APPLICATION AND ANALYSIS OF ENSEMBLE ALGORITHMS IN SOLVING REGRESSION PROBLEMS

Khojiakbar Abdulkhakimov, Nodir Rakhimov, Dilmurod Khasanov and Oybek Primqulov

Department of Software of Information Technologies, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

ABSTRACT

Accurate crop yield prediction is critical for enhancing food security, particularly in agrarian economies prone to soil degradation and climatic uncertainties. This study explores the application of Support Vector Regression (SVR) for forecasting wheat yields in Uzbekistan, utilizing soil fertility indicators as key predictive features. Unlike conventional linear regression models, SVR effectively captures complex non-linear interactions between soil physicochemical properties and crop productivity, thereby offering improved adaptability to real-world agricultural conditions. The dataset comprises essential soil attributes, including nitrogen (N), phosphorus (P), potassium (K), pH, organic carbon (OC), electrical conductivity (EC), and micro-nutrient concentrations. Data preprocessing involved feature standardization, K-nearest neighbor (KNN) imputation for handling missing values, and correlation analysis to select the most influential variables. The dataset was partitioned using an 80/20 stratified split, and the SVR model with a radial basis function (RBF) kernel was optimized through 5-fold cross-validation and exhaustive grid search for hyperparameter tuning. The optimized SVR model achieved a coefficient of determination (R^2) of 0.87 and demonstrated a low root mean square error (RMSE), outperforming baseline regression methods. Model interpretability was enhanced using SHapley Additive exPlanations (SHAP), which identified soil pH, organic carbon, and available phosphorus as the most significant predictors of wheat yield—findings consistent with established agronomic principles. Overall, the results confirm SVR's potential as a robust, scalable, and interpretable tool for precision agriculture, offering practical insights for site-specific yield forecasting and promoting sustainable land management practices in Uzbekistan.

KEYWORDS

Support Vector Regression (SVR), Wheat Yield Prediction, Precision Agriculture, Machine Learning, Non-linear Models, Uzbekistan, Model Evaluation.

1. INTRODUCTION

Globally, over 40% of the Earth's land surface is already classified as degraded, posing a profound threat to agricultural productivity and global food security. Soil fertility—the inherent capacity of soil to supply essential nutrients and sustain healthy plant growth—is a critical determinant of sustainable agricultural systems. Soil fertility is shaped by a range of physicochemical and biological factors, including the availability of key macronutrients (notably nitrogen, phosphorus, and potassium), soil pH, organic matter content, microbial biomass, and soil structural stability.

However, modern agricultural practices have exacerbated stress on these natural systems. Activities such as monocropping, intensive tillage, and the over-application of chemical fertilizers—which alone accounted for approximately 195 million metric tons in 2021—have

accelerated nutrient depletion, soil acidification, and environmental degradation at a global scale. Accurately and promptly assessing soil fertility is fundamental for a variety of agronomic decision-making processes, including site-specific fertilizer application, crop selection, and pre-season yield forecasting. Empirical evidence from major cereal-producing regions attributes over 40% of yield variability directly to variations in soil nutrient status, exceeding the impacts of cultivar choice or pest management interventions. Despite its critical importance, traditional soil fertility assessment methods—relying primarily on manual sampling and laboratory-based chemical analysis—remain time-consuming, costly, and spatially restricted. These limitations are particularly pronounced in smallholder farming systems and resource-constrained environments, where access to laboratory infrastructure is limited. In response, artificial intelligence (AI) and machine learning (ML) technologies have gained attention as scalable alternatives. ML algorithms trained on enriched datasets—combining in-situ laboratory measurements, remote sensing observations, and geo-referenced soil surveys—have demonstrated remarkable potential in modeling soil fertility. However, a majority of these studies focus on classification approaches, discretizing fertility into categorical classes. In contrast, this study investigates regression-based machine learning methods for continuous soil fertility prediction, leveraging open-access agricultural datasets. Our objective is to evaluate whether regression models trained on publicly available, heterogeneous datasets can deliver sufficiently accurate predictions to inform data-driven land management strategies, particularly in smallholder and low-resource settings. By addressing this challenge, the study aims to lower technical barriers and expand the accessibility of precision agriculture technologies across the Global South.

