DOI: 10.20998/2079-0023.2026.01.11
UDC 004.9

A. M. KOPP, Doctor of Philosophy (PhD), Docent, National Technical University "Kharkiv Polytechnic Institute", Head of Software Engineering and Management Intelligent Technologies Department, Kharkiv, Ukraine, e-mail: andrii.kopp@khpi.edu.ua, ORCID: https://orcid.org/0000-0002-3189-5623
I. P. GAMAYUN, Doctor of Technical Sciences, Professor, National Technical University "Kharkiv Polytechnic Institute", Full Professor of Software Engineering and Management Intelligent Technologies Department, Kharkiv, Ukraine, e-mail: ihor.hamaiun@khpi.edu.ua, ORCID: https://orcid.org/0000-0003-2099-4658
R. B. DASHKIVSKYI, Postgraduate Student, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine; e-mail: roman.dashkivskyi@cs.khpi.edu.ua; ORCID: https://orcid.org/0009-0006-8066-3622
Ye. R. KOSTIN, Postgraduate Student, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine; e-mail: yehor.kostin@cs.khpi.edu.ua; ORCID: https://orcid.org/0009-0009-5042-0795

INFORMATION TECHNOLOGY FOR OPTIMAL SERVICE PLACEMENT PREDICTION IN A MULTI-CLOUD ENVIRONMENT USING MACHINE LEARNING

The relevance of the work is due to the need to improve the efficiency of service distribution management in multi-cloud infrastructures, where optimal service placement directly affects latency, performance, reliability, and rational use of resources. The object of the study is the process of placing cloud services in a multi-provider environment. The subject of the study includes machine learning methods and algorithms that are used to predict optimal decisions for placing cloud services in a multi-provider environment based on measured performance indicators. The purpose of the study is to develop and evaluate models for predicting optimal placement of cloud services in a multi-provider environment using historical data on latency, response time, and load balancing efficiency. The work uses an open dataset, the Multi-Cloud Service Composition Dataset, which contains characteristics of services from AWS, Azure, Google Cloud, and IBM providers. Six machine learning algorithms implemented using the Python programming language and the Scikit-learn library were used for prediction. The obtained results showed that models based on Gradient Boosting and Naive Bayes provide the highest consistency of the metrics Accuracy, Precision, Recall and F1-score, reaching values of about 0.97–0.98, which confirms their suitability for the tasks of optimizing the placement of cloud services in a multi-cloud environment. Other developed models demonstrated lower stability of results, which limits their application in real conditions. The conclusions substantiate the possibility of using machine learning methods and algorithms to build adaptive load management systems in multi-cloud environments, and also identify prospects for expanding the proposed information technology by including additional parameters, such as energy consumption, computing cost and fault tolerance.

Keywords: machine learning, cloud computing, cloud infrastructure, optimal service placement, prediction models, information technology.

Problem statement. The rapid growth of multi-cloud infrastructures creates a need for accurate and adaptive methods of deploying services that operate in environments with different providers, different performance metrics, and variable load conditions.

Optimal service placement determines latency, throughput, stability, and resource efficiency. However, current approaches remain fragmented. Some research focuses on search and heuristic methods for configuring multi-cloud solutions but does not take into account the complexity of real-world performance metrics [1].

Other works demonstrate the development of multi-cloud frameworks, but are mainly focused on specific applications, which does not allow them to be directly used for the general task of optimal service placement [2]. A separate area of research applies Machine Learning (ML) to solve placement problems in uncertain environments, but these studies mainly cover edge infrastructures rather than multi-cloud systems with heterogeneous providers [3].

Thus, the problem of creating a universal approach that combines the processing of performance metrics, machine learning, and the ability to accurately predict the optimal placement of services in a multi-cloud environment remains unresolved.

This creates a scientific challenge – to develop models capable of consistently determining optimal placement based on various indicators, ensuring high accuracy and practical applicability in dynamic IT systems.

