BIO Web of Conferences **130**, 01011 (2024) https://doi.org/10.1051/bioconf/202413001011
BFT-2024

Crop yield forecasting using neural networks trained on the basis of agrometeorological and agrochemical data

Ksenia Degtyareva[1,2*], *Vadim* Tynchenko[1,2], *Nikita* Stepanov[1], *Ekaterina* Kalmykova[1], and *Darya* Makarevskaya[1]

[1]Reshetnev Siberian State University of Science and Technology, 660037 Krasnoyarsk, Russia
[2]Bauman Moscow State Technical University, 105005 Moscow, Russia

Abstract. In this study, a neural network model was developed and investigated for predicting crop yields based on data on weather conditions, the use of fertilizers and the content of basic nutrients in the soil (nitrogen, phosphorus and potassium). The research is based on the use of a multilayer perceptron architecture with Rely activation functions for hidden layers and linear activation for the output layer. The evaluation of the model quality was carried out using the mean square error (MSE), which was 0.5783 in the test sample, demonstrating high accuracy of predictions. Visualization of the results included analysis of scatter plots, residuals, histograms of residuals and comparison of distributions of actual and predicted values. The results obtained confirm the effectiveness of the proposed model for yield forecasting tasks, which makes it a valuable tool for optimizing agricultural production.

1 Introduction

Agriculture is the main source of food for mankind, and its successful functioning directly depends on effective resource management and accurate forecasting of yields. Crop yields can be subject to significant fluctuations due to a variety of factors such as weather conditions, soil conditions, agrotechnical measures and biotic stressors (pests and plant diseases). The unpredictability of these factors creates the need to develop methods capable of predicting yields with high accuracy, which, in turn, contributes to a more rational use of resources, reducing risks and increasing the sustainability of agricultural production [1-4].

One of the promising approaches to solving this problem is the use of machine learning methods and neural networks. These methods allow you to analyze a large amount of data and identify hidden dependencies between various factors affecting productivity [5, 6]. In particular, artificial neural networks (ANS) have proven to be a powerful tool for modeling complex nonlinear dependencies in data [7-10].

The purpose of this work is to develop and study a neural network model for predicting crop yields based on data on weather conditions, the use of fertilizers and the content of basic nutrients in the soil (nitrogen, phosphorus and potassium).

* Corresponding author: sofaglu2000@mail.ru

2 Research methods

Neural networks (NS) are computational models inspired by biological neurons that are used to solve various machine learning tasks, including classification, regression, and clustering. Neural networks consist of many interconnected artificial neurons organized into layers [11-13]. The main components of a neural network are an input layer, one or more hidden layers, and an output layer [14].

The activation function is used to introduce non-linearity into the model, which allows the neural network to model complex dependencies [15-17]. In this study, ReLU (Rectified Linear Unit) activation functions and linear activation for the output layer were used. The loss function evaluates the quality of the model's predictions [18]. The mean square error (MSE) function was used in this study. The optimization algorithm is used to adjust the weights of the neural network in order to minimize the loss function. In this case, the Adam algorithm was used [19-22].

First, the model architecture was selected. A multilayer perceptron neural network (MLP) with three hidden layers was chosen to predict yield. The number of neurons in the hidden layers is 64, 32 and 16, respectively, were selected experimentally to achieve the best results.

The next stage involved training the model. The model was trained on a training sample using the Adam optimization algorithm and the MSE loss function. The learning process involved setting up hyperparameters such as the number of epochs and the batch size.

Next, the model was evaluated. The quality of the model was evaluated on a test sample using the MSE metric. The results were also visualized for a more detailed analysis of the model's operation.

The dataset includes information on the number of fertilizers applied per unit area, temperature conditions during the growing period, nitrogen, phosphorus and potassium content in fertilizers or soil, measured in pounds per acre, as well as crop productivity, expressed in the amount of crop harvested per acre.

3 Results

The average quadratic error (MSE) was chosen as the main metric for evaluating the quality of the model [23]. The MSE value in the test sample was 0.5783, which indicates a fairly high accuracy of the model in predicting crop yields.

