

УДК 004.852:519.876.5:004.93

МОДЕЛЮВАННЯ ВИЖИВАННЯ ПАСАЖИРІВ «ТИТАНІКА» З ВИКОРИСТАННЯМ БАГАТОКРИТЕРІАЛЬНОГО АНАЛІЗУ ТА МЕТОДІВ МАШИННОГО НАВЧАННЯ

І. М. Лях, В. С. Морохович, Р. В. Літвинчук, А. Ю. Ціпіньо

ДВНЗ «Ужгородський національний університет»,
пл. Народна, 3, Ужгород, 88000, Україна

У статті розглядається моделювання ймовірності виживання пасажирів лайнера «Титанік» із використанням методів машинного навчання та багатокритеріального аналізу (MCDA). Метою дослідження є виявлення чинників, які найбільше впливали на шанси пасажирів вижити в умовах катастрофи, а також побудова прогностичної моделі з високою точністю класифікації. Аналіз здійснено на основі відкритого датасету з платформи Kaggle, який містить відомості про вік, стать, клас квитка, вартість проїзду, наявність родичів та інші важливі ознаки. Проведено повний цикл попередньої обробки даних, включно з виявленням і обробкою пропусків, нормалізацією та формуванням навчальної і тестової вибірок. Для побудови моделей використано алгоритми логістичної регресії, дерева рішень, метод опорних векторів (SVM), випадковий ліс, а також градієнтний бустинг. Оцінка важливості ознак здійснювалась за допомогою методів LIME і TOPSIS. Результати підкреслюють важливість впливу на виживання пасажирів соціально-демографічних чинників, зокрема статі, віку та класу квитка. Отримані висновки мають практичне значення для аналізу ризиків, побудови моделей прийняття рішень і вдосконалення систем безпеки у кризових ситуаціях, де швидкість та точність мають вирішальне значення.

Ключові слова: багатокритеріальний аналіз, машинне навчання, аналіз даних, прогнозування, метод опорних векторів, випадковий ліс.

Постановка проблеми. Виживання людей у критичних ситуаціях є однією з ключових тем для аналізу в сучасних дослідженнях. Вивчення поведінки та факторів, які впливають на шанси вижити, дозволяє краще зрозуміти, як приймаються рішення в екстремальних умовах, що має широку практичну цінність.

Катастрофа лайнера «Титанік» слугує історичним прикладом для дослідження цих аспектів, оскільки вона дозволяє моделювати вплив різних змінних, таких як: вік, стать, соціальний статус чи клас квитка. Застосування сучасних методів, включно з машинним навчанням та багатокритеріальним аналізом, дозволяє глибше оцінити ці фактори та створити узагальнені моделі виживання, які можна використовувати для прогнозування і розробки стратегій підвищення безпеки в майбутньому.

Аналіз останніх досліджень і публікацій. Дослідження виживання пасажирів «Титаніка» активно використовує машинне навчання для аналізу критичних

ситуацій. Результати досліджень даної тематики пропонують різноманітні підходи до виявлення ключових факторів, що впливають на шанси виживання пасажирів під час катастрофи.

У статті [1] підкреслюється важливість попередньої обробки даних та вибору ознак для досягнення високої точності моделей, зокрема через випадковий ліс і градієнтний бустинг. Автори дослідження [2] акцентують увагу на соціально-демографічних аспектах, таких як вік і стать, які суттєво впливають на ймовірність виживання, підкреслюючи важливість контекстуальних елементів.

Сучасні дослідження, присвячені моделюванню виживання пасажирів під час катастрофи «Титаніка», зосереджуються на застосуванні методів машинного навчання для аналізу факторів, що впливають на ймовірність виживання. Наприклад, автори статей [3, 4] виявили, що вік, стать та клас квитка є ключовими факторами, які суттєво впливають на шанси виживання, підкреслюючи важливість соціально-економічного статусу в умовах катастрофи.

Дослідження [5] акцентує на значущості багатокритеріального аналізу (MCDA) у прийнятті рішень, що може бути корисним для моделювання виживання пасажирів «Титаніка». Вони вказують на необхідність ідентифікації ключових критеріїв, які впливають на результати, що може бути застосовано для оцінки характеристик пасажирів, таких як: вік, стать та клас квитка. Це свідчить про потенціал MCDA у поєднанні з методами машинного навчання для глибшого аналізу факторів, що впливають на виживання в умовах катастроф.