Related work. The problem of service placement in hybrid and multi-cloud environments is actively being researched in the context of combining cloud and fog infrastructure. In their work, Azizi et al. propose the FLEX platform for scalable and flexible service deployment in multi-fog and multi-cloud environments, where the task is formulated as an integer linear programming problem and solved by a heuristic algorithm to minimize cost and delays [4]. A similar approach is developed by Dogani et al., who consider a two-layer scheme for deploying services in containerized fog-cloud platforms and apply NSGA-II to simultaneously optimize latency, power consumption, and cost [5]. Both works focus on optimization and evolutionary methods, but do not use classical machine learning methods for direct prediction of the “optimality” of placement based on empirical performance metrics.

A separate branch of research is devoted to dynamic service placement in 5G/6G scenarios and edge infrastructures. Tabatabaei et al. analyze dynamic service placement in 6G multi-cloud scenarios, taking into account low latency and reliability requirements, and propose an approach to placement management in complex network environments [6]. Lu et al. develop the Dynamic Service Placement with Deep Reinforcement Learning (DSP-DRL) framework for dynamic service placement in mobile edge computing using deep reinforcement learning to minimize total delay under resource and cost constraints [7]. These works demonstrate the potential of reinforcement learning

Research Article: This article was published by the publishing house of ***NTU "KhPI"*** in the journal *Bulletin of the National Technical University "KhPI". Series: System Analysis, Control and Information Technologies*. This article is distributed under the Creative Common Attribution (CC BY 4.0) international license. **Conflict of Interest:** The author/s declared no conflict of interest.

for adaptive placement, but mostly work in the context of edge/fog, rather than for optimal service placement tasks between several large cloud providers with explicit performance metrics.

Another area is related to the application of machine and deep learning for resource allocation tasks in cloud systems. In a recent review, Zhou et al. systematize methods based on deep reinforcement learning for resource planning tasks in cloud environments, demonstrating their advantage over classical heuristics in scenarios with dynamic loads and multi-criteria objectives [8]. Bodra and Khairnar perform a comparative analysis of modern machine learning algorithms for cloud resource allocation, including DRL, neural networks, and traditional ML methods, and demonstrate significant improvements in execution time, cost, and energy consumption compared to classical approaches [9]. However, these reviews focus primarily on general resource planning and allocation tasks (job scheduling, virtual machine placement, container allocation) and do not directly address the prediction of optimal service placement between specific multi-cloud providers based on integrated performance characteristics.

Thus, existing works demonstrate the importance of optimization, evolutionary, and DRL approaches for service placement and resource allocation tasks in cloud and fog-cloud infrastructures. At the same time, there remains a gap in the use of classical machine learning models for predicting the binary measure of placement optimality (e.g., optimal service placement) in multi-cloud environments based on measured metrics of latency, response time, and load balancing efficiency.

Research objective. The proposed research aims to fill this niche by using an open multi-cloud dataset and comparing several ML algorithms to support decisions on optimal service placement.

Materials and methods. The research was based on the open Multi-Cloud Service Composition Dataset [10], which contains measured performance characteristics of services in a multi-cloud environment. The dataset includes data on Service Latency (ms), Response Time (ms), and Load Balancing (%), as well as a target variable that indicates the optimality of service placement as a binary indicator. The data covers various types of services running on AWS, Azure, Google Cloud, and IBM infrastructures [10]. The dataset was used to build prediction models that allow the optimal placement of services in heterogeneous multi-provider environments to be evaluated.

The research methodology (Fig. 1) was based on the use of ML methods implemented in the Google Colab [11] using the Scikit-learn [12]. Preliminary data processing and preparation was performed, which included the selection of relevant attributes, their scaling, and distribution into training (75 %) and test subsets (25 %). This stage allowed us to form a consistent sample for correct comparison of models and obtaining stable results. The study selected six machine learning algorithms belonging to different model classes: Decision Tree, Random Forest, Naive Bayes, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Gradient Boosting (Fig. 2).