The scatter plot helps to visually compare how well the model predicts real values. If the model works well, the points on the graph should be located close to the perfect match line (diagonal line). The scattering diagram is shown in Figure 1.

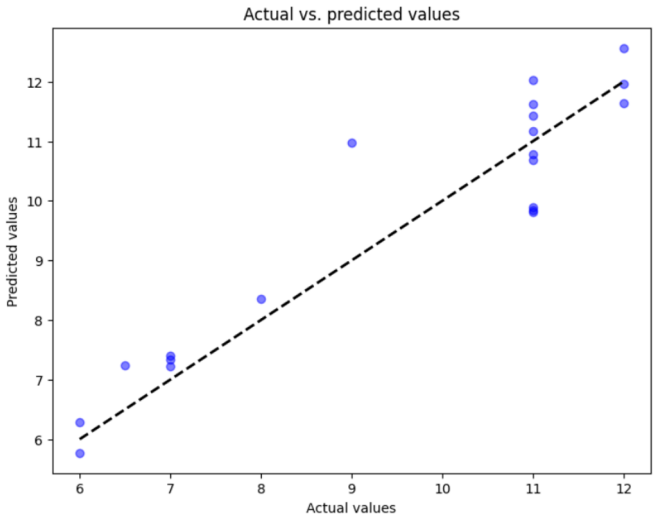

Fig. 1. Scatter plot.

The remainder graph is a scatter plot that shows the residuals (the differences between the actual and predicted values) on the vertical axis and the actual values on the horizontal axis [24, 25]. It is shown in Figure 2.

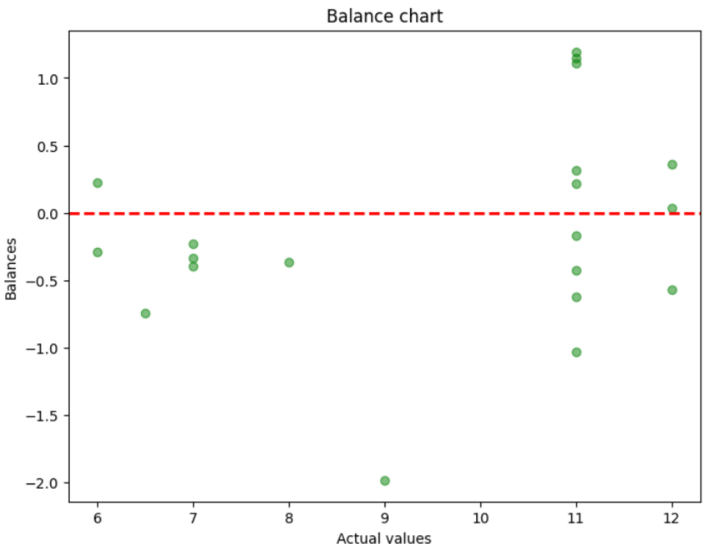

Fig. 2. Remainder graph.

Most of the residuals are concentrated in the range from -1 to 1, which indicates that most of the model's predictions are fairly accurate. However, there are significant outliers, especially noticeable at the bottom of the graph (about -2). There is no obvious trend or pattern in the distribution of residuals, which is good. This indicates that the model does not systematically overestimate or underestimate the values.

The histogram of the residuals shows the distribution of model errors (the difference between the actual values and the predicted values) [26-28]. It is shown in Figure 3.

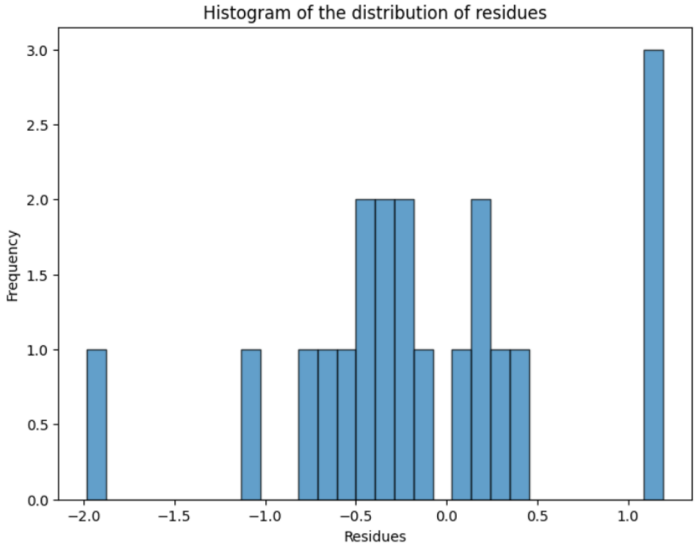

Fig. 3. Histogram of the residuals.