У статті [6] розглядається використання алгоритмів машинного навчання для аналізу катастрофи лайнера, ілюструючи, як сучасні аналітичні підходи можуть бути застосовані для розв'язання задач, що мають широку актуальність для кризового менеджменту, аналізу поведінки та розробки моделей прийняття рішень.

Автори дослідження [7] порівнюють алгоритми, виявляючи, що випадковий ліс демонструє найвищу точність. Аналіз, проведений дослідниками [8], виявив приховані залежності між ознаками, що підкреслює важливість візуалізації. У статті [9] акцентується увага на CatBoost для роботи з категоріальними ознаками. Дослідження [10] демонструє, як градієнтний бустинг покращує прогнози, а в статті [11] пропонуються нові підходи до предикативних моделей.

Формулювання мети дослідження: застосування багатокритеріального аналізу для комплексної оцінки впливу різних чинників на виживання пасажирів лайнера «Титанік» та побудова прогностичної моделі з високою точністю класифікації.

Виклад основного матеріалу. Kaggle є глобальною платформою для аналітиків даних, дослідників і розробників у галузі машинного навчання та штучного інтелекту [12]. Вона забезпечує доступ до бібліотеки датасетів, які охоплюють широкий спектр галузей, від економіки до медицини та екології. Крім того, платформа пропонує інструменти для побудови моделей, обміну досвідом у спільноті експертів, участі в змаганнях і вдосконалення навичок у реальних проектах.

Однією з найвідоміших задач на платформі є «Titanic Dataset», яка використовується як навчальний приклад для вирішення задач класифікації у сфері машинного навчання. У даному дослідженні було обрано датасет [13], який містить

12 колонок, кожна з яких відповідає окремій ознаці, і включає 869 записів про пасажирів (рис. 1). Це дозволило провести аналіз даних з метою вивчення факторів, що вплинули на виживання.

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.0500	NaN	S
6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349989	21.0750	NaN	S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.1333	NaN	S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.0788	NaN	C

Рис. 1. Набір даних пасажирів «Titanic Dataset» на платформі Kaggle

Початковий етап дослідження передбачав первинний огляд датасету, що дозволило зрозуміти його структуру та виявити аномалії. Серед них – наявність пропусків у деяких ознаках і значну різноманітність соціально-демографічних характеристик пасажирів, які вимагали обов'язкової попередньої обробки (рис. 2). Дані дії забезпечили збереження та покращення цінної інформації, спрощення подальшого аналізу і зменшення ризику отримання негативних результатів.

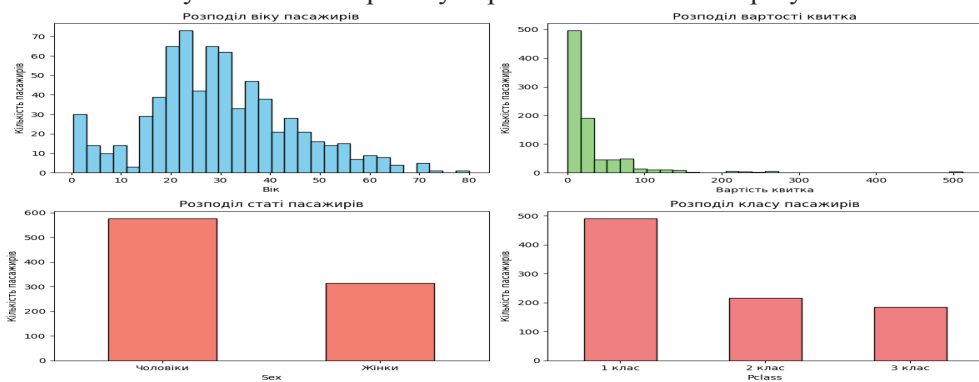


Рис. 2. Розподіл соціально-демографічних характеристик даних

Важливим етапом у підготовці даних стала нормалізація ознак, яка забезпечує приведення всіх значень до єдиного масштабу. Це особливо важливо в контексті машинного навчання, де різні ознаки можуть мати різні одиниці виміру та діапазони значень. Нормалізація дозволяє уникнути ситуацій, коли ознаки з більшими значеннями домінують над меншими, що може призвести до спотворення результатів моделювання. У даному дослідженні використовувався стандартний метод `StandardScaler`, який перетворює дані за формулою (1):

$$X' = \frac{X - \mu}{\sigma}, \quad (1)$$

де X' – нормалізоване значення; X – початкове значення; μ – середнє значення ознаки; σ – стандартне відхилення.