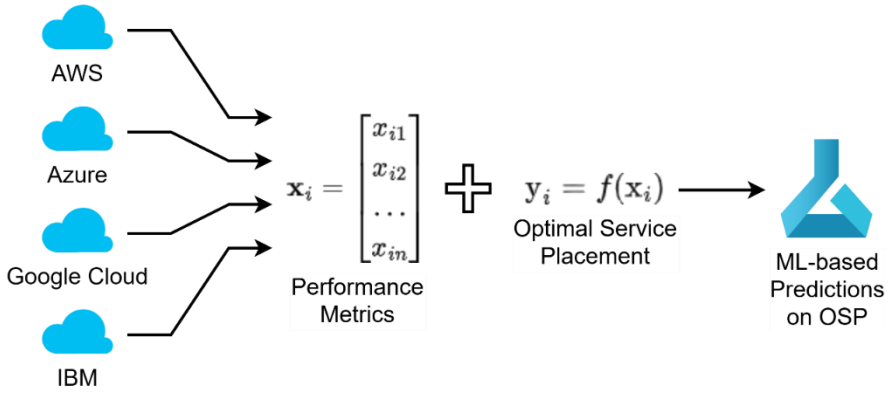

Fig. 1. Using machine learning to predict Optimal Service Placement (OSP) based on multi-cloud performance metrics

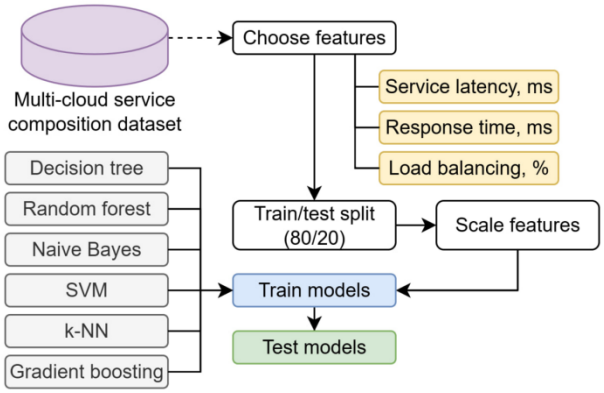

Fig. 2. ML-based workflow used in this study

The ML models were trained based on historical performance measurements [10] and then evaluated using standard classification performance metrics – Accuracy, Precision, Recall, and the F1-score [13]. We identified the algorithms that demonstrated the best consistency of results for the task of predicting optimal service placement.

The results presented in the figure demonstrate the high efficiency of the Decision Tree model (Fig. 3), which provided an Accuracy of 0.97, indicating an almost error-free classification of most cases.

A Precision of 0.96 means that the model rarely makes mistakes in determining the positive class. The Recall is

also 0.97, confirming the model's ability to detect almost all relevant objects. The F1-score is 0.96, indicating a balanced performance between precision and recall.

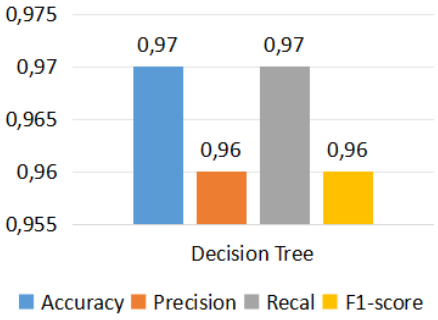

Fig. 3. Decision Tree model performance

The results for Random Forest indicate the high overall quality of the model (Fig. 4), which achieved an Accuracy of 0.96, meaning that most objects were classified correctly. The Precision is 0.91, which means a slightly higher number of false positives compared to the decision tree. At the same time, the Recall is 0.96, demonstrating the model's ability to effectively find almost all objects of the positive class. The F1-score of 0.93 reflects balance, but with a slight decrease compared to other metrics.

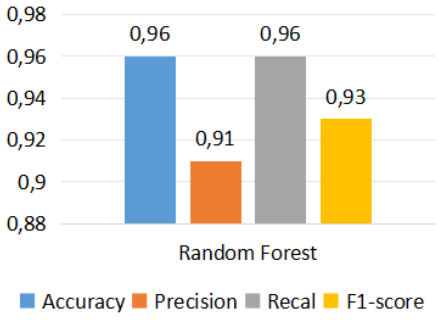

Fig. 4. Random Forest model performance

The results for Naive Bayes demonstrate consistently high model quality across all indicators, demonstrating its stability and consistency (Fig. 5).