Most of the residuals are in the range from -1 to 1, which suggests that the model is generally good at predicting values, however, there are several significant errors (outliers) in the range from -2 to 1. Several values of the residuals fall outside the range from -1 to 1, which may indicate the presence of outliers. These outliers may be the result of data that the model failed to correctly predict, or data that does not match the overall trend in the dataset. The highest density of residues is observed in the range from -0.5 to 0, which indicates the tendency of the model to underestimate predictions in this range.

Comparing the density of the distribution of the actual and predicted values will help to assess how much the distributions of the predicted values coincide with the actual values [29-31]. The graph is shown in Figure 4.

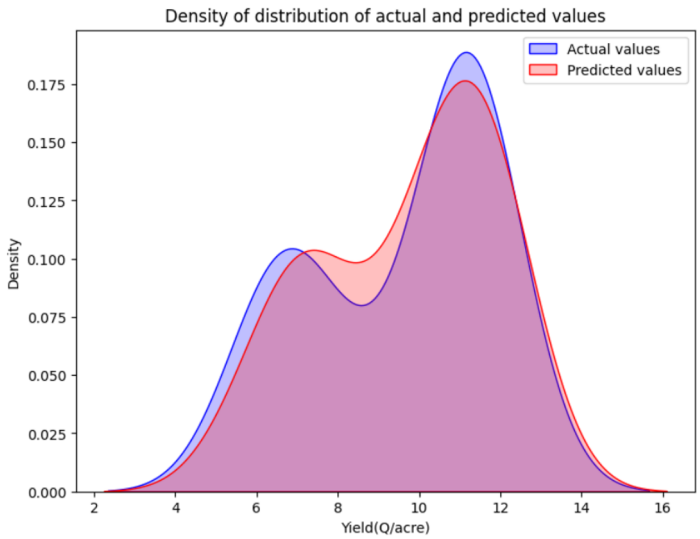

Fig. 4. Density of the distribution of the actual and predicted values.

It can be seen from the graph that the actual values of the indicator have a higher distribution density in the range of values above the predicted ones. This means that the model on the basis of which the forecasts were made underestimates the real values of the indicator. In other words, forecasts tend to be lower than the actual values.

Box plot helps to visualize the distribution of data and identify the presence of outliers [32, 33]. Figure 5 shows the distribution of actual and predicted values.

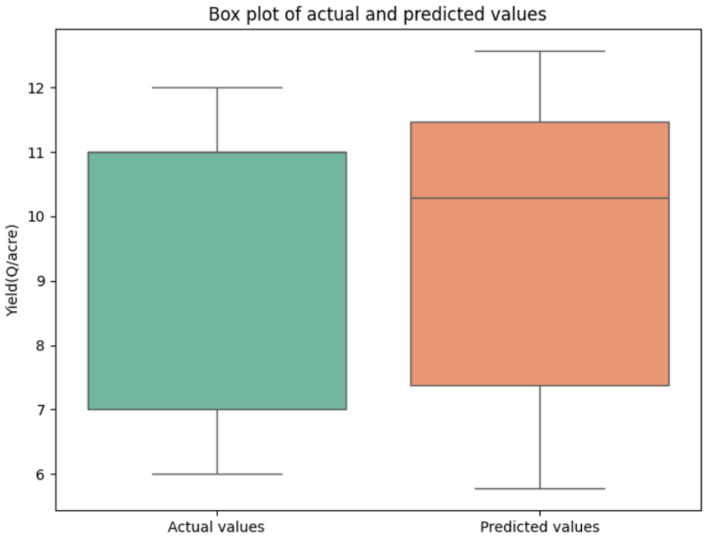

Fig. 5. Box plot.