Цей підхід не лише покращує продуктивність моделей, але й забезпечує їхню стабільність, що є важливим для отримання надійних і точних прогнозів. Результати нормалізації підтверджують, що дані успішно підготовлені для подальшого аналізу та навчання моделі (рис. 3).

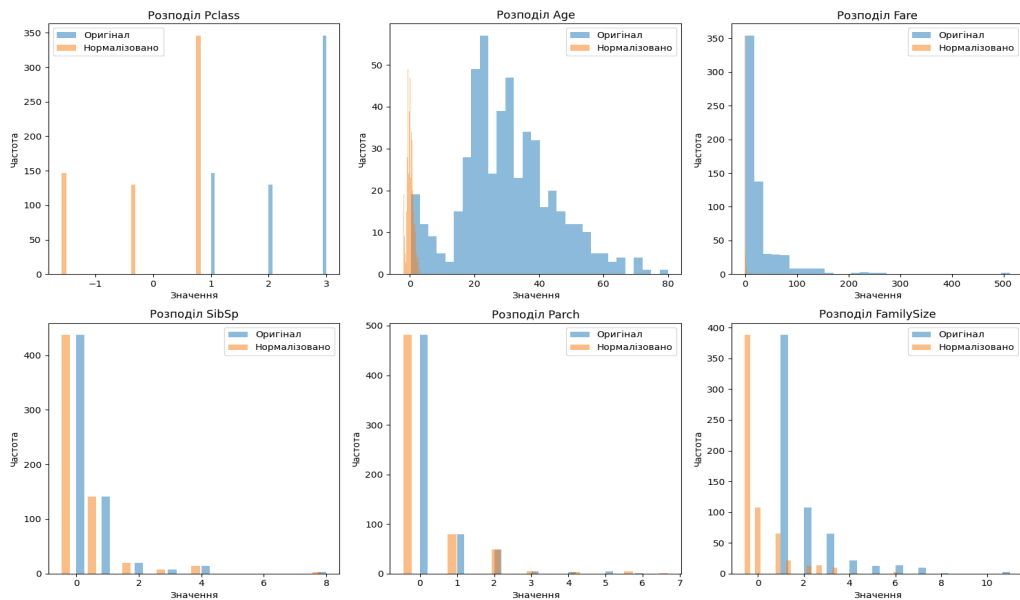


Рис. 3. Розподіл даних до та після нормалізації

Після обробки даних було проведено розподіл вибірки на навчальну та тестову, включаючи перемішування даних для уникнення упередженості в розподілі класів. Збереження пропорцій класів у вибірках забезпечило коректну оцінку моделей на невідомих даних. Це гарантувало, що пропорції класів у навчальній вибірці відповідають початковому датасету, що є критично важливим для надійних результатів моделювання.

Для оцінки важливості ознак у дослідженні застосовано методи випадковий ліс, LIME і TOPSIS. Випадковий ліс використовує ансамблевий підхід, обчислюючи важливість ознаки на основі покращення точності моделі (2):

$$Importance_j = Accuracy_{base} - Accuracy_{permuted} \quad (2)$$

де $Accuracy_{base}$ – точність моделі без перетворення ознаки; $Accuracy_{permuted}$ – точність моделі після випадкового перетворення ознаки.

Метод LIME (Local Interpretable Model-agnostic Explanations) дозволяє пояснити передбачення моделі, створюючи локальні інтерпретовані моделі. Формула для розрахунку ваги для кожної ознаки представлена у вигляді (3):

$$\omega_i = \frac{1}{\sigma_i^2} \cdot e^{-\frac{(x_i - z_i)^2}{2\sigma_i^2}}, \quad (3)$$

де x_i – значення ознаки; z_i – значення ознаки в зразку; σ_i – стандартне відхилення для ознаки i .

