DOI: https://doi.org/10.36910/6775-2524-0560-2024-55-20
УДК 004.651.4:629.5
Морохович Василь Степанович, к.ф.-м.н., доцент
https://orcid.org/0000-0002-4939-6566
Лях Ігор Михайлович, к.т.н., доцент
https://orcid.org/0000-0001-5417-9403
Хом'як Максим Ігорович, студент
https://orcid.org/0009-0003-6769-7009
Морохович Богдан Васильович, магістр
https://orcid.org/0000-0002-3498-6547
Ужгородський національний університет, м. Ужгород, Україна

ПРОГНОЗУВАННЯ ПАСАЖИРІВ, ЯКІ ВИЖИЛИ ПІД ЧАС КАТАСТРОФИ «ТИТАНІКА», ЗА ДОПОМОГОЮ ДЕРЕВА ПРИЙНЯТТЯ РІШЕНЬ

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. У статті запропоновано застосування методу дерева прийняття рішень для прогнозування пасажирів, які вижили в результаті катастрофи лайнера «Титанік». У якості вхідних даних використано набори даних «Titanic – Machine Learning from Disaster», що наявні у відкритому доступі на платформі Kaggle. Дерева прийняття рішень добре підходять для вирішення задач класифікації, а їхня простота інтерпретації робить цей метод кращим вибором серед інших алгоритмів машинного навчання. У навчальному наборі даних було виконано модифікацію, щоб заповнити відсутні значення. Оцінку розподілу якісних і кількісних ознак даних та пошуку в них закономірностей здійснено за допомогою візуального аналізу даних, що дало змогу виділити ознаки пасажирів, які корелюють з їхнім виживанням найбільше, і відповідно покращити набір даних. Дерево прийняття рішень для кінцевого набору даних побудовано за допомогою бібліотеки scikit-learn (sklearn), яка надає потужні інструменти для машинного навчання в Python. Точність побудованого дерева рішень становить 77% на відкладеній вибірці. Подальше дослідження застосування дерева прийняття рішень для даного набору даних може бути проведено шляхом використання методу налаштування гіперпараметрів дерева (hyperparameter tuning), який допоможе покращити точність побудованого дерева рішень.
Ключові слова: машинне навчання, аналіз даних, дерева прийняття рішень, прогнозування, платформа Kaggle.

Morokhovych V., Liakh I., Khomyak M., Morokhovych B. Predicting passengers who survived the Titanic disaster using a decision tree. The article proposes the use of the decision tree method for predicting the passengers who survived the Titanic liner disaster. The Titanic - Machine Learning from Disaster dataset, which is publicly available on the Kaggle platform, is used as input. Decision trees are well suited for solving classification and forecasting problems, and their ease of interpretation makes this method the best choice among other machine learning algorithms. On training data set, modification has been performed to fill the missing values. The distribution of qualitative and quantitative data features and the search for patterns in the data were evaluated using visual data analysis, which allowed us to identify the passenger features that correlate with their survival the most and improve the data set accordingly. The decision tree for the final dataset was built using the scikit-learn library (sklearn), which provides powerful tools for machine learning Python. The accuracy of the built decision tree is 77% of the deferred sample. Further study of the application of the decision tree for this dataset can be done by using the hyperparameter tuning method, which will help to improve the accuracy of the constructed decision tree.
Keywords: machine learning, data analysis, decision trees, forecasting, Kaggle platform.

Постановка проблеми. Катастрофа «Титаніка» сталася більше 100 років тому, але все ще приваблює дослідників зрозуміти та вивчити, як одні пасажири вижили, а інші загинули. Лайнер «Титанік» був найбільшим судном свого часу. Його трюм складався з шістнадцяти частин, навіть повне затоплення чотирьох із яких не могло призвести до його затоплення [7]. Не дивлячись на високі стандарти безпеки, які вкладали в цей корабель його конструктори, однак несприятливі події призвели до зіткнення «Титаніка» з айсбергом і подальшої жахливої катастрофи, яка забрала життя багатьох людей. Не зважаючи на давнину подій, дослідження даних про катастрофу продовжується і понині, в тому числі із використанням методів штучного інтелекту та машинного навчання. Різні атрибути пасажирів, такі як стать, вік, категорія, до якої вони належать, і їхній соціальний клас тощо, дозволяють побудувати прогностичні моделі, забезпечуючи якісну базу даних для аналізу. Дані дослідження надають можливість передбачити виживання пасажирів на «Титаніку» за допомогою різних методів на основі даних платформи Kaggle «Titanic – Machine Learning from Disaster».