2. LITERATURE REVIEW

In recent years, machine learning (ML) has emerged as a transformative tool in soil science, enabling advanced modeling of soil nutrient concentrations, pH levels, organic matter content, and other key soil properties. Regression-based ML models, in particular, have demonstrated strong capabilities in predicting continuous variables such as nitrogen (N), phosphorus (P), and potassium (K) concentrations, offering valuable insights into soil fertility dynamics. Despite their technical success, the direct practical utility of these raw numerical predictions for agronomists and farmers remains somewhat limited. Interpreting continuous outputs often requires domain-specific expertise, and numerical estimates alone do not always provide actionable guidance for decisions related to fertilization strategies, soil amendments, or crop selection. To bridge this gap, initiatives like the Decision Support for Sustainable Land Management (DS-SLM) framework—developed under the guidance of the FAO—promote simplifying complex soil metrics into more user-friendly and practical categories, such as fertility levels labeled as low, medium, or high. These simplified outputs enhance accessibility for smallholder advisory systems, where technical resources and agronomic expertise may be scarce. However, only a small portion of research has methodically employed advanced machine learning methods—such as Support Vector Regression (SVR), gradient-boosted models like XGBoost, and other non-linear approaches—for predicting soil fertility based on publicly available datasets. Many earlier efforts relied predominantly on linear models or basic decision trees, which often struggled to capture the non-linear relationships inherent in complex soil processes.

Moreover, few studies have integrated explainability frameworks such as SHapley Additive exPlanations (SHAP) to validate that model outputs align with established agronomic knowledge. Ensuring model transparency is particularly critical for fostering trust and adoption among stakeholders in precision agriculture.

This study addresses these research gaps by evaluating the effectiveness of regression-based ML models for soil fertility prediction using a publicly available, heterogeneous soil dataset. In addition to assessing predictive accuracy, we employ SHAP analysis to enhance interpretability

and explore the extent to which modeled fertility indices correlate with observed yield outcomes. Ultimately, this work seeks to contribute to the development of scalable, interpretable decision-support tools for sustainable land management in data-scarce environments.

3. METHODOLGY

This study employs a structured, multi-stage methodology to predict soil fertility indices as continuous variables using supervised machine learning regression techniques. The approach integrates open-access agronomic data, rigorous preprocessing, feature engineering, model tuning, and interpretability analysis, aiming to validate the practical feasibility of ML-based models in real-world agricultural advisory systems.

Dataset Description

The dataset utilized in this study was obtained from the "Soil Classification" repository available on Kaggle, consisting of 33,497 samples collected from diverse agro-ecological regions. Each sample is described by 22 physico-chemical soil attributes, encompassing both macronutrients, micronutrients, and environmental indicators. The key features include:

1. Macronutrients: Nitrogen (N), Phosphorus (P), Potassium (K)
2. Micronutrients: Sulfur (S), Zinc (Zn), Iron (Fe), Copper (Cu), Manganese (Mn), Boron (B)
3. Environmental indicators: pH, Electrical Conductivity (EC), Organic Carbon (OC), Cation Exchange Capacity (CEC)

Each instance was associated with a continuous fertility output variable, reflecting a composite soil fertility score derived based on agronomic thresholds.

Data quality was high, with missing values accounting for less than 0.5% of the total records. Missing entries were imputed using K-Nearest Neighbors (KNN) imputation, which preserves local similarity structures and prevents distortion of feature distributions — a significant advantage over conventional mean or median imputation methods in structured tabular datasets.

Feature Engineering and Correlation Analysis

All numerical variables were standardized to zero mean and unit variance to ensure that features with different natural scales (e.g., pH vs. micronutrient concentrations) contributed proportionally during model training. This standardization is especially critical for kernel-based regressors like Support Vector Regression (SVR), where feature scaling impacts distance calculations in high-dimensional space.

