation when the temperature was in the range of 15 to 28°C, which is favorable for plant growth. The histogram has a pronounced peak in the range of 2.0-3.8 t/ha. This indicates that the largest number of days in the range of 15 to 28°C falls on this yield. The resulting histogram makes it possible to assess the efficiency of cultivation, identify optimal conditions for winter wheat and take measures to increase yields in future seasons.

The next step is to select the attributes having the highest correlation with the target feature "yield". This is done with the help of a correlation matrix, where their average value is determined for each input factor:

$$\bar{X}_i = \frac{1}{N}\sum_{i=1}^{N} X_{ij}, j = 1, m, \qquad (1)$$

The correlation matrix is determined by using the formula:

$$K_{ij} = \frac{\mathrm{cor}\left(X_{ij}, Y_1\right)}{\sigma\left(X_{ij}\right), \sigma\left(Y_1\right)}, \quad (2)$$

where $\mathrm{cor}\left(X_{ij}, Y_1\right)$ – correlation between the input data and the target feature Y_1.

The correlation between the quantitative values of the input factors (X1...X7) and the target feature "yield" (Y1) is determined by the formula:

$$\mathrm{cor}\left(X_{ij}, Y_1\right) = \frac{1}{N-1}\sum_{l=1}^{N}\left(X_{li} - \bar{X}_i\right)\left(X_{lj} - \bar{X}_j\right), i, j = 1, n, \quad (3)$$

At the next stage, the attributes that are most correlated with the resulting attribute “yield” are selected using the correlation matrix. The obtained research results indicate that the data set is qualitative and can be used to build a machine-learning model at the next stage of the research.

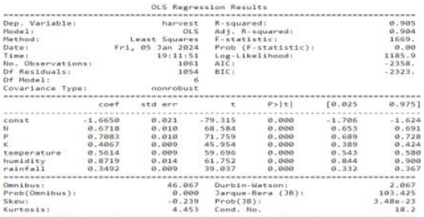
In the present research, it is accepted that the type and architecture of the model are chosen following the specific requirements for the task of forecasting the cultivation of an agricultural crop using the example of winter wheat.

Based on pre-loading into the interactive Jupyter Notebook environment, processed and analyzed data characteristics, key parameters can be determined and optimal learning algorithms can be selected to achieve the highest accuracy and efficiency of the model. Further selection and adjustment of hyperparameters will allow optimizing the model for the given task, providing the best forecasting results within the intelligent information system. During the research, it was established the feasibility of choosing the following algorithms for training the winter wheat yield forecasting model: least squares (OLS); gradient boosting (XGBoost); linear regression with polynomial features.

To begin with, a model based on the least squares algorithm was chosen. Building a model based on this algorithm is quite a popular approach in regression problems, as it allows finding a linear function that best reflects the relationship between independent and dependent variables, minimizing the sum of squared deviations.

After determining the coefficients of the model, it can be used to predict the values of the dependent variable based on new input data. Using the least squares algorithm provides an efficient and accurate way to build regression models (Figure 6).

The construction of a forecast model of winter wheat yield based on the gradient boosting algorithm, in particular XGBoost, included several stages. Based on the analysis and data preparation, the parameters of the model were determined.

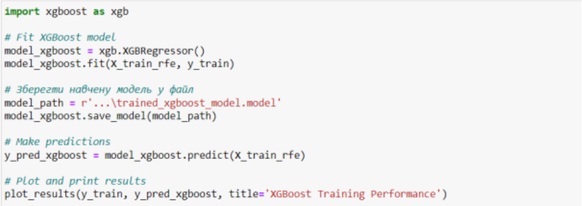


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                harvest   R-squared:                       0.905
Model:                            OLS   Adj. R-squared:                  0.904
Method:                 Least Squares   F-statistic:                     1669.
Date:                Fri, 05 Jan 2024   Prob (F-statistic):               0.00
Time:                        19:11:51   Log-Likelihood:                 1185.9
No. Observations:                1061   AIC:                            -2358.
Df Residuals:                    1054   BIC:                            -2323.
Df Model:                           6
Covariance Type:            nonrobust
===============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const         -1.6650      0.021    -79.315      0.000      -1.706      -1.624
N              0.6718      0.010     68.584      0.000       0.653       0.691
P              0.7083      0.010     71.759      0.000       0.689       0.728
K              0.4067      0.009     45.954      0.000       0.389       0.424
temperature    0.5614      0.009     59.696      0.000       0.543       0.580
humidity       0.8719      0.014     61.752      0.000       0.844       0.900
rainfall       0.3492      0.009     39.037      0.000       0.332       0.367
==============================================================================
Omnibus:                       46.067   Durbin-Watson:                   2.067
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              103.425
Skew:                          -0.239   Prob(JB):                     3.48e-23
Kurtosis:                       4.453   Cond. No.                         18.2
==============================================================================
```

Fig. 6. Model based on ordinary least squares (OLS) algorithm

Рис. 6. Модель на основі звичайного алгоритму найменших квадратів (OLS)

After that, the model training process was implemented using gradient descent to improve the quality of forecasts (Figure 7). Using an ensemble of decision trees, XGBoost improves over multiple iterations, weighting errors and correcting them at each step. The final step involves tuning the hyperparameters to ensure optimal model performance.

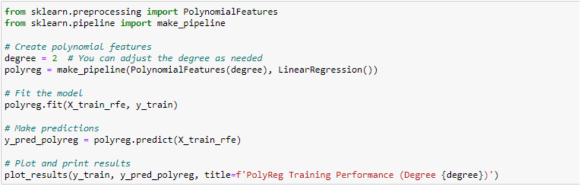


```
import xgboost as xgb

# Fit XGBoost model
model_xgboost = xgb.XGBRegressor()
model_xgboost.fit(X_train_rfe, y_train)

# Зберегти навчену модель у файл
model_path = r'...\trained_xgboost_model.model'
model_xgboost.save_model(model_path)

# Make predictions
y_pred_xgboost = model_xgboost.predict(X_train_rfe)

# Plot and print results
plot_results(y_train, y_pred_xgboost, title='XGBoost Training Performance')
```

Fig. 7. Code for creating a gradient boosting model (XGBoost)

Рис. 7. Код для створення моделі посилення градієнта (XGBoost)

Construction and training of a predictive model using a linear regression algorithm with polynomial features was also carried out.



```
from sklearn.preprocessing import PolynomialFeatures
from sklearn.pipeline import make_pipeline

# Create polynomial features
degree = 2  # You can adjust the degree as needed
polyreg = make_pipeline(PolynomialFeatures(degree), LinearRegression())