Аналіз останніх досліджень і публікацій. Дослідження прогнозування розподілу виживання пасажирів під час катастрофи «Титаніка» із використанням набору даних Kaggle проводились науковцями, використовуючи різні алгоритми машинного навчання. Приміром, для прогнозування рівня виживання пасажирів лайнера дослідники застосовували різні алгоритми,

включаючи логістичну регресію, *k*-найближчих сусідів, метод опорних векторів, «випадкового лісу» (random forest), штучних нейромереж та дерева рішень [4, 5]. У роботі [3] була проведена класифікація з двома класами (пасажири, які вижили і не вижили) за допомогою дерева рішень, де виживання аналізувалося на кожному рівні. Кластеризація виконувалась за допомогою алгоритму машинного навчання KMeans, а його реалізацію виконано за допомогою програмування Python. Науковцями встановлено, що пасажири, які подорожували з невеликою сім'єю, яка складала від 2 до 4 осіб, мали більше шансів на виживання. У статті [2] розглянуто створення систем прогнозування на основі методу опорних векторів (SVM). Дослідження полягало в тому, щоб побудувати серію моделей машинного навчання з точністю f-вимірювання понад 80% на заданому наборі демографічної інформації на Kaggle. Автором було досягнуто найкращого результату, тобто 82,82% правильних прогнозів.

Сумнозвісний інцидент затоплення найбільшого круїзного лайнера змушує дослідників і надалі заглиблюватися у масиви даних та проводити дослідницький аналіз даних, щоб зрозуміти вплив ключових параметрів на виживання людей на його борту.

Формулювання мети дослідження. Метою статті є аналіз даних про пасажирів круїзного лайнера «Титанік» та побудова дерева рішень з метою прогнозування пасажирів, які вижили під час його катастрофи, використовуючи набори даних, що наявні у відкритому доступі на платформі Kaggle.

Виклад основного матеріалу. Одним із методів розв'язання задач класифікації та прогнозування є дерева прийняття рішень, що представляє собою ієрархічну структуру наборів правил, які послідовно дають відповіді «так» або «ні». Структура дерева починається з кореневого вузла, який не має жодних вхідних гілок. Гілки, що виходять з кореневого вузла, потрапляють до внутрішніх вузлів, відомих як вузли прийняття рішень. На основі існуючих ознак обидва типи вузлів виконують оцінки і формують однорідні підмножини, які позначаються листовими вузлами. Саме вони являють собою всі можливі результати набору даних.

Метод дерева прийняття рішень має такі переваги над іншими алгоритмами машинного навчання: їх легше інтерпретувати, може працювати з якісними та кількісними ознаками, не потребує попередньої нормалізації даних.

Kaggle є відомою онлайн-платформою для спільної роботи у сфері аналізу даних, машинного навчання і штучного інтелекту [1]. Вона надає користувачам можливість обмінюватися знаннями, виконувати проекти та приймати участь у них, які пов'язані з розробкою моделей машинного навчання. Однією із поставлених задач є «Titanic – Machine Learning from Disaster», яка є навчальною.

Набір даних на платформі Kaggle складається з двох груп: дані для навчання моделі та для її тестування. Структура цих даних однакова, за виключенням відсутності ознаки виживання пасажирів в тестових даних (рис. 1).

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Рис. 1. Набір даних пасажирів «Титаніка» на платформі Kaggle

Загалом навчальний набір даних містить інформацію про 814 пасажирів та 12 ознак. Також із цих даних за необхідності можна одержати інформацію про середні значення, стандартні відхилення та інші статистичні характеристики ознак.

Для подальшого використання даних необхідно провести їх очищення, тобто позбутися неінформативних ознак, розібратися з відсутніми значеннями в даних і модифікувати їх [6].

Суттєвою проблемою є наявність в наборі даних відсутніх значень, а саме – інформації про вік деяких пасажирів. Цей недолік було усунено шляхом заповнення відсутніх даних середнім значенням віку пасажирів. На рис. 2 представлено набір даних після їх очищення.