The median of the actual yield values is higher than the median of the predicted values. This means that, on average, the actual harvest was higher than predicted. The range of the actual yield values (interquartile range) is greater than the range of the predicted values. This means that the actual harvest was more variable than predicted. There are outliers in both samples. However, there are more outliers in the sample of actual values than in the sample of predicted values.

The graph of the learning curve shows how the error changes on the training and test datasets depending on the number of training epochs [34, 35]. This helps to understand whether the model is under-trained or over-trained. The graph is shown in Figure 6.

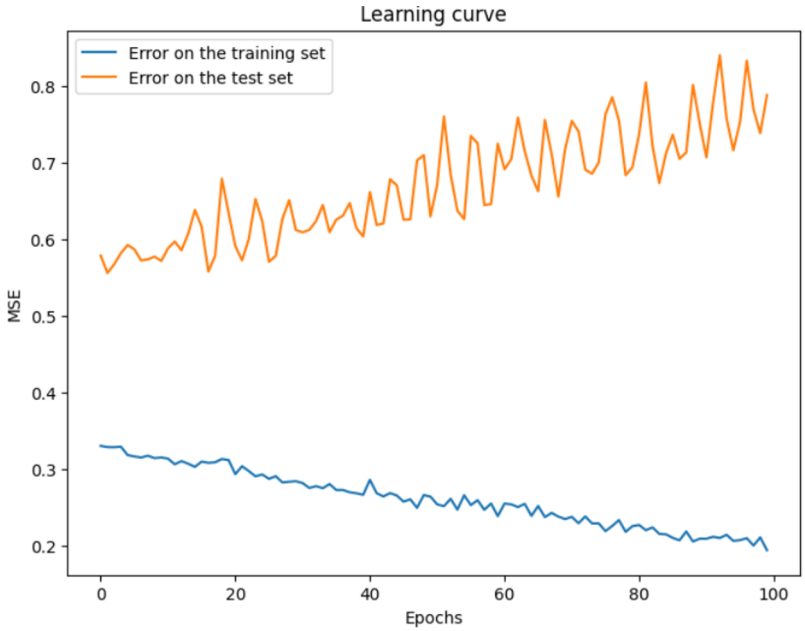

Fig. 6. Learning curve.

The learning error decreases rapidly during the first 20 epochs. This means that the model learns quickly from the training data. Validation error also decreases, but more slowly than learning error. This means that the model begins to retrain on the training data. The validation error reaches a minimum at the 40th epoch, and then begins to grow. This means that the 40th epoch is optimal for this model. In general, this graph indicates that the model is well trained.

4 Conclusion

In this study, a neural network model was developed and tested to predict crop yields based on data on weather conditions, fertilizers and the content of nutrients in the soil. The main goal was to create an effective tool for agronomists and farmers, which would make it possible to more accurately predict yields and, consequently, optimize the processes of growing crops [36].

The neural network model showed good results, which is confirmed by the low value of the mean square error (MSE) in the test sample. The use of a multilayer perceptron architecture with ReLU activation functions and linear activation for the output layer made it possible to create a sufficiently powerful and accurate model. Visualization of the results, including graphs of predicted versus actual values, residue distribution, and distribution density, showed that the model is able to predict crop yields fairly accurately [37, 38].

In general, the developed neural network model has demonstrated its effectiveness in predicting crop yields. It represents a significant step forward in the use of artificial intelligence technologies in agronomy. Further research and model improvements can lead to even more accurate and useful tools for the agricultural sector.

