Modeling and forecasting tasks of agriculture based on machine learning

Baratbek Sabitov[1], *Asel* Kartanova[2,*], *Talant* Kurmanbek uulu[3], *Nazgul* Seitkazieva[3], *Ainura* Dyikanova[4], and *Aida* Orozobekova[2]

[1]Kyrgyz National University named after Zh.Balasagyn, Bishkek, Kyrgyz Republic
[2]Kyrgyz State Technical University named after I.Razzakov, Bishkek, Kyrgyz Republic
[3]Kyrgyz State University named after I.Arabaev, Bishkek, Kyrgyz Republic
[4]Kyrgyz National Agrarian University named after K.I.Scriabin, Bishkek, Kyrgyz Republic

Abstract. Continuous advances in computer technology have provided good support for the expansion of agricultural research using machine learning. This article considered the current problem of yield forecasting using methods and algorithms of machine learning to support management decision-making in the agricultural sector. For a set of data collected from five districts of the Issyk-Kul region, such as weather conditions, soil characteristics and pre-processing of the sowing area, a study of the yield of various crops using advanced machine learning algorithms, such as the support vector method, k-nearest neighbors, variants of gradient boosting and random forest, etc., is demonstrated. To assess the accuracy of the models, a comparative analysis with the results of multiple regression was carried out. It is shown that powerful regression machine learning algorithms like k-nearest neighbors (KNN), random forest (RF), support vector method (SVR) and gradient boosting (GBR) give tangible results in prediction compared to other machine learning methods (MAPE=10%). The calculation results showed the effectiveness of using algorithms with ensemble methods to solve the problems of yield forecasting, and that environmental factors (weather conditions) have a greater impact on yield than soil genotype.

Keywords: yield, algorithms, model, machine learning, agricultural problems.

1 Introduction

Food security plays a key role worldwide and is an important lever in the success of a country's economy as a whole. In this case, its main component is the yield of crops. Yields found in many agricultural projects belong to complex categories in modeling and forecasting tasks. It is a trait that, in general, can be defined as an amalgamation of many factors of nature and the natural conditions of the environment where a particular crop is grown and requires comprehensive research. Accurate crop yield prediction requires a fundamental understanding of the functional relationship between crop yield and factors

[*] Corresponding author: a.kartanova@gmail.com

related directly to specific natural phenomena like temperature, humidity, and precipitation, as well as indirectly affecting yields. As indirect factors, the characteristics of the sowing areas, the quality and composition of the soil treated with fertilizers, and the results of seed studies of plant genotype for the region under study can be taken into account. At the same time, it is very important to have reliable data sets, as well as powerful algorithms for modeling and predicting such processes in order to identify the relationship.

To build a yield model, it is also necessary to establish other relationships, for example, with the phenomena of climate change, which is currently typical for many regions of Kyrgyzstan. For example, despite the dry year of 2021, many regions of Kyrgyzstan were characterized by a shortage of irrigation water, abnormally hot days. But in Issyk-Kul region in the period from August 15 to September 15, 2021 there were incessant heavy rains, ambient humidity, favorable temperature conditions, coincided with the period of fruit bearing and ripening of fruits, rapid growth of plants and greatly affected the yield of crops, especially potatoes, fruit trees and other crops. At the same time, along with the quantity of agricultural products, the quality of products obtained from these areas has improved greatly. This factor of nature suggests that weather conditions, rather than genotype or consideration of soil characteristics, play an important role for yield.

In this direction for wheat yields in [1] considered, the application of k-nearest neighbors and the decision tree. In [2], using naive Bayesian process and k-nearest neighbor algorithms to the data of companies on the development of new crop varieties, the prediction problem was considered. In [3], based on the random forest algorithm and linear regression, the problem of forecasting, taking into account regional and global features, when choosing to grow certain species of plants was solved. To study rice yield in [4], using the method of reference vectors obtained the results of forecasting, taking into account the topographic features [5] of the area with subtropical climate based on algorithms of decision tree, logistic regression and k-nearest neighbors. In [6], using methods of naive Bayesian process, random forest, neural networks, decision tree, and support vector machines, the classification problem for yield problems was investigated. In [7-8], yield prediction was investigated using neural network analysis of fruit images. Specifically, the recognition problem was solved using image segmentation for fruit detection and yield estimation in apple orchards. Proper agronomic farming with the use of fertilizers, continuous observation and timely identification of the needs of agricultural plants, plays a key role in obtaining the desired yield was studied in [9].

Research in recent years shows that the results of a set of algorithms applying ensemble learning are attractive because ensemble methods attempt to construct a set of hypotheses and combine them for use to many applied problems, including yield prediction problems [10,11,12,13].

There are many factors that affect crop yields, such as the area planted with proper pre-tillage, the effective use of irrigation systems and its improvement, take into account changes in weather and characteristics of rainfall and temperature. With all of the above conditions, this paper investigated crop yields using machine learning methods and algorithms.

The main emphasis was placed on strong factors affecting yields such as, tillage and basic weather conditions. In collecting data, these factors were taken into account individually and regionally. When forming the database, the peculiarities of soil characteristics of the regions under study were also studied. In our case, soil acidity and composition were selected for the northern coast, eastern and southern regions of the Issyk-Kul region for each separately. When creating forecasts, the scale of the main agricultural regions of the study area was taken into account.