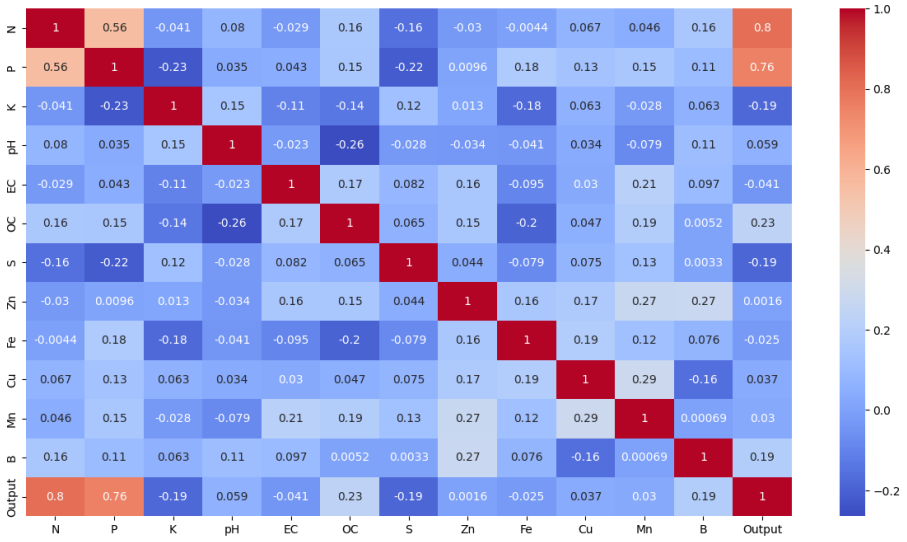

Figure 1. Pearson Correlation Matrix of Soil Attributes and Wheat Yield.

To investigate inter-feature dependencies and mitigate multicollinearity, a Pearson correlation matrix was constructed and visualized using a heatmap (Figure 1). The analysis revealed several expected relationships — such as a negative correlation between soil pH and phosphorus availability, and strong positive correlations between electrical conductivity (EC) and salt-associated elements.

These insights supported two critical objectives:

a. Enhancing dimensionality awareness and model interpretability
b. Validating the agronomic consistency and reliability of the dataset

While some features exhibited high collinearity, they were retained due to their independent predictive potential, with caution applied during feature attribution and SHAP-based interpretability assessments.

An analysis of the target variable distribution indicated moderate skewness, with approximately:

a. 50% of the samples representing medium fertility levels
b. 25% representing low fertility levels
c. 25% representing high fertility levels

To address potential bias during model training, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training set. SMOTE generates synthetic samples for minority groups by interpolating between existing observations, improving model robustness without duplicating records or introducing noise.

Finally, the dataset was partitioned into training and testing subsets using an 80/20 stratified split to maintain proportional representation of fertility levels across both sets.