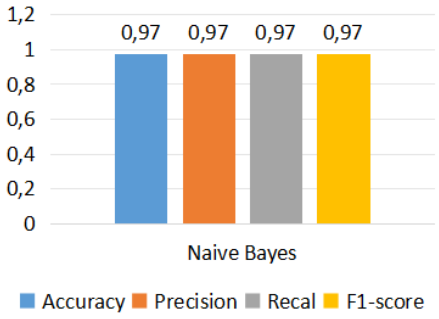

Fig. 5. Naive Bayes model performance

Accuracy is 0.97, meaning that the model correctly classifies the vast majority of cases. The Precision, Recall, and F1-score metrics are also 0.97, indicating no significant imbalance between the number of false positives and false negatives. Such consistency in metrics indicates that Naive Bayes performs well in classification tasks and provides predictable behavior across the entire dataset.

The results for SVM show high overall model performance (Fig. 6), with an Accuracy of 0.96, indicating correct classification of most objects. The Precision is 0.91, indicating a slightly higher number of false positives compared to other metrics. At the same time, the Recall is 0.96, meaning that the model successfully detects almost all relevant cases. The F1-score of 0.93 demonstrates a balance between precision and recall, but with a slight decrease due to lower accuracy.

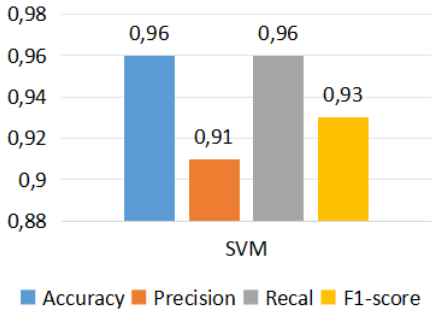

Fig. 6. SVM model performance

The results for k-NN demonstrate stable and consistent classification quality (Fig. 7). The model achieved an Accuracy of 0.95, which means that most examples were predicted correctly. The Precision score is 0.94, meaning that k-NN accurately identifies the positive class with a minimum number of false positives. Recall is 0.95, indicating effective detection of relevant cases. The F1-score of 0.95 confirms the balance of the model and the harmonious combination of accuracy and completeness.

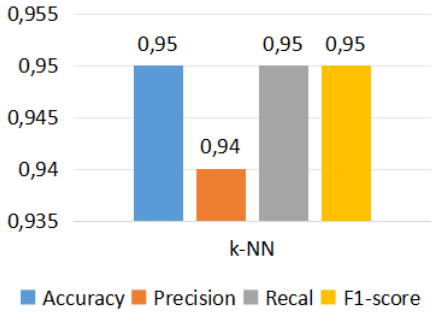

Fig. 7. k-NN model performance

The results for Gradient Boosting demonstrate the highest quality among all models considered (Fig. 8), as all key metrics have the same value of 0.98. An Accuracy of 0.98 indicates an almost error-free classification of examples. Precision and recall are also at 0.98, which means a very low number of false positives and false negatives. The F1-score of 0.98 confirms the complete balance of the model.

Such uniformity and the high performance measures show that Gradient Boosting provides the best consistency and stability among the algorithms, making it particularly effective for optimal service placement classification tasks in the experimental environment.

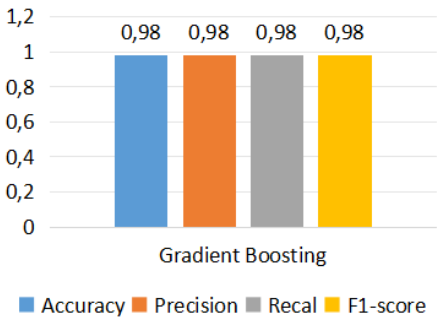

Fig. 8. Gradient Boosting model performance

The proposed intelligent information technology based on ML-algorithms provides automated collection of multi-cloud performance metrics using APIs (Application Programming Interface) and provider monitoring tools of the open access (Fig. 9):

- The system prepares and labels data according to predefined rules, forming a consistent set of measurements for training ML models.
- Models are trained and validated on the prepared dataset using specified hyperparameters and frameworks that implement ML algorithms.
- The information technology provides predictions for the OSP based on current performance metrics (e.g. latency, response time, and load balancing efficiency) and optimization policies.
- The obtained results are integrated with multi-cloud orchestration systems to support decision-making and automated service management.

Conclusions. The study confirmed the possibility of effectively applying ML methods and algorithms to predict the optimal service placement in multi-cloud environments based on the developed intelligent information technology (Fig. 9). Analysis of the results showed that Gradient Boosting and Naive Bayes models provide the highest accuracy and consistency of performance indicators, making them suitable for practical tasks of optimizing performance and minimizing delays.

The use of an open multi-cloud dataset [12] made it possible to evaluate the behavior of algorithms in realistic conditions when working with different types of services and different cloud providers. The results demonstrate the promise of such ML models for building adaptive resource management and decision support information technology in modern cloud architectures.

Further research may focus on expanding the set of parameters, including energy and computing costs, and fault tolerance indicators. The use of hybrid models that combine "classical" ML with reinforcement learning algorithms to improve adaptability to dynamic loads is also promising. A separate problem is to test the models on significantly larger and more heterogeneous production datasets, which will allow us to evaluate the scalability of the proposed approach. In the future, a prototype could also be developed that would predict the optimal service placement in real time and interact with the infrastructure tools of cloud platforms.

Declaration on the use of generative AI. During the preparation of this work, the authors used ChatGPT for rephrasing and reformulating the text, as well as DeepL for grammar and spell checking. After using these tools, the authors reviewed and edited the content as necessary and take full responsibility for the content of this publication.

References

1. Łazuka M., Parnell T., Anghel A. *Search-based methods for multi-cloud configuration*. Available at: https://arxiv.org/abs/2204.09437 (accessed: 21.01.2026). DOI: https://doi.org/10.48550/arXiv.2204.09437.
2. Brum R. C., Stelling de Castro M. C., Arantes L., Drummond L. M. de A., Sens P. *Multi-FedLS: A framework for cross-silo federated learning applications on multi-cloud environments*. Available at: https://arxiv.org/abs/2308.08967 (accessed: 21.01.2026). DOI: https://doi.org/10.48550/arXiv.2308.08967.

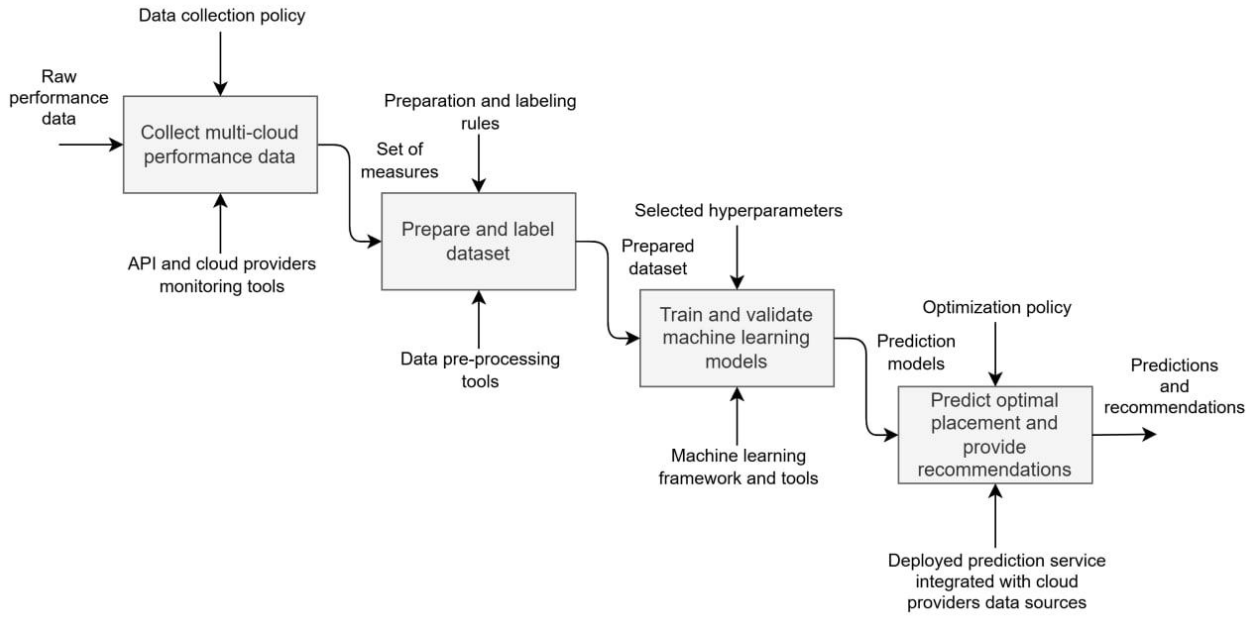

Fig. 9. Information technology to predict optimal placing of cloud services in a multi-provider environment

3. Hejazi T. H., Ghadimkhani Z., Borji A. *A learning-based solution approach to the application placement problem in mobile edge computing under uncertainty*. Available at: https://arxiv.org/pdf/2403.11259 (accessed: 21.01.2026). DOI: https://doi.org/10.48550/arXiv.2403.11259.
4. Azizi S., Farzin P., Shojafar M., Rana O. *A Scalable and Flexible Platform for Service Placement in Multi-Fog and Multi-Cloud Environments. ResearchSquare*. Available at: https://link.springer.com/article/10.1007/s11227-023-05520-9 (accessed: 21.01.2026). DOI: https://doi.org/10.1007/s11227-023-05520-9.
5. Dogani J., Yazdanpanah A., Zare A., Khunjush F. *A Two-tier Multi-Objective Service Placement in Container-based Fog-Cloud Computing Platforms. Preprint*. Available at: https://scispace.com/pdf/a-two-tier-multi-objective-service-placement-in-container-34szay2e.pdf (accessed: 21.01.2026). DOI: https://doi.org/10.21203/rs.3.rs-3130299/v1.
6. Tabatabaei F., Khalili H., Requena M., Kahvazadeh S., Mangues-Bafalluy, J. (2023). *Dynamic Service Placement in 6G Multi-Cloud Scenarios*. Available at: https://zenodo.org/records/10959481 (accessed: 21.01.2026). DOI: https://doi.org/10.1109/ICTON59386.2023.10207547.
7. Lu S., et al. *A Dynamic Service Placement Based on Deep Reinforcement Learning in Mobile Edge Computing*. Available at: https://www.mdpi.com/2673-8732/2/1/8 (accessed: 21.01.2026). DOI: https://doi.org/10.3390/network2010008.
8. Zhou G., Tian W., Buyya R., Xue R., Song L. *Deep reinforcement learning-based methods for resource scheduling in cloud computing: A review and future directions*. Available at: https://clouds.cis.unimelb.edu.au/papers/DRLCloudReview2024.pdf (accessed: 21.01.2026). DOI: https://doi.org/10.1007/s10462-024-10756-9.
9. Bodra D., Khairnar R. *Machine learning-based cloud resource allocation algorithms: A comprehensive comparative review*. Available at: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2025.1678976/pdf (accessed: 21.01.2026). DOI: https://doi.org/10.3389/fcomp.2025.1678976.
10. *Supervised learning – scikit-learn 0.22 documentation*. Available at: https://scikit-learn.org/stable/supervised_learning.html (accessed: 21.01.2026).
11. *Google Colab*. Available at: https://colab.google/ (accessed: 21.01.2026).
12. Ziya. *Multi-Cloud Service Composition Dataset*. Available at: https://www.kaggle.com/datasets/ziya07/multi-cloud-service-composition-dataset (accessed: 21.01.2026).
13. Dalianis H. *Clinical Text Mining*. Available at: https://link.springer.com/chapter/10.1007/978-3-319-78503-5_6 (accessed: 21.01.2026). DOI: https://doi.org/10.1007/978-3-319-78503-5_6.