На рис. 4 зображено приклади зразків, які демонструють ймовірність виживання пасажирів залежно від впливу певних ознак.

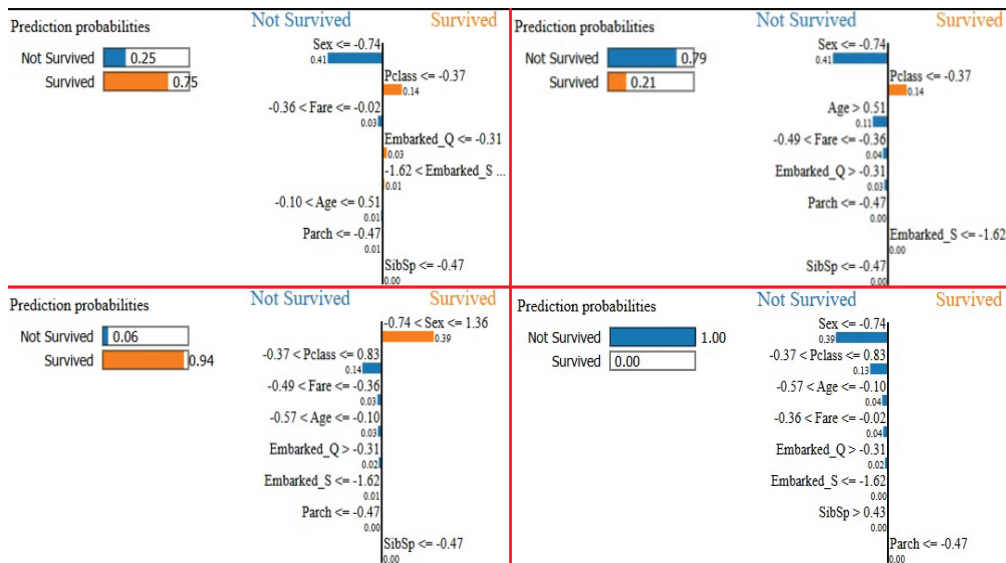


Рис. 4. Зразки передбачуваних моделей

Присутність різних ознак у зразках пояснюється їхньою здатністю впливати на ймовірність виживання залежно від контексту. Наприклад, вік може бути критично важливим у випадках, коли молодші пасажери мали вищі шанси на виживання, тоді як наявність родичів на борту може свідчити про соціальну підтримку, що також впливає на шанси виживання. Таким чином, різні ознаки можуть мати різноманітний вплив залежно від умов і характеристик пасажирів.

Аналіз чутливості, проведений шляхом зміни ваг ознак, виявив вплив цих змін на точність моделей. Моделі, такі як логістична регресія та SVM, показали стабільну точність, тоді як дерева рішень (рис. 5) та випадковий ліс (рис. 6) продемонстрували різну точність.