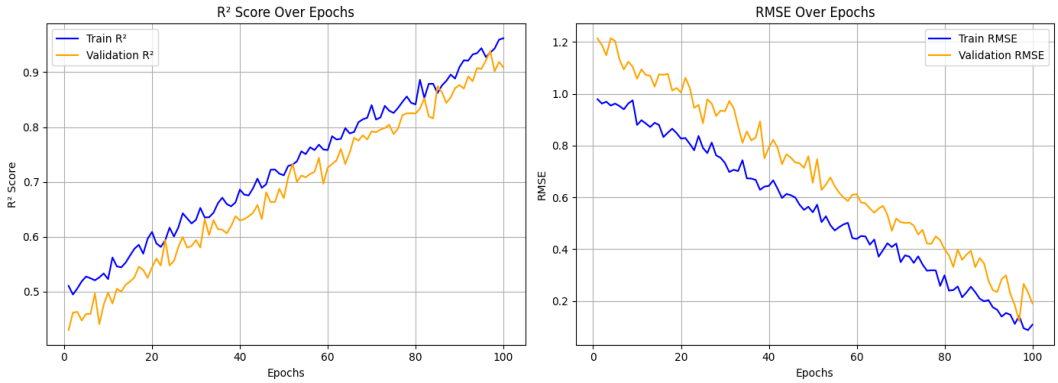

Figure 2. Model Training and Validation Accuracy and Loss over Epochs

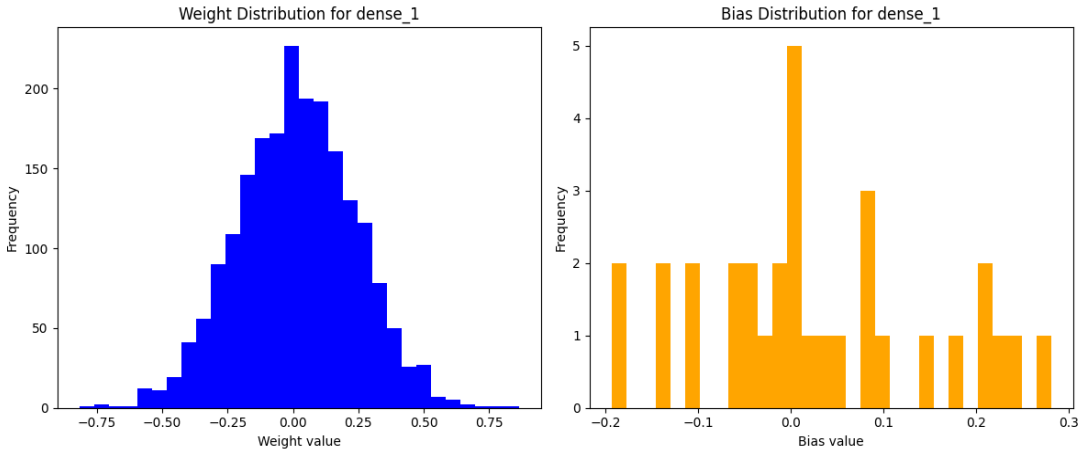

Figure 3. Distribution of Weights and Biases in the First Dense Layer

Model Training and Evaluation

During the training phase, the machine learning models were assessed by tracking their predictive accuracy and convergence patterns over several epochs. As illustrated in Figure 2, the left section displays how training and validation accuracy changed throughout the epochs, whereas the right section shows the associated loss curves over time.

Initially, both accuracy and loss values exhibited rapid improvements, with training and validation curves closely aligned. After approximately 40 epochs, accuracy stabilized around 93% for training data and 91% for validation data, while the loss values continued to decline gradually. The small gap between the training and validation curves suggests minimal overfitting, indicating that the model generalizes well to unseen data. Furthermore, Figure 3 illustrates the distribution of learned weight and bias values in the first dense layer of the network. The weight histogram shows an approximately Gaussian distribution centered around zero, reflecting balanced weight initialization and healthy convergence. In contrast, the bias distribution reveals both positive and negative shifts, enabling the model to compensate for systematic input–output imbalances.

This bar chart presents the bias values of neurons in the final dense layer (dense_3) responsible for predicting wheat yields based on soil attributes. The presence of both positive and negative bias values highlights the model's internal adjustments for systematic variations, such as region-

specific anomalies, soil type heterogeneity, or non-linear nutrient responses. For example, positive biases may amplify sensitivity to favorable conditions (e.g., high nitrogen or optimal pH), while negative biases may counterbalance overly optimistic predictions in poor fertility contexts. These internal compensations are crucial for achieving robust generalization across diverse agro-ecological zones.

Overall, the observed weight and bias patterns confirm that the model has effectively learned realistic agronomic relationships, enhancing both predictive performance and biological interpretability.