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	1	22.0	1	0	7.2500	0
1	1	1	0	38.0	1	0	71.2833	1
2	1	3	0	26.0	0	0	7.9250	0
3	1	1	0	35.0	1	0	53.1000	0
4	0	3	1	35.0	0	0	8.0500	0

Рис. 2. Набір даних після їх очищення

Перед застосуванням методів машинного навчання є корисним застосування методу візуального аналізу даних для досліджуваного набору даних. Цей метод полягає в аналізі вигляду розподілів ознак даних та пошуку в них закономірностей. Це дозволить виділити ознаки пасажирів, які корелюють з їхнім виживанням найбільше, і відповідно покращити набір даних, що призведе до покращення результатів передбачення дерева рішень.

Для оцінки розподілу якісних та кількісних ознак побудовано їх графіки (рис. 3, 4).

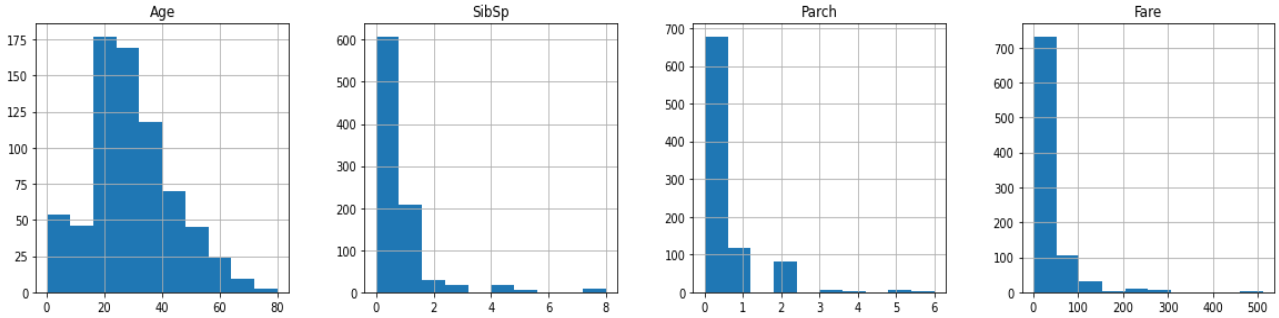

Рис. 3. Гістограми кількісних ознак даних

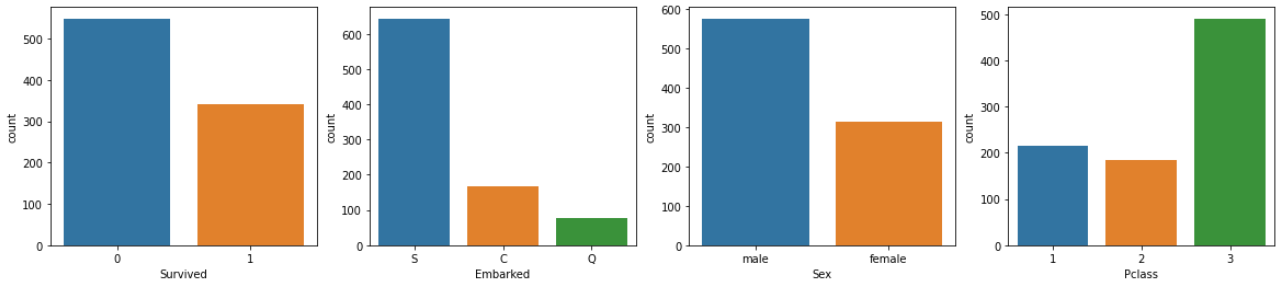

Рис. 4. Гістограми якісних ознак даних

Аналізуючи графіки, можна зазначити, що тільки вік пасажирів (Age) розподілений нормально на відміну від інших ознак, таких як: стать (Sex), кількість братів і сестер/подружжя (SibSp), порт посадки (Embarked), вартість квитка (Fare), клас квитка (Pclass), кількість батьків/дітей на борту (Parch), кількість пасажирів, які вижили (Survived).

Гістограми вказують, що в ознаках статі та кількості пасажирів, які вижили, майже є баланс, проте в ознаці «порти посадки» домінує порт Cherbourg, а серед класу квитків пасажирів – найбільше 3-го класу. Такий дисбаланс у двох ознаках міг би негативно вплинути на навчання моделі ШІ для застосування її на великих наборах даних.