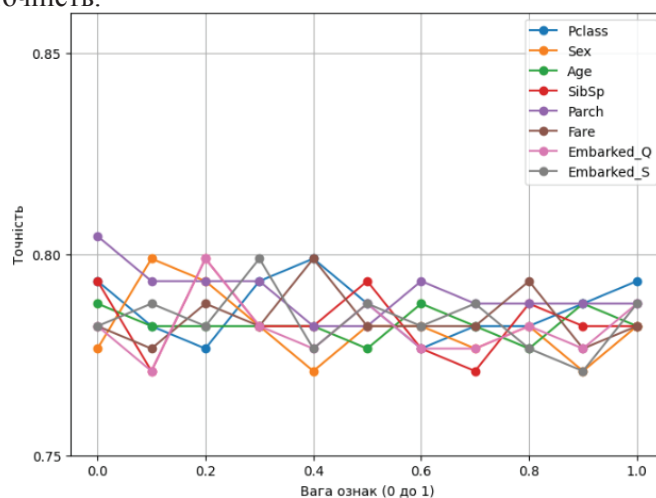


Рис. 5. Аналіз чутливості для дерева рішень

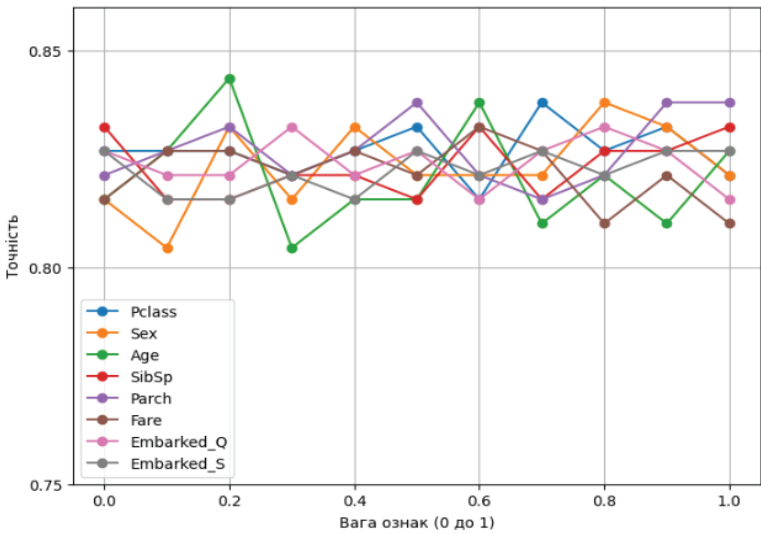


Рис. 6. Аналіз чутливості для випадкового лісу

Боксплот розкриває статистичні характеристики TOPSIS-індексів: медіану, квартилі та викиди. Кольорове групування за статусом виживання підкреслює диференціацію життєвих шансів пасажирів (рис. 7).

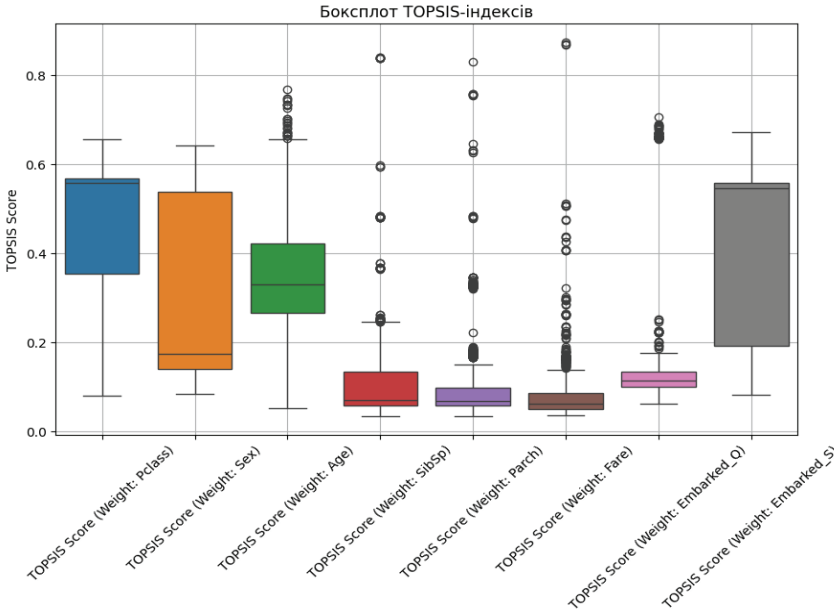


Рис. 7. Боксплот TOPSIS-індексів

Важливим етапом стало вивчення кореляцій між критеріями (рис. 8). Теплова карта демонструє взаємозв'язки між ознаками, виявляючи приховані залежності й потенційно значущі фактори впливу на виживання пасажирів.

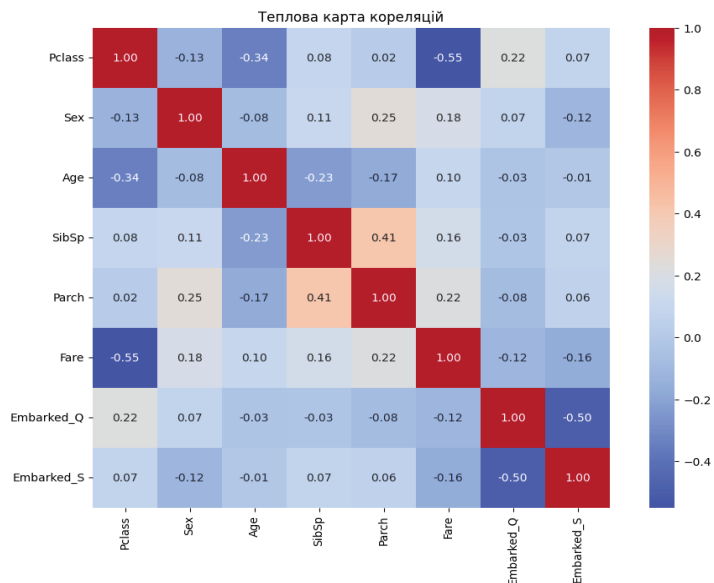


Рис. 8. Теплова карта кореляцій між критеріями

Завершальним етапом стало порівняння прогнозованих результатів із фактичними даними про виживання пасажирів лайнера. Аналіз тестової вибірки, яка містила 268 записів, показав точність моделі на рівні 86%. Порівняння моделей класифікації продемонструвало, що випадковий ліс виявився найефективнішим із AUC на рівні 0.86. Логістична регресія досягла AUC – 0.84, а SVM – 0.83, тоді як дерево рішень має найменш ефективний результат із AUC – 0.8.

ROC-крива підтвердила перевагу використання випадкового лісу у розділенні класів виживання пасажирів (рис. 9). Отримані результати свідчать про потенціал удосконалення моделі шляхом залучення додаткових ознак і використання складніших алгоритмів машинного навчання.

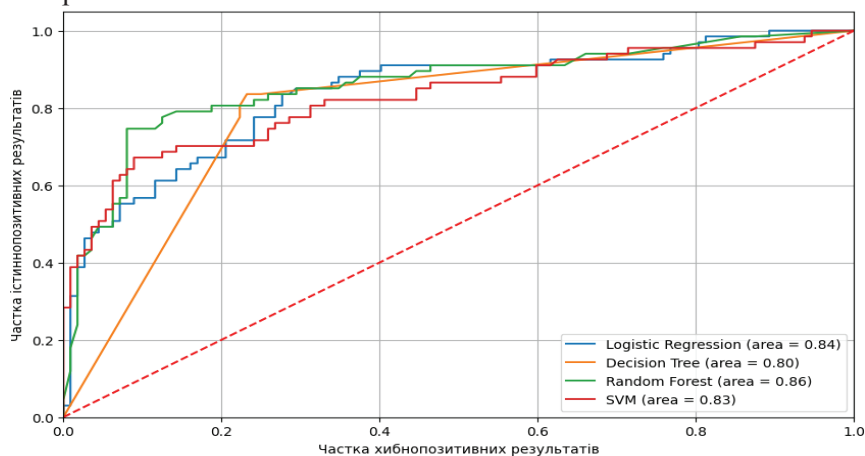


Рис. 9. ROC-криві для різних моделей

Висновки та перспективи подальшого дослідження. Результати дослідження підтверджують ефективність багатокритеріального аналізу для оцінки чинників виживання пасажирів лайнера «Титанік». Застосування методів випадкового лісу, LIME та TOPSIS дозволило ідентифікувати найбільш значущі ознаки, серед яких критично важливими виявилися стать пасажирів (0.2785), вартість квитка (0.2634) та вік (0.2508). Інші фактори, такі як клас, кількість родичів на борту та місце посадки, мають менший вплив.

Попередня обробка даних, включно з нормалізацією та розподілом вибірки на навчальну і тестову, забезпечила точність та надійність моделей. Випадковий ліс показав найвищу продуктивність із AUC 0.86, що підтверджено ROC-кривими, тоді як логістична регресія (AUC 0.84) та SVM (AUC 0.83) також продемонстрували стабільні результати. Дерево рішень виявилось менш ефективним (AUC 0.8).

Важливим аспектом дослідження став аналіз чутливості моделей та вивчення кореляцій між ознаками, візуалізованих у вигляді теплової карти. Ці методи дозволили краще зрозуміти взаємозв'язки між факторами та їх вплив на результати прогнозування. Зокрема, топологічний аналіз за допомогою TOPSIS-індексів виявив диференціацію життєвих шансів пасажирів, що відображено на боксплоті.

Практичне значення отриманих результатів виходить за межі даної задачі: запропоновані підходи можуть бути адаптовані для аналізу інших кризових ситуацій, сприяючи створенню інтегрованих систем підтримки прийняття рішень. Подальші дослідження можуть зосередитися на впливі соціально-економічних факторів на виживання пасажирів, а також на розробці інтерактивних платформ для візуалізації результатів багатокритеріального аналізу.

Практичне значення отриманих результатів виходить за межі даної задачі: запропоновані підходи можуть бути адаптовані для аналізу інших кризових ситуацій, сприяючи створенню інтегрованих систем підтримки прийняття рішень. Подальші дослідження можуть зосередитися на впливі соціально-економічних факторів на виживання пасажирів, а також на розробці інтерактивних платформ для візуалізації результатів багатокритеріального аналізу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Ai Yu. Predicting Titanic Survivors by Using Machine Learning. *Highlights in Science, Engineering and Technology*. 2023. № 34. С. 360–367. DOI: <https://doi.org/10.54097/hset.v34i.5494>.
2. Zhang Y., Haghani A. A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*. 2015. № 58. С. 308–324.
3. Titanic Disaster Prediction Based on Machine Learning Algorithms. URL: https://www.researchgate.net/publication/370569058_Titanic_Disaster_Prediction_Based_on_Machine_Learning_Algorithms.
4. Морохович В. С., Лях І. М., Хом'як М. І., Морохович Б. В. Прогнозування пасажирів, які вижили під час катастрофи «Титаніка», за допомогою дерева прийняття рішень. *Computer-integrated technologies: education, science, production*. 2024. № 55. С. 161–166.

5. Singh A., Saraswat S., Faujdar N. Analyzing Titanic disaster using machine learning algorithms. *International Conference on Computing, Communication and Automation (ICCCA)*. 2017. C. 406–411. URL: <https://doi.org/10.1109/CCAA.2017.8229835>.
6. An Overview of Multi-Criteria Decision Analysis (MCDA) Application in Managing Water-Related Disaster Events. URL: <https://www.mdpi.com/2073-4441/13/10/1358>.
7. Haque M.A., Shivaprasad G., Guruprasad G. Passenger data analysis of Titanic using machine learning approach in the context of chances of surviving the disaster. *IOP Conference Series: Materials Science and Engineering*. 2021. № 1065(1). URL: <https://doi.org/10.1088/1757-899X/1065/1/012042>.
8. Singh K., Nagpal R., Sehgal R. Exploratory Data Analysis and Machine Learning on Titanic Disaster Dataset. *10th International Conference on Cloud Computing, Data Science and Engineering*. 2020. C. 1–6.
9. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*. 2018. C. 6638–6648.
10. Analysis of Titanic Disaster using Machine Learning Algorithms. URL: https://www.researchgate.net/publication/353352089_Analysis_of_Titanic_Disaster_using_Machine_Learning_Algorithms.
11. Bogle S., Potter W. A machine learning predictive model for the Jamaica frontier market. *Proceedings of the 2015 International Conference of Data Mining and Knowledge Engineering*. 2015.
12. Платформа Kaggle. URL: <https://www.kaggle.com/>.
13. Датасет Титаніку на платформі Kaggle. URL: <https://www.kaggle.com/datasets/yasserh/titanic-dataset>.