2 Materials and Method

When solving the problem of yield forecasting, the most commonly used algorithms were considered, such as:

Logistic regression (LR). As one of the most widely used classification algorithms, logistic regression shows satisfactory performance at relatively low computational cost. The logistic function for logistic regression with several variables can be written as follows (1):

$$P(X)=\frac{e^{\beta_0+\beta_0X_1+\beta_2X_2+...+\beta_nX_n}}{1+e^{\beta_0+\beta_0X_1+\beta_2X_2+...+\beta_nX_n}}, \tag{1}$$

where $X = (X_1, X_2, ..., X_n)$ is n predictors of the set P, they have the corresponding training parameters β, and P(X) is the probability of being positive given the predictors X. The maximum likelihood method is often used to estimate the training parameters β_i for the regression problem, i.e. the L (2) function is constructed to determine the parameters:

$$L(a_j,\sigma_j)=\prod_{i=1}^{n} P(X_i^j, a_j, \sigma_j)=\frac{1}{(\sigma_j\sqrt{2\pi})^n}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(X_i^j-a_j)^2}{\sigma_j^2}\right), j=0,n; \tag{2}$$

where $\hat{a}_j$ и $\hat{\sigma}_j^2 - options$, which are determined by the maximum likelihood method (3):

$$\hat{a}_j=\frac{1}{n}\sum_{i=1}^{n} X_i^j=\hat{X}^j, \hat{\sigma}_j^2=\frac{1}{n}\sum_{i=1}^{n}(X_i^j-\hat{X}^j)^2, \tag{3}$$

Support Vector Machine (SVM). The support vector method algorithm, as one of the popular and most adaptable algorithms with its performance under various conditions, is among the powerful machine learning methods for model building. Usually the support vector machine uses a kernel function of the form:

$$K(x_i, x_j)= e^{-\lambda\sum_{k=1}^{m}(x_{ij}-x_{jk})^2}$$

for splitting the nonlinear decision boundary in classification problems [14], where x_{ij} and $x_{i'j}$ are the *i*-th pair of observations of the *j*-th feature, m is the number of features, λ is a tuning parameter that accounts for the smoothness of the decision boundary, and $K(x_i, x_j)$ is a kernel function.

Random Forest (RF). One of the most advanced and powerful machine learning algorithms, Random Forest is an ensemble learning-based machine learning classification algorithm. This algorithm uses multiple decision trees to perform the classification task. In random forest classification problems, all constituent decision trees are weak learners, the outputs of these weak decision trees are combined, and make the final prediction of all these classifiers as a voting classifier. Random Forest is essentially a tree-based method that has robust and good prediction performance by combining a large number of decision trees to produce a single consistent prediction. The main feature of the random forest is that it is not allowed to consider most of the available features in every partition of the tree [15].

Note that the resulting classifier, such as the classifier a(x), is defined as $a(x)=\frac{1}{N}\sum_{i=1}^{N} b_i(x)$. Simply for the classification problem, we choose the solution by majority vote, and in the regression problem, we choose the mean. Classifiers b_i(x) constructed as a decisive tree generated from a sample X_i using bootstrap. It is recommended to take m=$\sqrt{n}$ in classification tasks, and m= $\frac{n}{3}$ in regression tasks, where n is the number of features.

Gradient Boosting. As a branch of ensemble methods, boosting is a way of combining the performance of a number of weak classifiers to create a powerful "voting classifier", so it is considered a strong classifier. As an example, consider the prediction (4) given by Discrete AdaBoost, which is described as follows:

$$\mathrm{F(x)} = \mathrm{sign}\ (\sum_{m=1}^{M} c_m f_m(x)), \tag{4}$$

where f_m(x) is the weak classifiers, which gives either positive or negative predictions, c_m is the coefficient calculated using training weights, M is the number of weak classifiers, the sign function here returns either positive or negative predictions, and F(x) is the corresponding prediction. By combining multiple models, the boost method can provide better prediction performance than a single model. The basic boosting procedure is to fit a sequence of weak learners (e.g., discriminant analysis, k-nearest neighbors, decision tree, etc.) to weighted versions of the training data. In this paper, three popular boosting algorithms, namely Discrete AdaBoost, LogitBoost, and Gentle AdaBoost, are chosen to investigate the problem.