Отже, в наборі даних є всього вісім дійсно корисних для аналізу ознак, чотири із яких – кількісні, а інші чотири – якісні, причому однією із якісних є цільова ознака пасажирів, які вижили. Саме для цієї ознаки необхідно визначити фактори, які найбільше на неї впливають. Порівнювати між собою всі ознаки одночасно не можливо, тому є необхідність побудувати графіки для різних типів ознак. Нижче побудовано графік типу «boxplot», який дозволяє виявляти зв'язки між якісною ознакою пасажирів, які вижили, та кількісними ознаками (Age, SibSp, Parch, Fare) (рис. 5).

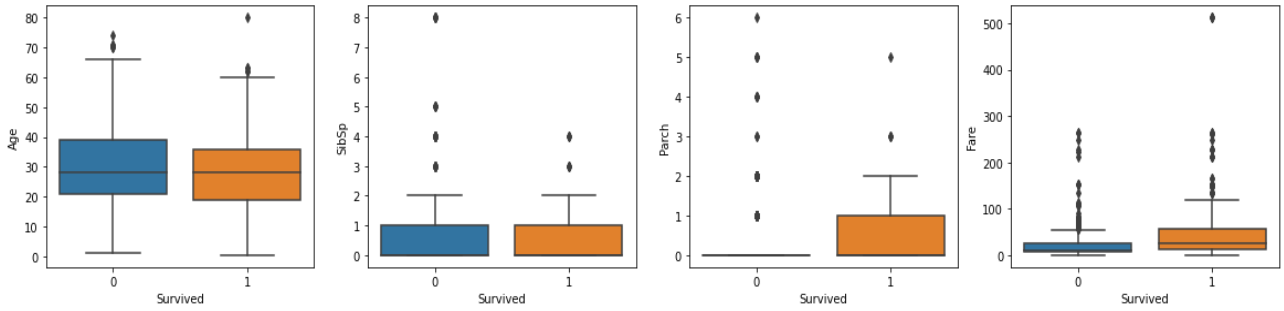

Рис. 5. Графіки «boxplot» порівняння кількісних ознак з ознакою пасажирів, які вижили

На основі отриманих графіків, можна зробити висновок, що вік пасажирів незначно впливає на їх виживання – у молодших пасажирів трохи більші шанси вижити. Ознака «кількість братів і сестер/подружжя» не впливає на виживання пасажирів. Наявність у пасажира батьків/дітей на борту сильно корелюється з ознакою пасажирів, які вижили, а також, вартість квитка пасажира корелюється з його виживанням.

Для аналізу якісної ознаки виживання пасажирів з іншими якісними ознаками побудовано наступні гістограми (рис. 6).

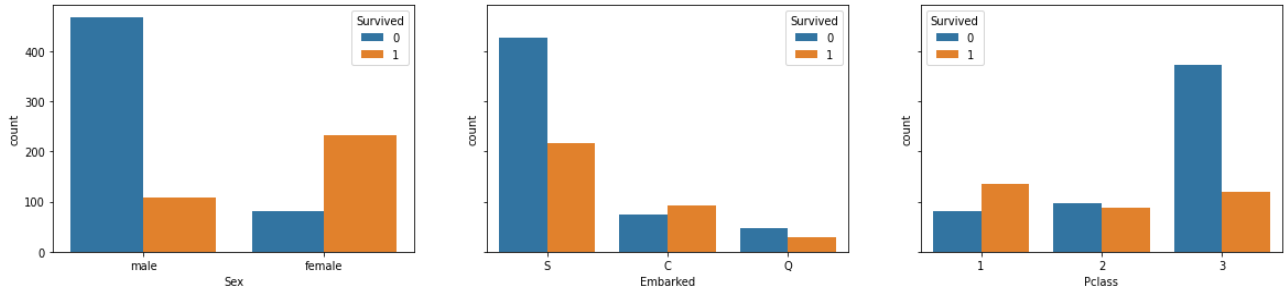

Рис. 6. Гістограми порівняння якісних ознак з ознакою виживання пасажирів
(0 – пасажири, які загинули; 1 – пасажири, які вижили)

Проаналізувавши гістограми, можна відмітити, що стать пасажира сильно впливає на його виживання – серед жінок набагато більше змогли врятуватися. Найбільше загинуло людей серед пасажирів третього класу. Цікавим є графік з портами посадки пасажирів – серед пасажирів, які сіли в порту «Southampto», найбільше загиблих, тобто ознака порту посадки пасажирів також виявилась важливою. Тому доцільно побудувати гістограми для порівняння порту посадки пасажирів та їх класу квитків і статі (рис. 7).