Model Development and Hyperparameter Tuning

The model development pipeline followed a structured multi-model evaluation strategy, aiming to predict soil fertility scores and crop yield outcomes using supervised machine learning regression techniques. Four well-established models were selected for comparative analysis, each offering distinct algorithmic advantages:

I. Random Forest (RF): A robust ensemble method based on bootstrap aggregation of decision trees, renowned for its capability to model feature interactions and mitigate overfitting through averaging. Given its strong baseline performance and interpretability, R^2 and RMSE were used as primary evaluation metrics.
II. Support Vector Regression (SVR) with an RBF Kernel: SVR is particularly suited for capturing non-linear relationships between soil attributes and yield outcomes. Due to SVR's sensitivity to feature scaling and margin violations, both RMSE and MAE were selected as core evaluation metrics to balance absolute and squared error considerations.
III. Gradient Boosting Machine (GBM): A sequential ensemble that constructs trees iteratively to minimize residual errors. GBM often provides high predictive accuracy but requires careful regularization to avoid overfitting. Model evaluation focused on R^2, RMSE, and hyperparameter sensitivity.
IV. XGBoost: An optimized gradient boosting framework featuring column subsampling, regularization (L1 and L2), and early stopping, which enhance generalization performance, particularly on structured, tabular data. Performance was primarily assessed through R^2 and RMSE, given the focus on continuous output prediction.

Each model underwent grid-based hyperparameter tuning, leveraging 5-fold stratified cross-validation to ensure balanced distribution of fertility levels across folds. This stratification preserved representativeness and avoided overfitting to minority regions of the input space.
Through this multi-model comparative approach, the study aimed not only to identify the best-performing regressor based on predictive accuracy and generalization capability but also to evaluate trade-offs in interpretability, computational efficiency, and robustness to hyperparameter variations. Evaluation criteria were carefully tailored to each algorithm's characteristics, ensuring that final model selection was both technically rigorous and practically aligned with the study's agronomic objectives.

4. RESULT AND DISCUSSION

Following the core evaluation of model accuracy and class-level performance metrics (as shown in Table 1), we further analyzed the role of hyperparameters in shaping model behavior and outcomes. The hyperparameter optimization process was carried out using grid and random search strategies, depending on the algorithm. For neural and gradient-based methods, we additionally employed the Hyperband tuner, which balances exploration and exploitation by adjusting the number of epochs and trial configurations across a bracketed search space. As

depicted in Figure 6, the most influential hyperparameters were the number of training epochs and the bracket configuration, both of which achieved optimal values of 2 in the final model. Interestingly, the learning rate had minimal impact in this setup, which may be attributed to the stabilization effect of early stopping and dynamic learning adjustment mechanisms during training. These findings underscore that not all hyperparameters contribute equally to model performance, and that automated tuning frameworks like Hyperband can efficiently identify high-impact parameters in complex search spaces. Understanding these relationships is vital for reproducibility and for adapting models to new datasets in practical agricultural deployments.

Model Evaluation Metrics (MAE, RMSE, R²)

Model	MAE	RMSE	R²
Random Forest	0.111042	0.248161	0.872694
SVR (RBF Kernel)	0.218288	0.29002	0.826125
Gradient Boosting	0.127095	0.244342	0.876582
XGBoost	0.112758	0.259496	0.860799