The data were extracted from the sources of five districts of the Issyk-Kul region of the Kyrgyz Republic and represent the fertilizer (nitrogen, phosphorus and potassium) applied and cultivated fields, temperature, humidity, rainfall, soil acidity, for five names of districts were used, four names of soils (two districts have the same soil data) belonging to each district, and potato yields for each district. For the convenience of working with Python libraries when collecting data, the peculiarities of the data of each region were taken into account and was presented in the form of a generalized .csv file. Pre-processing of yield data, showed that all factors obey the normal law of distribution. Correlation matrix, shown in Figure 1 of the database under study without dummy factors and distributed as follows:

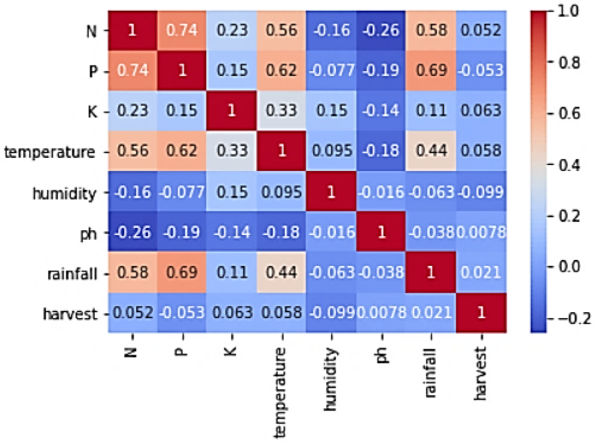


Fig. 1. Results of the correlation matrix of the predicted factor with other variables

Here we can see that the dependent variable harvest – yield – does not correlate with the other independent variables. The dependent variable yield as an initial representation in the form of a multiple regression model has the following form (5):

$$Y=\omega_0+\omega_1 X_1+\omega_2 X_2+\ldots+\omega_p X_p+\varepsilon, \tag{5}$$

where Y is the dependent variable of yield, X_p is independent variables composing the multiple regression, ω_p is unknown coefficients, and ε is the model error.

The algorithm for determining the unknown coefficients ω_i in (5), which are part of this model, can be analytically determined from the minimum of the following quadratic functional (6):

$$L=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\hat{\omega}_0-\hat{\omega}_1 X_{i1}-\hat{\omega}_2 X_{i2}-\ldots-\hat{\omega}_p X_{ip})^2, \tag{6}$$

where $\hat{y}_i$ is the predicted value for the exact value y_i. Here, from the minimum condition of the functional (6) $\partial L/\partial \underline{\omega}=0$, the vector (7) is defined

$$\vec{\omega}=(X^T X)^{-1} X^T \vec{y}, \tag{7}$$

However, the algorithm (7) is unstable to small changes in the input data, i.e. the problem in this case is incorrectly posed. The model created by this algorithm will be retrained. The trained model in this case begins to interpolate the data instead of extrapolating, and it loses its generalizing ability, i.e. generalizing the model on the test data will not work. To solve this problem, we introduce some regularizing function $R(\vec{\omega})$ [16]. Let us now formulate the problem of determining the weighting coefficients $\vec{\omega}$ with the transformed error functional in the form (8):

$$\Omega(X,\vec{y},\vec{\omega})= L(X,\vec{y},\vec{\omega})+\lambda R(\vec{\omega}) \Rightarrow min, \tag{8}$$

where λ is called the regularization coefficient, E is a unit matrix. In the implementation of this problem (8) we usually use (9) as $R(\vec{\omega})$ as for the Ridge problem.

$$R(\vec{\omega})=\frac{1}{2}\left\|\vec{\omega}\right\|_2^2=\frac{1}{2}\vec{\omega}^T\vec{\omega}\ , \tag{9}$$

Then, the algorithm for finding the coefficients $\vec{\omega}$ is defined by the formula

$$\vec{\omega}=(X^TX+\lambda E)^{-1}X^T\vec{y}.$$

3 Discussion

In the calculations performed below for the retrained models, this technique was applied everywhere. Here are the results of applying machine learning algorithms, as components of some ensemble, to the study of regression models, as well as their estimates of accuracy on yield prediction. The resulting yield regression results using machine learning algorithms are shown below in Table 1 and as a visualization of the yield calculations in Figures 2-3. Table 1 shows that the algorithms of gradient boosting with MAPE =10.14% and random forest with MAPE =10.24% gave the best results than the other algorithms. The leader of the prediction was the support vector method with a result of 10.12%.

Below are the results of model building using eight machine learning algorithms to solve yield regression problems. The most proven regression model algorithms for machine learning were used. The mean_squared_error, r2_score, mean_absolute_error, and max_error libraries from the corresponding sklearn.metrics library were used to evaluate the model. We relied on the following classes to analyze the process of regression models:

lin = LinearRegression(),
dtr = DecisionTreeRegressor(),
sgd = SGDRegressor (loss='squared_loss'),
gbr = GradientBoostingRegressor(),
knn = KNeighborsRegressor(n_neighbors=5),
rfr = RandomForestRegressor(),
svr=SVR(), and xgb=XGBRegressor().

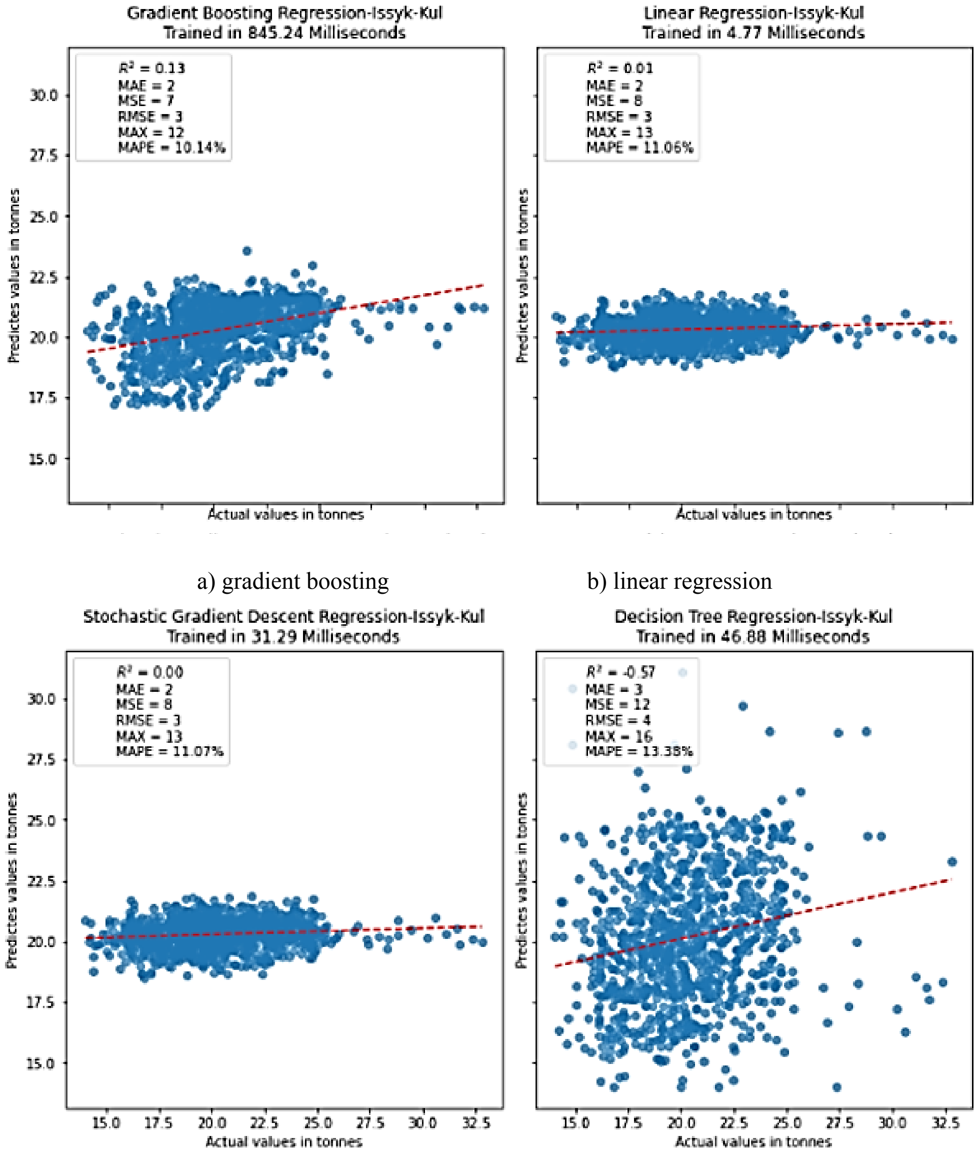
The results of the evaluation of models built with machine learning algorithms are presented in Table 1.

Table 1. Results of model evaluations obtained by machine learning algorithms

№	**Estimates/ML Algorithm**	R^2	**MAE**	**MSE**	**RMSE**	**MAX**	**MAPE in %**
1	Linear Regression	0.01	2	8	3	13	11.06
2	Decision Tree Regression	-0.61	3	12	4	16	13.49
3	Stochostic Gradient Descent Regression	0	2	8	3	13	11.12
4	K – Nearest Neighbour (n_neighbors=5)	0.03	2	8	3	12	10.58
5	SVR	0.10	2	9	3	13	10.12

6	Gradient Boosting Regression	0.12	2	7	3	12	10.14
7	Random Forest Regression	0.11	2	7	12	12	10.19

Figure 2 shows the results of applying machine learning to yield regression problems using the following algorithms: gradient boosting (a), multiple linear regression (b), stochastic gradient descent (c), and decision tree regression (d).



a) gradient boosting

b) linear regression

c) stochastic gradient descent

d) decision tree regression

Fig. 2. Results of applying machine learning to yield regression problems using various algorithms

The remaining results using four random forest machine learning algorithms, XGBoost boosting, K-nearest neighbors, and the reference vector method showed the result – average absolute percentage error MAPE = 10% deviation from the exact prediction. Here is a visualization of the prediction data representation in the form of the following regressions.

Figure 3 shows the results of applying machine learning to the yield regression problems of the following algorithms random forest, k-nearest neighbor in the case of five nearest neighbors, XGBoost boosting, and support vector method.

As can be seen from the results, the intensity of data arrangement around the diagonal function is most densely located in the stochastic gradient descent, random forest, gradient boosting and support vector method. It is known that the more negative the MAE, the better, and a perfect model is available at MAE=0. The estimate of the mean absolute deviation of the MAE in all of our cases, ranges around 2. Shown below are some calculated model performance estimates for neg_mean_absolute_error (NMAE) with regression=5:

$$R^2\text{: } 0.142+0.024,\ 0.142-0.024,\ \text{MAX:}10+0,\ 10-0,$$
$$\text{MAE:}2+0,\ 2-0,\ \text{MSE: } 6+0{,}6-0,\ \text{RMSE: } 3+0{,}3-0$$

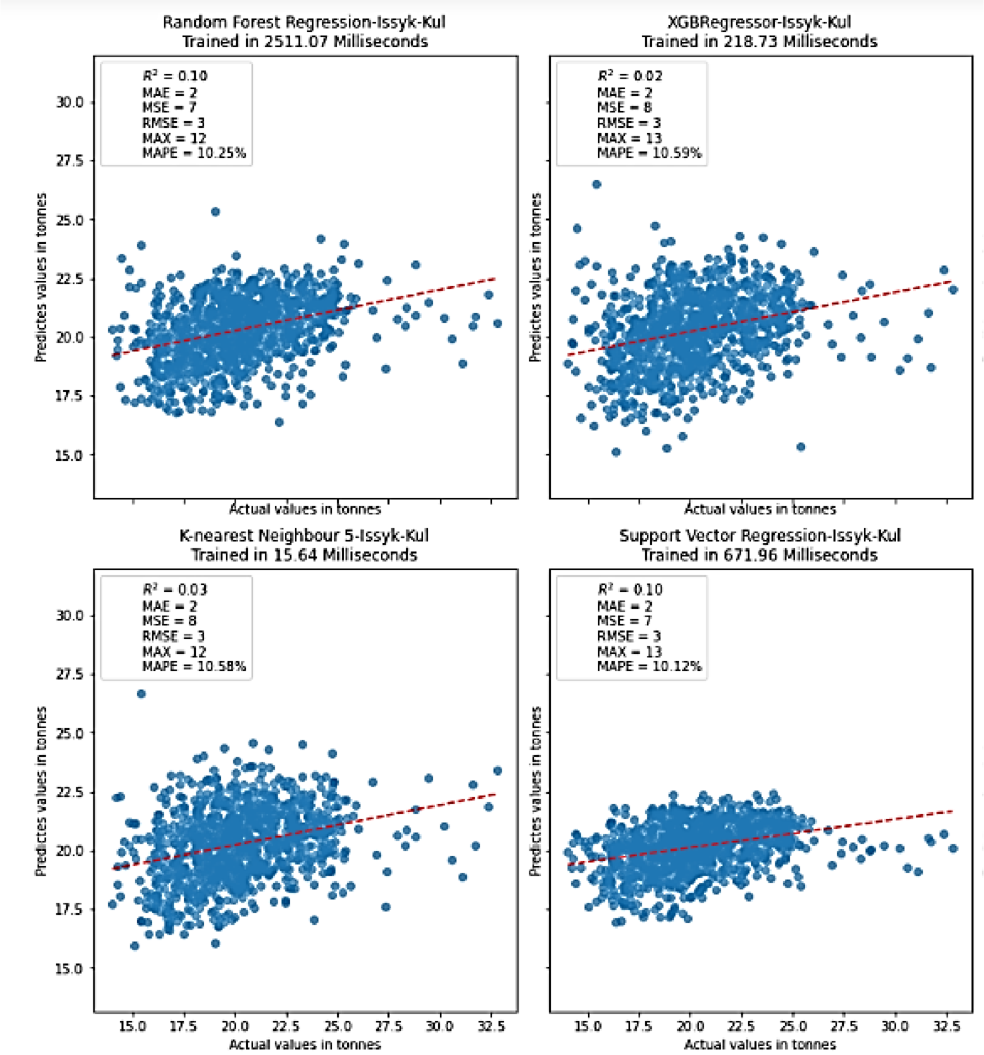


Fig. 3. Results of applying machine learning to yield regression problems of the following algorithms random forest, k-nearest neighbors in the case of 5 nearest neighbors, XGBoost boosting, and support vector method

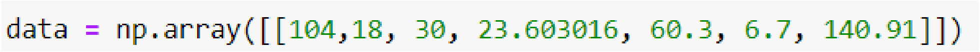
In this case, the ensemble result can give a tangible result for prediction. Next, we used ensemble algorithms, or as it is called Lazy Prediction, for the task of predicting crop yields with multiple, ensemble components. Here it is appropriate to note that the constituent algorithms can be any. As shown below, using this library with a special selection of parameters of machine learning algorithms, we calculated the performance of the 10 most powerful standard classifications or regression models on several performance matrices. Calculations on ensemble algorithms for the regression problems below are computed in Google Colaboratory.

As the next task, the number of fictitious functions – features was expanded to study the features of soil characteristics for these five regions. For the new updated database, expanded district soil data were included. Four cases of the ensemble algorithm were considered. The following ensembles of algorithms and VotingRegressor results were implemented to analyze the voting algorithm, which gave the following results, as shown in Table 2:

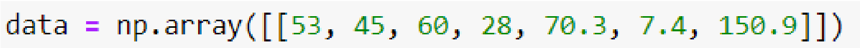
Table 2. Ensemble method results

№	Estimates/ML Algorithm	R^2	MAE	MSE
1	RandomForestRegressor AdaBoostRegressor VotingRegressor	-3.200329	1.981150	6.039916
2	AdaBoost GradientBoostingRegressor VotingRegressor	-3.798058	1.9922521	6.11885
3	Random Forest and GradientBoostingRegressor VotingRegressor	-2.21814	1.9646782	6.03540
4	Random Forest AdaBoost and GradientBoostingRegressor VotingRegressor	-2.96578	1.971011	5.98589

Except for the latter case, where there are more constituent algorithms of the voting algorithm, the results turned out to be close to each other. Using this database for advisory purposes, a web-based system has been created for farmers to study, which predicts from climatic data and relevant soil characteristics of the area, what to grow and what crops are essential for the region. The results showed that machine learning algorithms with test data from a particular area, specifying the relevant parameters of the region, give an appropriate prediction. The next two machine learning algorithms, gradient boosting (XGBoost) and Random Forest, performed well in the tests. For example, the following test data was used for the selected area:



```
data = np.array([[104,18, 30, 23.603016, 60.3, 6.7, 140.91]])
```

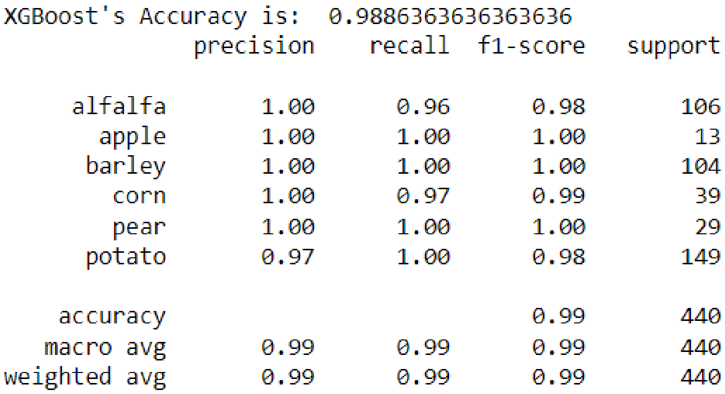


```
data = np.array([[53, 45, 60, 28, 70.3, 7.4, 150.9]])
```

,

where the Numpy arrays contain data reflecting the weather conditions and soil characteristics of the area, the result of the calculation in both cases showed the forecast – to plant potatoes (potato). Table 3 below shows the results of calculations using two algorithms for this problem.

Table 3. Prediction result with the gradient boosting algorithm

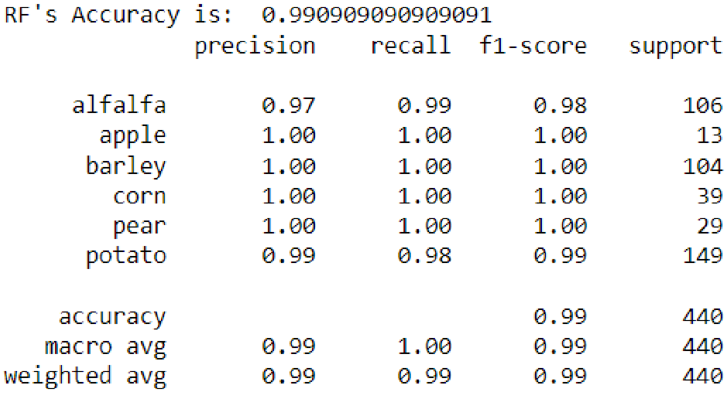


```
XGBoost's Accuracy is:  0.9886363636363636
              precision    recall  f1-score   support

     alfalfa       1.00      0.96      0.98       106
       apple       1.00      1.00      1.00        13
      barley       1.00      1.00      1.00       104
        corn       1.00      0.97      0.99        39
        pear       1.00      1.00      1.00        29
      potato       0.97      1.00      0.98       149

    accuracy                           0.99       440
   macro avg       0.99      0.99      0.99       440
weighted avg       0.99      0.99      0.99       440
```

Table 4. Prediction result with the random forest algorithm

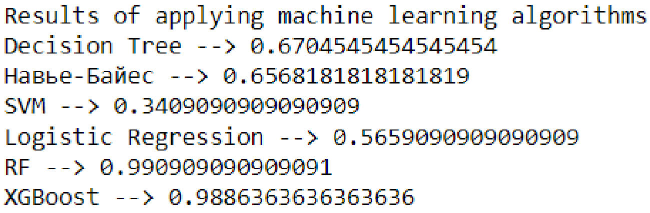


```
RF's Accuracy is:  0.990909090909091
              precision    recall  f1-score   support

     alfalfa       0.97      0.99      0.98       106
       apple       1.00      1.00      1.00        13
      barley       1.00      1.00      1.00       104
        corn       1.00      1.00      1.00        39
        pear       1.00      1.00      1.00        29
      potato       0.99      0.98      0.99       149

    accuracy                           0.99       440
   macro avg       0.99      1.00      0.99       440
weighted avg       0.99      0.99      0.99       440
```

Now let's compare the accuracy of the models and the results of calculations with other machine learning algorithms.



```
Results of applying machine learning algorithms
Decision Tree --> 0.6704545454545454
Навье-Байес --> 0.6568181818181819
SVM --> 0.3409090909090909
Logistic Regression --> 0.5659090909090909
RF --> 0.990909090909091
XGBoost --> 0.9886363636363636
```

A comparative analysis of the accuracy of machine learning algorithms. The accuracy diagram of the algorithm models is shown in Figure 4.

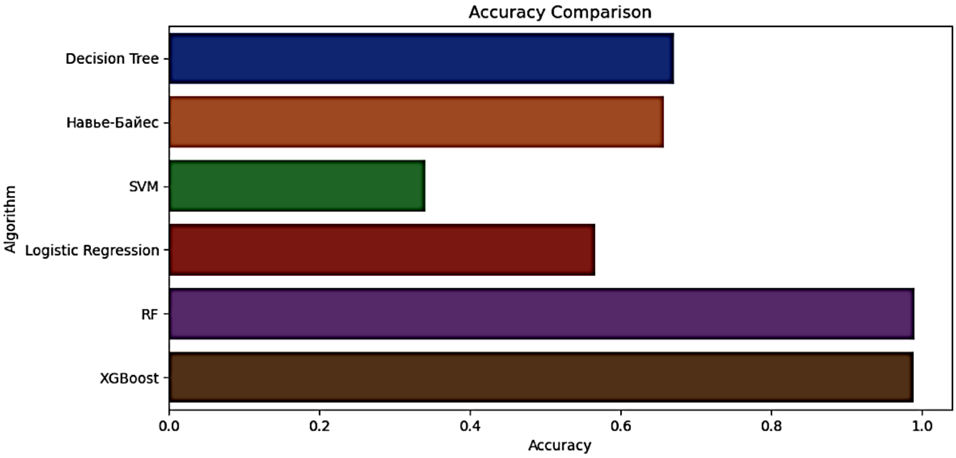


Fig. 4. Machine learning algorithm comparison chart

As can be seen from Figure 4, the highest prediction result of the random forest algorithm with a prediction result of 0.990909 (99.09%).

The paper shows the effectiveness of using machine learning algorithms for modeling and forecasting agricultural problems. Research on data from five districts of the Issyk-Kul region on the yield of potatoes and other crops, using machine learning, showed the effectiveness of identifying the main characteristics to build models and forecasts. In the databases collected on the yield of potatoes and other plants about 5 thousand subjects – farmers in Issyk-Kul region were taken into account important factors of the use of weather conditions in these regions, as well as taking into account the indicators of acidity, which in the selected regions ranges mainly from 7 to 7.5 and other characteristics of the soil. The temperature regime for these five regions, due to the influence of Lake Issyk-Kul, fluctuates around the average monthly temperature by a value of plus minus 4-5 degrees. For the indicators of precipitation and humidity, the data were taken into account according to the meteorological observations of the republic. Calculations based on machine learning algorithms gave the results of the deviation of the exact and predictive data by the values based on Table 1. Figures 2-3 show the results of applying machine learning algorithms to regression problems in the form of a graphical representation, the potato yield indicator for the region as a whole, with different accuracies. Figure 4 shows the performance of the XGBoost algorithm model. For yield prediction the best performance was obtained, accuracy score with average absolute error MAPE = 10.2% for the random forest algorithm. In this case, the intensity of data accumulation around the regression line, as seen in Figures 2-3, is the greatest. It can be noticed that other algorithms like, reference vector method, gradient binning, nearest neighbor method are close with MAPE values to the result of random forest. The paper also studied ensembles of algorithms using the Lazy Prediction library. Table 2 shows the results of applying the ensemble method to yield prediction problems. The top ten models were selected and results were obtained. Table 3-4 shows the results of applying machine learning algorithms to identify the most acceptable crops according to certain test data, according to the localization of the crop area and soil properties. The results of the evaluation of VotingRegressor ensemble algorithms for model building using Random Forest, AdaBoost, and GradientBoostingRegressor for an extended database that takes into account localization of areas and soil characteristics are obtained.

4 Conclusions

For a set of data collected from five districts of the Issyk-Kul region, such as weather conditions, soil characteristics and pre-treatment of the sowing area, in the work, the task of predicting the yield of various crops using advanced machine learning algorithms, as the method of reference vectors, to – nearest neighbors, options of gradient boosting and random forest was studied. For comparative analysis, model accuracy estimates were compared with multiple regression results. The effectiveness of algorithms for solving the problems of yield forecasting is shown. The calculation results showed the efficiency of the used algorithms using ensemble methods. At the same time, the requirements for parameter selection were removed for the set of constituent ensemble algorithms. This study has also shown that the developed models can also be successfully used to predict yield for multiple plants. What has not been realized is the application of deep neural networks to this particular case. The accuracy of the raw data and the nature of the choice of data collected should most fully and realistically reflect the environment of the crop being grown, which in general, is a costly way to collect big data. The convenience of using machine learning is identifying some of the hidden complex and non-linear features of the factors affecting yields for the selected region. Further research for the selected regions is the collection of data with deeper factors such as minimum and maximum temperature, daylight hours, soil characteristics, soil composition of the region and the use of data from seed companies to create new crop varieties to predict future yields. The issue of climate change, the risks associated with debris flows, hail warnings through weather forecasts, which are not uncommon for this region of the phenomenon, increased solar activity, the duration of abnormally hot days and low water levels remain untouched. All of these factors lead to soil erosion and weathering. According to the authors [17,18,19,20,21], these studies in the future should use deep learning technologies for big data with local conditions. In general, the calculations and results obtained in the work showed the completeness of the research using machine learning algorithms for the tasks of modeling and predicting yields using region-specific data.

The authors state that this study was conducted without any assistance from funded parties and does not represent a commercial or financial relationship that could be interpreted as a conflict of interest.

References

1. J. Romero, P. Roncallo, P. Akkiraju, I. Ponzoni, V. Echenique, J. Carballido, Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires Comput. Electron. Agric., **96** (2013), pp. 173-179, https://doi.org/10.1016/j.compag.2013.05.006
2. M. Paul, S. Vishwakarma, A. Verma, Analysis of soil behaviour and prediction of crop yield using data mining approach. In: 2015 International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, pp. 766–771. https://doi.org/10.1109/CICN.2015.156.
3. J. Jeong, J. Resop, N. Mueller, D. Fleisher, K. Yun, E. Butler, S. Kim, Random forests for global and regional crop yield predictions. PLoS ONE, **11** (6) (2016), https://doi.org/10.1371/journal.pone.0156571
4. N. Gandhi, O. Petkar, L. Armstrong, A. Tripathy, Rice crop yield prediction in India using support vector machines. In: 2016 13th International Joint Conference on

Computer Science and Software Engineering, JCSSE 2016. https://doi.org/10.1109/JCSSE.2016.7748856

5. N. Gandhi, L., Armstrong,. Applying data mining techniques to predict yield of rice in humid subtropical climatic zone of India. In: Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016, 1901–1906. Retrieved from https://ieeexplore.ieee.org/abstract/document/7724597/
6. R. Sujatha, P. Isakki, A study on crop yield forecasting using classification techniques. In: 2016 International Conference on Computing Technologies and Intelligent Data Engineering, ICCTIDE 2016. https://doi.org/10.1109/ICCTIDE.2016.7725357
7. H. Cheng, L. Damerow, Y. Sun, M. Blanke, Early yield prediction using image analysis of apple fruit and tree canopy features with neural networks J. Imag., **3** (1) (2017), p. 6, https://doi.org/10.3390/jimaging3010006
8. S. Bargoti, J. Underwood, Image segmentation for fruit detection and yield estimation in apple orchards J. Field Rob., **34** (6) (2017), pp. 1039-1060, https://doi.org/10.1002/rob.21699
9. A. Shekoofa, Y. Emam, N. Shekoufa, M. Ebrahimi, E. Ebrahimie, Determining the Most Important Physiological and Agronomic Traits Contributing to Maize Grain Yield through Machine Learning Algorithms: A New Avenue in Intelligent Agriculture. PLoS ONE **9**(5): e97288. https://doi.org/10.1371/journal.pone.0097288
10. G. Brown, (2017). “Ensemble learning” in Encyclopedia of Machine Learning and Data Mining. Ed. Sammut, K., Webb, G.I. (Boston, MA: Springer, USA), 393–402.
11. L. Breiman, Bagging predictors. *Machine Learning* **24,** 123–140 (1996). https://doi.org/10.1007/BF00058655
12. L. Breiman, Random Forests. *Machine Learning* **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
13. You, J., Li, X., Low, M., Lobell, D., and Ermon, S. (2017). “Deep gaussian process for crop yield prediction based on remote sensing data,” in Thirty-First AAAI Conference on Artificial Intelligence (San Francisco, CA), 4559–4566.
14. G. James, D. Witten, T. Hastie, R. Tibshirani, Introduction to Statistical Learning, Vol. **112**, (2013) (New York Heidelberg Dordrecht London: Springer). https://doi.org/10.1007/978-1-4614-7138-7
15. Friedman, J., Hastie, T., Tibshirani. R. The Elements of Statistical Learning Springer Series in Statistics, New York (2001). Google Scholar
16. James, G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning Springer, New York (2013). Google Scholar
17. Khaki S and Wang L (2019) Crop Yield Prediction Using Deep Neural Networks. Front. Plant Sci. **10**:621.https://doi.org/10.3389/fpls.2019.00621
18. Taherei-Ghazvinei P., Hassanpour-Darvishi H., Mosavi A., Yusof K.W., Alizamir M., Shamshirband S., Chau K. Sugarcane growth prediction based on meteorological parameters using extreme learning machine and artificial neural network Eng. Appl. Comput. Fluid Mech., **12** (1) (2018), pp. 738-749, https://doi.org/10.1080/19942060.2018.1526119
19. M. Villanueva, M. Louella, M. Salenga, Bitter Melon Crop Yield Prediction using Machine Learning Algorithm. IJACSA) International Journal of Advanced Computer Science and Applications, Vol. **9**. (2018) Retrieved from www.ijacsa.thesai.org.
20. Gandhi, N., Petkar, O., Armstrong, L.J., Tripathy, A.K., Rice crop yield prediction in India using support vector machines. In: 2016 13th International Joint Conference on Computer Science and Software Engineering, JCSSE 2016.

21. Abhishek, K., Singh, M., Ghosh, S., and Anand, A. (2012). Weather forecasting model using artificial neural network. Procedia Technol. **4**, 311–318. https://doi.org/10.1016/j.protcy.2012.05.047