# Fit the model
polyreg.fit(X_train_rfe, y_train)

# Make predictions
y_pred_polyreg = polyreg.predict(X_train_rfe)

# Plot and print results
plot_results(y_train, y_pred_polyreg, title=f'PolyReg Training Performance (Degree {degree})')
```

Fig. 8. Building a model using a linear regression algorithm with polynomial features

Рис. 8. Побудова моделі на основі алгоритму лінійної регресії з поліноміальними ознаками

In the beginning, the original features of the data are considered, and then polynomial features are generated and raised to powers to obtain non-linear dependencies. This model is trained using a training set, which decides the optimal weight values for each feature, minimizing the loss function.

A predictive model based on a linear regression algorithm with polynomial features is evaluated on the test set to evaluate its predictive ability and avoid overtraining. This approach allows linear regression to adapt to non-linear patterns in the data, making it effective for modeling complex relationships.

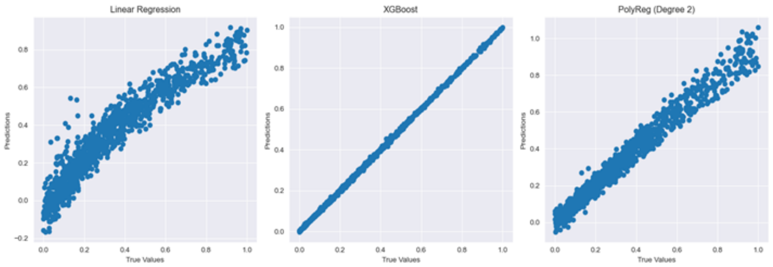


Fig. 9. Comparative visual representation of learning results of three types of predictive models
Рис. 9. Порівняльне візуальне представлення результатів навчання трьох типів прогностичних моделей

To choose the optimal model for building forecasts, the quality of the models was evaluated based on the metrics discussed in the previous sections of the qualification work.

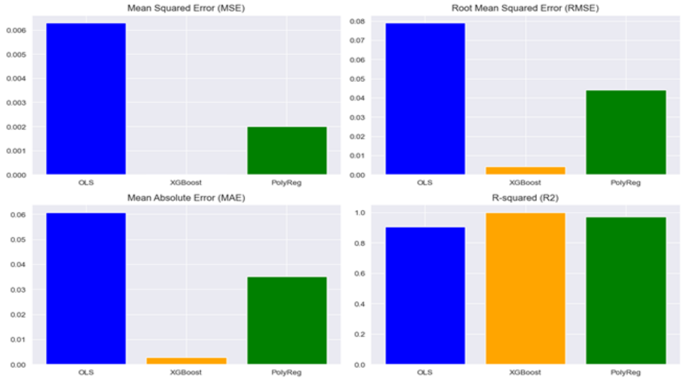


Fig. 10. Results of evaluation of accuracy indicators of the obtained forecast models of winter wheat yield
Рис. 10. Результати оцінювання показників точності отриманих прогнозних моделей урожайності озимої пшениці

The MSE metric measures the mean of the squared differences between observed and predicted values. RMSE is the square root of MSE and measures the root mean square error. MAE measures the average absolute value of the differences between observed and predicted values. R-squared indicates how much of the variability in the explained variable is explained by the model.

Based on the histograms presented in Fig. 10, it can be concluded that the XGBoost model showed the best results for all evaluated indicators - MSE = 0.0001, RMSE = 0.003, MAE = 0.003, and R^2 = 0.99. This indicates its good ability to provide accurate winter wheat yield forecasts. PolyReg also demonstrates good results (MSE=0.002, RMSE=0.043, MAE=0.035, and R^2=0.96), although less accurate than XGBoost. OLS, although a basic model, performed the worst. This analysis emphasizes the importance of model selection to ensure forecast accuracy, especially in complex systems such as agriculture.

Based on the obtained data, tge authors can conclude that the best model is the model of gradient boosting (XGBoost). It has the lowest values of all quality metrics, including MSE, RMSE, MAE, and R-squared. This means that it makes the most accurate forecasts of the values of the dependent variable, namely the yield index of winter wheat. This is because the XGBoost model is good at processing different types of data, such as numerical indicators (nitrogen, phosphorus, potassium) and meteorological conditions (temperature, humidity, precipitation). This allows the model to learn to find relationships between these indicators for accurate forecasting. In particular, the XGBoost model builds an ensemble of decision trees that allow it to identify complex nonlinear interactions between given parameters and their impact on winter wheat yield. In addition, XGBoost has built-in regularization methods that prevent overfitting.

Different climatic, soil, and agro-climatic conditions in other regions vary significantly and affect the relationship between yield and temperature, humidity, precipitation and soil composition. Due to the difficulty of collecting similar data, which was collected for the study from only one region (Lviv region), there are limited opportunities to use the model in other regions of Ukraine. The data cover only one growing season (2021), which limits the ability to use the model for long-term yield forecasts. Different climatic conditions can vary significantly over the years, so research based on long-term observations is needed to improve the model's accuracy.