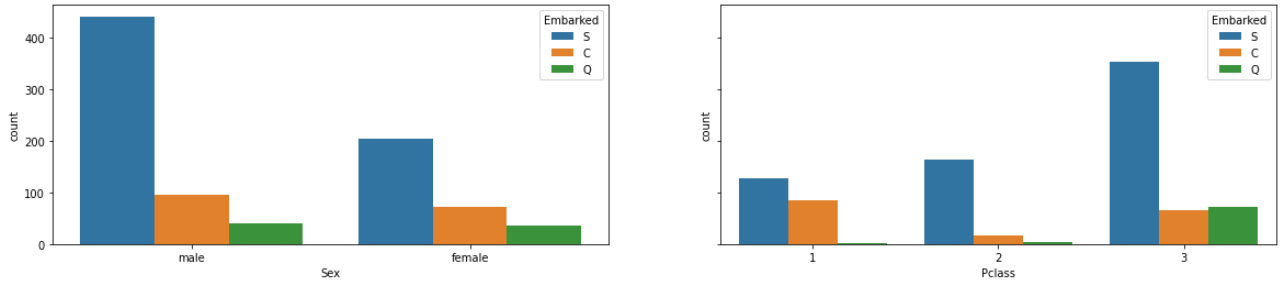

Рис. 7. Порівняння портів посадки пасажирів з їх класом квитків та статтю

З рис. 7 видно, що в порту «Southampto» на лайнер «Титанік» сіли, в основному, чоловіки та пасажири з квитками третього класу. Тому буде доцільно вилучити ознаку порту посадки з набору даних, оскільки вона є зайвою. Також було вилучено ознаку класу квитка, оскільки вона прямо корелює з його вартістю.

Переконавшись, що дані переведено в числову форму і в них немає відсутніх значень, можна перейти до побудови дерева рішень для прогнозування цільової ознаки, а саме пасажирів, які вижили.

Максимальна глибина побудованого дерева дорівнює трьом, а його ефективність вимірювалась за допомогою метрики точності та становить 77%. Графічне представлення дерева рішень наведено на рис. 8.