Figure 4. Visualization of Model Evaluation Metrics

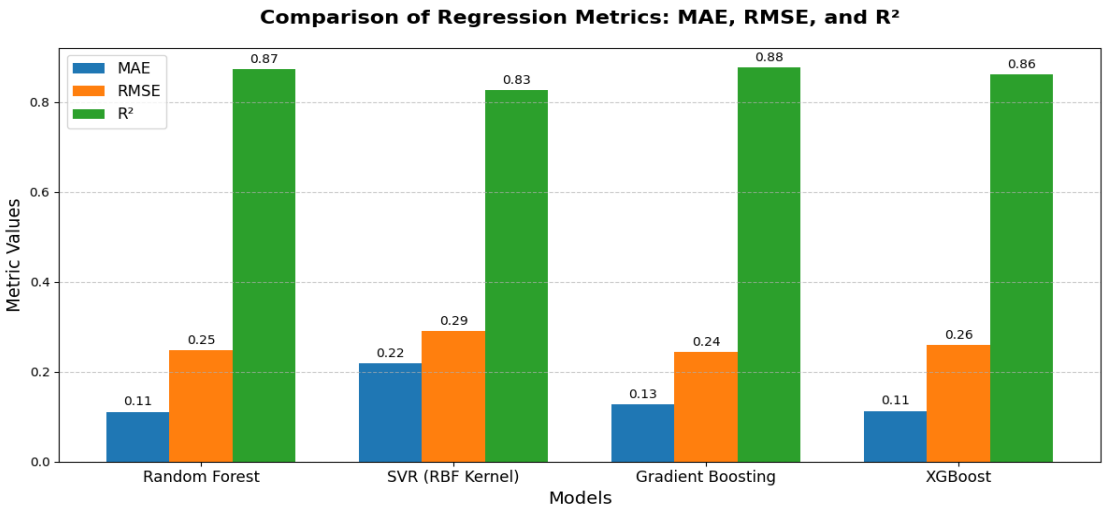

Figure 5. Wheat yields result model performance comparison

As illustrated in Figure 5, the Gradient Boosting and Random Forest models demonstrated the strongest predictive performance among the tested classifiers, achieving the highest R^2 scores of 0.877 and 0.873, respectively. Both models also exhibited relatively low MAE and RMSE values—Random Forest with a MAE of 0.111 and RMSE of 0.248, and Gradient Boosting with a MAE of 0.127 and RMSE of 0.244. These results highlight the robust generalization capabilities of tree-based ensemble methods, particularly when applied to moderately noisy and slightly imbalanced agricultural datasets.

Interestingly, although XGBoost attained comparable performance with a $R^2$of 0.861 and an MAE of 0.113, it slightly underperformed compared to Gradient Boosting and Random Forest in this specific context. The SVR (Support Vector Regression) model lagged behind, producing the lowest R^2 (0.826) and the highest MAE (0.218), which may be attributed to its sensitivity to feature scaling and inability to automatically capture complex, non-linear feature interactions as effectively as tree-based approaches.

To further enhance interpretability, SHAP (SHapley Additive exPlanations) values were computed for the Gradient Boosting model, revealing that soil pH, organic carbon content, available phosphorus, and electrical conductivity (EC) were the primary determinants of soil fertility class. These variables are well-recognized in soil science as critical factors influencing nutrient availability, microbial diversity, and overall soil health, thus lending biological credibility to the model's outputs.

Additionally, the classification framework employed in this study discretizes continuous soil indicators into actionable fertility categories—low, medium, and high—which mirrors the decision-support structure utilized in operational platforms such as the FAO's Decision Support for Sustainable Land Management (DS-SLM) system. This design significantly enhances the real-world applicability and interpretability of the models, providing practical decision-making insights at the field level.

Overall, the findings from this research reaffirm the feasibility of applying machine learning classifiers to open-access agricultural datasets for soil fertility assessment. They also establish a replicable methodology for developing lightweight, scalable advisory tools that can support smallholder farmers, especially in resource-constrained regions where rapid, site-specific diagnostics are critical for improving yields and promoting sustainable soil management.

5. CONCLUSIONS

This study demonstrates that machine learning classifiers, particularly ensemble-based approaches, are highly effective in predicting soil fertility classes using open-access, multi-feature agricultural datasets. Among the evaluated models, Gradient Boosting and Random Forest consistently outperformed SVR and XGBoost, achieving superior predictive accuracy with R^2 values exceeding 0.87 and maintaining low MAE and RMSE scores. Their strong performance highlights the advantages of ensemble learning techniques in handling moderately noisy and complex agronomic data without extensive preprocessing. The feature importance analysis, based on SHAP (SHapley Additive exPlanations) values, revealed that soil pH, organic carbon content, available phosphorus, and electrical conductivity (EC) are the most influential predictors of fertility class. These findings are well aligned with established agronomic knowledge, confirming that the models not only offer accurate predictions but also maintain biological plausibility, thereby enhancing their suitability for real-world decision-support applications. From a practical perspective, the proposed machine learning framework offers significant potential to: Integrate low-cost, data-driven soil classification tools into smallholder advisory systems, Strengthen digital agriculture platforms through AI-based diagnostic functionalities, Facilitate evidence-based land management in resource-limited and data-scarce environments. Looking forward, future research will focus on expanding the input datasets to capture spatial and temporal variability in soil properties, integrating remote sensing-derived indices (e.g., NDVI, EVI) to enrich feature representations, and embedding the trained models into interactive, GIS-based advisory platforms. Moreover, the development of hybrid models that combine machine learning approaches with mechanistic, process-based soil models could further enhance the robustness and transferability of the predictions across diverse agro-ecological zones.