While XGBoost performed well on the data in this study, the model can be sensitive to the choice of characteristics, such as data size and the presence of outliers. In cases of significant changes in the data or increasing persistence, other machine learning methods may provide better results.

The results were obtained and the approach was used to predict winter wheat yields. They can be adapted for other crops, allowing for the expansion of the use of crop yield prediction models based on machine learning methods.

Based on the selected predictive model, it is possible to create an intelligent information system for planning processes in agriculture with a module for planning the harvest of winter wheat with a known predicted yield, which includes:

1. Determination of the expected volume of the gross collection of crops by using the forecasted yield and planned sown areas.

2. Plan the need for agricultural machinery and workers involved in technological processes in crop production.

3. Planning routes and harvest schedules. This will make it possible to ensure efficient and timely harvesting and reduce its losses.

4. Planning of crop storage, and cleaning of harvested wheat. This is necessary to keep the harvest in good condition before its sale on the agricultural market.

CONCLUSIONS

1. The proposed approach to forecasting winter wheat yields is based on the use of historical data and machine learning algorithms that take into account the specifics of the processes performed and the resources involved in agriculture. The choice of an effective model for predicting winter wheat yields is based on the developed algorithm, which involves the systematic implementation of seven stages. The peculiarity of this approach is that the formation of historical data for model training is based on attributes that reflect the peculiarities of agricultural processes and characterize the state of natural resources. To prepare the data, the authors of the study used intelligent analysis algorithms that provide an assessment of the relationships between the factors that affect the yield of winter wheat.

2. Based on the qualitatively prepared data, the researchers substantiated the model for predicting winter wheat yields by evaluating the accuracy indicators. Three algorithms (least squares (OLS), gradient boosting (XGBoost), and linear regression with polynomial features) were chosen for the study. As a result, separate models were created and subsequently compared by quality indicators. Based on the results, it was found that the best model is the gradient boosting model (XGBoost). It has the lowest values of all quality metrics - MSE, RMSE, MAE, and R-squared. This means that it provides the most accurate predictions of the dependent variable, namely the winter wheat yield index. Further research should be carried out in the direction of creating an intelligent information system for planning processes in agriculture with a module for forecasting winter wheat yield based on the model the authors have substantiated.

REFERENCES

1. Shahbandeh M. Global market size of smart farming 2021-2027. Statista, 2023. URL: https://www.statista.com/statistics/720062/market-value-smart-agriculture-worldwide/.

2. Precision Farming Market. Market & Market's, 2023. URL: https://www.marketsandmarkets.com/Market-Reports/precision-farming-market-1243.html.

3. Hrynevych O., Blanco Canto M., Jiménez García M. Tendencies of Precision Agriculture in Ukraine. *Disruptive Smart Farming Tools as Cooperation Drivers. Agriculture*. 2022. No 12(5). P. 698.

4. Zahmatkesh Z., Goharian E. Comparing Machine Learning and Decision Making Approaches to Forecast Long Lead Monthly Rainfall. *The City of Vancouver, Canada. Hydrology*. 2018. No 5(1). P. 10.

5. Yurynets R., Yurynets Z., Grzebyk M., Kokhan M., Kunanets N., Shevchenko M. Neural Network Modeling of the Social and Economic. *Investment and Innovation Policy of the State, Proceedings of the 4nd International Workshop on Modern Machine Learning Technologies and Data Science Workshop MoMLeT&DS* 2022. Leiden, The Netherlands, November 25-26, 2022. Pp. 252-262.