Received 11.03.2026
Accepted 05.04.2026
Published 20.05.2026

УДК 004.9

А. М. КОПП, доктор філософії (PhD), доцент, Національний технічний університет «Харківський політехнічний інститут», завідувач кафедри програмної інженерії та інтелектуальних технологій управління, м. Харків, Україна, e-mail: andrii.kopp@khpi.edu.ua, ORCID: https://orcid.org/0000-0002-3189-5623
І. П. ГАМАЮН, доктор технічних наук, професор, Національний технічний університет «Харківський політехнічний інститут», професор кафедри програмної інженерії та інтелектуальних технологій управління, м. Харків, Україна, e-mail: ihor.hamaiun@khpi.edu.ua, ORCID: https://orcid.org/0000-0003-2099-4658
Р. Б. ДАШКІВСЬКИЙ, аспірант, Національний технічний університет «Харківський політехнічний інститут», м. Харків, Україна; e-mail: roman.dashkivskyi@cs.khpi.edu.ua; ORCID: https://orcid.org/0009-0006-8066-3622
Є. Р. КОСТІН, аспірант, Національний технічний університет «Харківський політехнічний інститут», м. Харків, Україна; e-mail: yehor.kostin@cs.khpi.edu.ua; ORCID: https://orcid.org/0009-0009-5042-0795

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПРОГНОЗУВАННЯ ОПТИМАЛЬНОГО РОЗМІЩЕННЯ ПОСЛУГ У БАГАТОХМАРНОМУ СЕРЕДОВИЩІ НА ОСНОВІ МАШИННОГО НАВЧАННЯ

Актуальність роботи зумовлена потребою підвищення ефективності управління розподілом сервісів у багатохмарних інфраструктурах, де оптимальне розміщення послуг безпосередньо впливає на затримки, продуктивність, надійність та раціональне використання ресурсів. Об’єктом дослідження є процес розміщення хмарних послуг у середовищі з декількома провайдерами. Предметом дослідження виступають методи та алгоритми машинного навчання, які застосовуються для прогнозування оптимальних рішень щодо розміщення хмарних послуг у середовищі з декількома провайдерами на основі вимірюваних показників продуктивності. Метою дослідження є розробка та оцінювання моделей прогнозування оптимального розміщення хмарних послуг у середовищі з декількома провайдерами з використанням історичних даних про затримку, час відгуку та ефективність балансування навантаження. У роботі використано відкритий набір даних Multi-Cloud Service Composition Dataset, що містить характеристики сервісів провайдерів AWS, Azure, Google Cloud та IBM. Для прогнозування застосовано шість алгоритмів машинного навчання, реалізованих із застосуванням мови програмування Python та з використанням бібліотеки Scikit-learn. Отримані результати показали, що моделі на основі Gradient Boosting та Naive Bayes забезпечують найвищу узгодженість метрик Accuracy, Precision, Recall та F1-score, досягаючи значень близько 0.97–0.98, що підтверджує їхню придатність для задач оптимізації розміщення хмарних послуг у багатохмарному середовищі. Інші розроблені моделі продемонстрували нижчу стабільність результатів, що обмежує їх застосування в реальних умовах. У висновках обґрунтовано можливість використання методів і алгоритмів машинного навчання для побудови адаптивних систем керування навантаженням у багатохмарних середовищах, а також визначено перспективи розширення запропонованої інформаційної технології шляхом включення додаткових параметрів, таких як енергоспоживання, вартість обчислень та відмовостійкість.

Ключові слова: машинне навчання, хмарні обчислення, хмарна інфраструктура, оптимальне розміщення послуг, моделі прогнозування, інформаційна технологія.

Повні імена авторів / Author's full names

Автор 1 / Author 1: Копп Андрій Михайлович / Kopp Andrii Mykhailovych
Автор 2 / Author 2: Гамаюн Ігор Петрович / Gamayun Igor Petrovych
Автор 3 / Author 3: Дашківський Роман Борисович / Dashkivskyi Roman Borysovych
Автор 4 / Author 4: Костін Єгор Романович / Kostin Yehor Romanovych