References

1. A. Gladkov, et al., *Development of Requirements for AIS Aimed at Controlling High Turnover.* 2023 IEEE International Conference on Computing (ICOCO). IEEE (2023)
2. Ya. Zhilkina, et al., *Strategy of introduction of information system in trade and logistics company.* E3S Web of Conferences **458** (2023)
3. V.V. Kukartsev, et al., *Advancements in network-based management systems for enhanced business services.* E3S Web of Conferences **460** 92023)
4. A. Kozlova, et al., *Finding dependencies in the corporate environment using data mining.* E3S Web of Conferences **431** (2023)
5. V.V. Kukartsev, et al., *Control system for personnel, fuel and boilers in the boiler house.* E3S Web of Conferences **458** (2023)
6. K.A. Bashmur, et al., Sustainability **14.20:** 13083 (2022)
7. O.A. Kolenchukov, et al., Energies **15.22:** 8346 (2022)
8. Ya.A. Tynchenko, et al., Sustainable Development of Mountain Territories **16.1:** 56-69 (2024)
9. V. Kukartsev, et al., Sustainable Development of Mountain Territories **15.3:** 784-797 (2023)
10. V. Brigida, et al., Resources **13.2:** 33 (2024)
11. A.A. Sokolov, MIAB **11.1:** 278-291 (2023)
12. K. Degtyareva, D.A. Ageev, V.V. Kukartsev. *Finding patterns in employee attrition rates using self-organizing Kohonen maps and decision trees.* 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES). IEEE (2023)
13. B.V. Malozyomov, et al., Energies **16.9**: 3909 (2023)
14. D.M. Strateichuk, et al., Crystals **13.5:** 825 (2023)
15. N.V. Martyushev, et al., Energies **16.2:** 729 (2023)
16. V.A. Rezanov, et al., Metals **12.12:** 2135 (2022)
17. V.A. Kukartsev, et al., Metals **13.2:** 337 (2023)
18. N.V. Martyushev, et al., Materials **16.9:** 3490 (2023)
19. V. Kukartsev, et al., *Intelligent Data Analysis as a Method of Determining the Influence of Various Factors on the Level of Customer Satisfaction of the Company.* Proceedings of the Computational Methods in Systems and Software. Cham: Springer Nature Switzerland, 109-128 (2023)
20. K. Degtyareva, et al., *Data analysis using neural networks and Kohonen maps in a comparative perspective.* 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES). IEEE (2023)
21. V. Nelyub, et al., *Machine learning to identify key success indicators.* E3S Web of Conferences **431** (2023)
22. A. Borodulin, et al., *Using machine learning algorithms to solve data classification problems using multi-attribute dataset.* BIO Web of Conferences **84** (2024)
23. V. Kukartsev, et al., *Using digital twins to create an inventory management system.* E3S Web of Conferences **431** (2023)
24. V. Kukartsev, S.A. Zamolockiy, V.V. Khramkov. News of the Tula state university. Sciences of Earth **3:** 101-111 (2023)

25. I.I. Bosikov, et al., Fire **6.3:** 95 (2023)
26. V. Vasileva, et al., *Integration of automated information systems and architectural solutions in industrial enterprises.* E3S Web of Conferences **458** (2023)
27. A. Gladkov, et al., *Development of an automation system for personnel monitoring and control of ordered products.* E3S Web of Conferences **458** (2023)
28. V. Orlov, et al., *Designing an information system to automate service management at the enterprise.* E3S Web of Conferences **458** (2023)
29. O. Kolenchukov, *Forecasting the technical condition of thermochemical reactor systems.* SOCAR Proceedings **1** (2023)
30. B.V. Malozyomov, et al., Energies **16.11:** 4276 (2023)
31. B.V. Malozyomov, et al., Micromachines **14.7:** 1288 (2023)
32. V.O. Gutarevich, et al., Applied Sciences **13.8:** 4671 (2023)
33. V.B. Zaalishvili, et al., Geosciences **14.4:** 102 (2024)
34. R.V. Klyuev et al., Mining informational and analytical bulletin **5**: 144-157 (2024)
35. V.V. Tynchenko, et al., Mathematics **12.2:** 276 (2024)
36. V.V. Kukartsev et al., *Application of non-parametric learning method in soil suitability assessment in present day economy.* Journal of Infrastructure, Policy and Development **8** (2024).
37. K. Degtyareva, et al., *Analyzing Credit Card Defaulters: A Comparative Study Using Kohonen Maps, Neural Networks, and Decision Trees.* 2023 International Conference on Information Technology and Computing (ICITCOM). IEEE (2023)
38. A.S. Borodulin, et al., *Analyzing Data by Applying Neural Networks to Identify Patterns in the Data.* Proceedings of the Computational Methods in Systems and Software. Cham: Springer Nature Switzerland, 99-108 (2023)