REFERENCES

1. Yu, A. (2023). Predicting Titanic Survivors by Using Machine Learning, *Highlights in Science, Engineering and Technology*, 34, 360-367, <https://doi.org/10.54097/hset.v34i.5494>.
2. Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction, *Transportation Research Part C: Emerging Technologies*, 58, 308-324.
3. Titanic disaster prediction based on machine learning algorithms. (n.d.). *ResearchGate*. https://www.researchgate.net/publication/370569058_Titanic_Disaster_Prediction_Based_on_Machine_Learning_Algorithms.
4. Morokhovych, V., Liakh, I., Khomyak, M., & Morokhovych, B. (2024). Prohnozuvannia pasazhyriv, yaki vyzhyly pid chas katastrofy «Titanika», za dopomohoiu dereva pryiniattia rishen. *Computer-integrated technologies: education, science, production*, (55), 161-166.
5. Singh, A., Saraswat, S., & Faujdar, N. (2017). Analyzing Titanic disaster using machine learning algorithms. In *2017 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 406–411). <https://doi.org/10.1109/CCAA.2017.8229835>.
6. An overview of multi-criteria decision analysis (MCDA) application in managing water-related disaster events. (n.d.). *MDPI Water*. <https://www.mdpi.com/2073-4441/13/10/1358>.
7. Haque, M. A., Shivaprasad, G., & Guruprasad, G. (2021). Passenger data analysis of Titanic using machine learning approach in the context of chances of surviving the disaster. *IOP*

- Conference Series: Materials Science and Engineering*, 1065(1). <https://doi.org/10.1088/1757-899X/1065/1/012042>.
8. Singh, K., Nagpal, R., & Sehgal, R. (2020). Exploratory data analysis and machine learning on Titanic disaster dataset. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 1–6).
 9. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems* (pp. 6638–6648).
 10. Analysis of Titanic disaster using machine learning algorithms. (n.d.). *ResearchGate*. https://www.researchgate.net/publication/353352089_Analysis_of_Titanic_Disaster_using_Machine_Learning_Algorithms.
 11. Bogle, S., & Potter, W. (2015). A machine learning predictive model for the Jamaica frontier market. In *Proceedings of the 2015 International Conference of Data Mining and Knowledge Engineering*.
 12. Kaggle. (n.d.). *Kaggle*. <https://www.kaggle.com/>.
 13. YasserH. (n.d.). *Titanic dataset*. Kaggle. <https://www.kaggle.com/datasets/yasserh/titanic-dataset>.

doi: 10.32403/1998-6912-2025-1-70-49-59

MODELING THE SURVIVAL OF TITANIC PASSENGERS USING MULTICRITERIA ANALYSIS AND MACHINE LEARNING METHODS

I. M. Liakh, V. S. Morokhovych, R. V. Litvynchuk, A. Yu. Tsipino

*Uzhhorod National University,
square Narodna, 3, Uzhhorod, 88000, Ukraine
igor.lyah@uzhnu.edu.ua; vasyi.morokhovych@uzhnu.edu.ua; roman.litvynchuk@uzhnu.edu.ua; tsipino.artemii@student.uzhnu.edu.ua*

The article explores modeling the probability of survival for Titanic passengers using machine learning and multicriteria decision analysis (MCDA) methods. The aim of the study is to identify key factors that influenced survival outcomes and to develop a predictive model with high classification accuracy.

The analysis is based on an open dataset from the Kaggle platform, which contains detailed information on passenger age, gender, ticket class, fare, family ties, and other relevant features. A complete data preprocessing cycle was carried out, including outlier detection, normalization, encoding of categorical variables, and splitting the dataset into training and testing sets.

The study applied a range of machine learning algorithms: logistic regression, decision trees, support vector machines (SVM), random forest, and gradient boosting. These models were evaluated based on standard metrics to determine their predictive performance. To interpret model outputs and assess feature importance, the LIME and TOPSIS

methods were used, allowing for a more transparent evaluation of which variables most influenced survival.

The findings confirm the critical role of socio-demographic factors in survival likelihood, particularly gender, age, and ticket class. Women, children, and first-class passengers had notably higher chances of survival. These insights mirror historical realities and have practical implications for modern risk modeling, especially in high-stakes environments where quick and accurate decision-making is essential.

Integrating MCDA with machine learning enhances both interpretability and reliability of predictive models. The approach can be valuable in designing decision-support systems for transport safety, disaster response, and other areas where forecasting human outcomes is crucial.

Keywords: multicriteria analysis, machine learning, data analysis, forecasting, support vector method, random forest.

Стаття надійшла до редакції 30.05.2025.

Received 30.05.2025.