6. Gilliland M. The value added by machine learning approaches in forecasting. *International Journal of Forecasting.* 2020.Volume 36, Issue 1, Pp. 161-166.

7. Liakos K.G., Busato P., Moshou D., Pearson S., Bochtis D. Machine Learning in Agriculture. *A Review. Sensors.* 2018. No18. P. 26-74.

8. Kunanets N., Vasiuta O., Boiko N. Advanced technologies of big data research in distributed information systems. *14th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT),* vol. 3, 71–76, September 2019.

9. Tryhuba A., Boyarchuk V., Tryhuba I., Ftoma O., Padyuka R., Rudynets M. Forecasting the Risk of the Resource Demand for Dairy Farms

Basing on Machine Learning. *Proceedings of the 2nd International Workshop on Modern Machine Learning Technologies and Data Science (MoMLeT+DS 2020).* Volume I: Main Conference. Lviv-Shatsk, Ukraine, June 2-3, 2020. Pp. 327-340.

10. Tryhuba A., Kondysiuk I., Tryhuba I., Boiarchuk O., Tatomyr A. Intellectual information system for formation of portfolio projects of motor transport enterprises. *CEUR Workshop Proceedings*, 2022, 3109. Pp. 44–52.

11. Koval N., Tryhuba A., Kondysiuk I., Tryhuba I., Boiarchuk O., Rudynets M., Grabovets V., Onyshchuk V. Forecasting the Fund of Time for Performance of Works in Hybrid Projects Using Machine Training Technologies. *Proceedings of the 3nd International Workshop on Modern Machine Learning Technologies and Data Science Workshop. Proc. 3rd International Workshop (MoMLeT&DS 2021).* Volume I: Main Conference. Lviv-Shatsk, Ukraine, June 5-6, 2021. pp. 196-206.

12. Lycett M., Marcos E., Storey V. Model-driven systems development: An introduction. *European Journal of Information Systems*. 2007. 16. 10.1057/palgrave.ejis.3000684.

13. Shiqiang Zh, Ting Y., Tao X., Hongyang Ch., Schahram D., Sylvain G., Deniz G., Ekram H., Yaochu J., Feng L. et al. Intelligent Computing: The Latest Advances, Challenges, and Future. *Intell Comput*. 2023. No 2:0006.

14. Kamilaris A., Prenafeta-Boldú F. Deep learning in agriculture. *A survey. Computers and Electronics in Agriculture*, 2018. Vol. 147. Pp. 70-90.

15. Nawarecki E., Kluska-Nawarecka S., Wilk-Kołodziejczyk D., Śnieżyński B., Legień G. Integrated Multi-functional LPR Intelligent Information System, 2018.

16. Zhelyeznyak A., Ptashnyk V. Modelling the architecture of a planning system for agricultural enterprises. Selected Papers from the Xth International Conference “Information technologies in energy and agro-industrial complex”, ITEA 2021, October 6-8, 2021, 2022. Pp. 32-37.

17. Hua Leong F., Farn Haur C. Deep Learning-Based Text Recognition of Agricultural Regulatory Document. In: Bădică, C., Treur, J., Benslimane, D., Hnatkowska, B., Krótkiewicz, M. (eds) Advances in Computational Collective Intelligence. ICCCI 2022. *Communications in Computer and Information Science*, vol. 1653. Springer, Cham.

18. Vyklyuk Y., Radovanovic M., Pasichnyk V., Kunanets N., Sydor P. Forecasting of forest fires in portugal using parallel calculations and machine learning. *Proceedings of the XVIII International Conference on Data Science and Intelligent Analysis of Information*, June 4–7, 2018, Kyiv, Ukraine, pp. 39-49.

19. Yan-e D. Design of Intelligent Agriculture Management Information System Based on IoT, Fourth International Conference on Intelligent Computation Technology and Automation, Shenzhen, China, 2011. Pp. 1045-1049.

20. Mohd J., Abid H., Pratap Singh R., Rajiv S. Enhancing smart farming through the applications of Agriculture 4.0 technologies. *International Journal of Intelligent Networks*. 2022. Volume 3, pp. 150-164.

21. Yaganteeswarudu A., Saroj Kumar B., Aruna V. Smart farming using artificial intelligence: A review Engineering Applications of Artificial Intelligence. Volume 120, April 2023, 105899.

22. Benos L., Tagarakis A., Dolias G., Berruto R., Kateris D., Bochtis D. Machine Learning in Agriculture. *A Comprehensive Updated Review*. *Sensors*. 2021. No 21(11). P. 37-58.

23. Elbasi E., Zaki C., Topcu AE., Abdelbaki W., Zreikat AI., Cina E., Shdefat A., Saker L. Crop Prediction Model Using Machine Learning Algorithms. *Applied Sciences*. 2023. No 13(16). P. 9288.

Стаття надійшла 21.08.2024