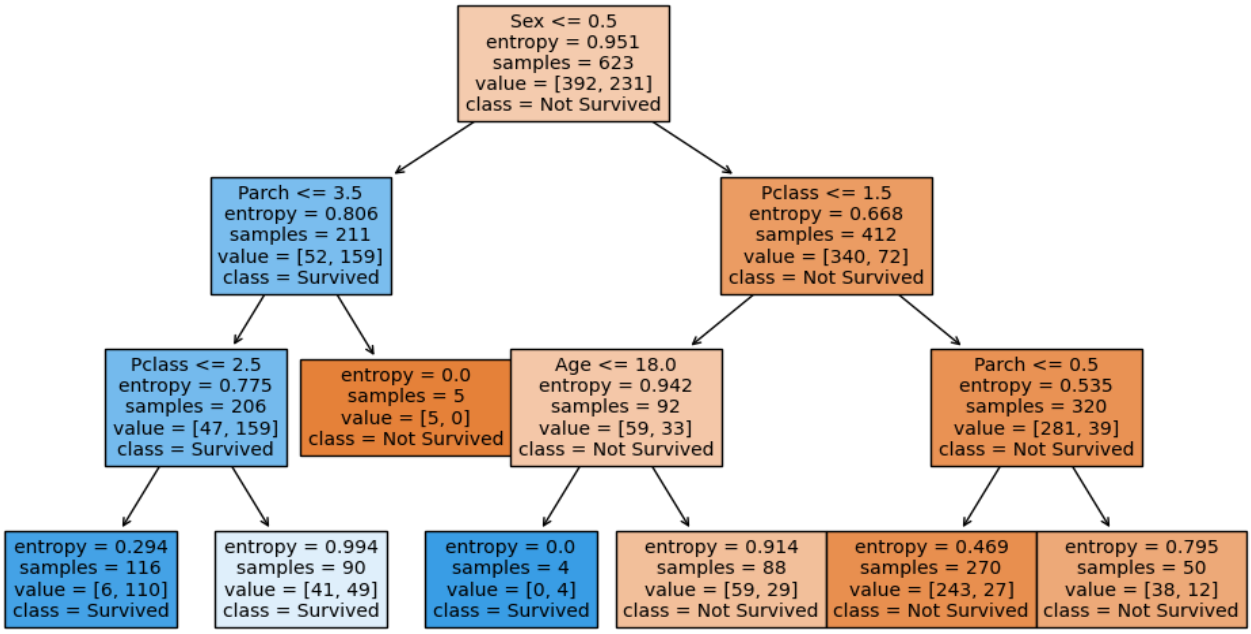

Рис. 8. Дерево прийняття рішень для набору даних

Дерево прийняття рішень для кінцевого набору даних було побудовано за допомогою бібліотеки scikit-learn (sklearn), яка надає потужні інструменти для машинного навчання в Python. У бібліотеці sklearn існує клас DecisionTreeClassifier, який дозволяє побудувати дерево рішень для класифікації, тобто передбачення категоріальної цільової ознаки.

Висновки та перспективи подальшого дослідження. У результаті проведеного дослідження було проаналізовано дані про пасажирів круїзного лайнеру «Титанік», які перебували на ньому під час його катастрофи 14 квітня 1912 року. Дані про пасажирів було очищено та використано для побудови дерева рішень з метою прогнозування пасажирів, які вижили. Точність побудованого дерева рішень становить 77% на відкладеній вибірці. Таким чином можна зробити висновок, що дерева прийняття рішень добре підходять для розв'язування задач класифікації, а легкість їх інтерпретації робить цей метод пріоритетним вибором серед інших алгоритмів машинного навчання, коли необхідне чітке розуміння, як саме приймається певне рішення.

Подальше дослідження застосування дерева рішень на даному наборі даних може бути проведено шляхом покращення точності вже побудованого дерева рішень хоча б до значення 80% на відкладеній вибірці. Метод, який допоможе покращити точність побудованого дерева рішень, є налаштування гіперпараметрів дерева (hyperparameter tuning). Наприклад, при побудові дерева рішень можна вказати такі гіперпараметри, як максимальна глибина дерева та максимальна кількість ознак набору даних, які враховуються алгоритмом при побудові чергових вузлів дерева.

Список бібліографічного опису

1. Лук'янюк В. (2017). Цей день в історії: Загибель «Титаніка». URL: https://www.jnsm.com.ua/h/0414M/.
2. Ai Yu. (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.
3. Haque MA, 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.
4. Singh A., Saraswat S. & Faujdar N. (2017). Analyzing Titanic disaster using machine learning algorithms, *International Conference on Computing, Communication and Automation (ICCCA)*, Greater Noida, India, 406-411, https://doi.org/10.1109/CCAA.2017.8229835.
5. Singh K., Nagpal R. & Sehgal R. (2020). Exploratory Data Analysis and Machine Learning on Titanic Disaster Dataset, *10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 320-326, https://doi.org/10.1109/Confluence47617.2020.9057955.

6. Song YY, Lu Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry*, Vol. 27(2), 130-135, https://doi.org/10.11919/j.issn.1002-0829.215044.
7. Titanic – Machine Learning from Disaster. URL: https://www.kaggle.com/c/titanic/data.

References

1. Lukyaniuk, V. (2017). This day in history: The sinking of the Titanic. URL: https://www.jnsm.com.ua/h/0414M/.
2. Ai, Yu. (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.
3. Haque, MA, 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.
4. Singh, A., Saraswat, S. & Faujdar, N. (2017). Analyzing Titanic disaster using machine learning algorithms, *International Conference on Computing, Communication and Automation (ICCCA)*, Greater Noida, India, 406-411, https://doi.org/10.1109/CCAA.2017.8229835.
5. Singh, K., Nagpal, R. & Sehgal, R. (2020). Exploratory Data Analysis and Machine Learning on Titanic Disaster Dataset, *10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 320-326, https://doi.org/10.1109/Confluence47617.2020.9057955.
6. Song, YY, Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry*, Vol. 27(2), 130-135, https://doi.org/10.11919/j.issn.1002-0829.215044.
7. Titanic – Machine Learning from Disaster. URL: https://www.kaggle.com/c/titanic/data.