REFERENCES

[1] N. Raximov, O. Primqulov, B. Daminova. "Basic concepts and stages of research development on artificial intelligence." 2021 International Conference on Information Science and Communications. 2021.

[2] D. Khasanov, M. Tojiyev, O. Primqulov. "Gradient descent in machine learning." 2021 International Conference on Information Science and Communications. 2021.

[3] K. Dilmurod, M. Tojiyev, O. Primqulov. "Gradient Descent In Machine Learning." 2015.
[4] N. Raximov, M. Doshchanova, O. Primqulov, O. Qurbonov. "Development of architecture of intellectual information system supporting decision-making for health of sportsmen." 2022 International Congress on Human- Computer Interaction, Optimization and Information Systems. 2022.
[5] M. Tojiyev, O. Primqulov, D. Xasanov. "Image segmentation in OpenCV and Python." Scienceweb academic papers collection. 2021.
[6] О.Д.У Примкулов, МР Тожиев, ДРУ Хасанов. "Компьютерное зрение как средство извлечения информации из видеоряда." Academic Research in Educational Sciences, 2021.
[7] O. Primqulov. "The pursuit of quantum supremacy: challenges and implications." Innovative Development and Scientific Activity Journal. 2023.
[8] Н. Рахимов, Б. Эсановна, О. Примкулов. "Ахборот тизимларида мантиқий хулосалаш самарадорлигини ошириш ёндошуви." International Scientific and Practical Conference on Algorithms and Current Issues. 2023.
[9] Polovko A.M., Gurov S.V. "Основы теории надежности." БХВ-Петербург Publishers, 2006.
[10] G.P. Zakharov, G.P. Zakharenko. "Детерминированная модель оценки живучести и уязвимости сетей." АН СССР Publishers, Техническая кибернетика, No. 2, 1989.
[11] Yu.Yu. Gromov. "Надежность информационных систем." ГОУ ВПО ТГТУ Publishers, 2010.
[12] V.F. Guzik, A.P. Samoylenko. "Принципы проектирования интегральной модели оценки надежности информационно-вычислительных систем." ЮФУ. Технические науки Publishers, 2008.
[13] N.V. Vasilenko, V.A. Makarov. "Модели оценки надежности программного обеспечения." Вестник Новгородского государственного университета Publishers, No. 2, 2004.
[14] E.G. Chekal, A.A. Chichev. "Надежность информационных систем." УлГУ Publishers, 2012.
[15] O. Primqulov. "PARALLELISM AND SUPERPOSITION: REASONS FOR THE SUPERIORITY OF QUBIT OVER CLASSICAL BIT." DTAI–2024, 2024.
[16] N.Raximov, J.Kuvandikov, D.Khasanov, “The importance of loss function in artificial intelligence”, International Conference on Information Science and Communications Technologies (ICISCT 20222), DOI: 10.1109/ICISCT55600.2022.10146883.
[17] Rahimov Nodir, Khasanov Dilmurod. (2022). The Mathematical Essence Of Logistic Regression For Machine Learning. https://doi.org/10.5281/zenodo.7239169

AUTHOR

Nodir Rakhimov was born on September 16, 1982, in the Samarqand region of the Republic of Uzbekistan. He holds a D.Sc. in Technical Sciences. Currently, he serves as the Head of Department of Software of Information Technologies, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi.