УДК 004.9

В. С. Морохович, к.ф.-м.н., доц.; М. І. Хом'як
(ДВНЗ «Ужгородський національний університет», Україна)

ВИКОРИСТАННЯ ДЕРЕВА РІШЕНЬ ДЛЯ РОЗВ'ЯЗАННЯ ЗАДАЧ КЛАСИФІКАЦІЇ НА ПРИКЛАДІ НАБОРУ ДАНИХ ПАСАЖИРІВ «ТИТАНІК»

V. S. Morokhovych, Ph.D., Assoc. Prof.; M. I. Khomyak

USING A DECISION TREE TO SOLVE CLASSIFICATION PROBLEMS ON THE EXAMPLE OF THE TITANIC PASSENGER DATA SET

Більше століття тому сталася одна з найбільших морських катастроф в історії людства. Велетенський круїзний лайнер «Титанік» під час свого першого рейсу затонув біля берегів Ньюфаундленду в Північній Атлантиці після того, як його корпус отримав значні пошкодження від зіткнення з айсбергом. З 2224 осіб, які перебували на борту, лише 712 людей змогли врятуватись. Корабель «Титанік» був найбільшим судном свого часу. Його трюм складався з шістнадцяти частин, навіть повне затоплення чотирьох із яких не могло призвести до його затоплення [1]. Не дивлячись на високі стандарти безпеки, які вкладали в цей корабель його конструктори, велика черга несприятливих подій призвела до зіткнення «Титаніка» з айсбергом і подальшої жахливої катастрофи. Незважаючи на давнину події, дослідження щодо «Титаніка» продовжуються і сьогодні, в тому числі із використанням методів штучного інтелекту та машинного навчання.

Під час дослідження було використано набори даних пасажирів лайнера «Титанік», що наявні у відкритому доступі на платформі Kaggle [2]. Набір даних складається з двох груп: дані для навчання моделі та дані для її тестування. Структура цих даних однакова, за виключенням відсутності ознаки виживання пасажира в тестових даних. Загалом 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 показники ознак.

Для подальшого використання даних необхідно провести їх очищення: позбутися неінформативних ознак, розібратися з відсутніми значеннями в даних та модифікувати дані, які цього потребують. Проаналізувавши набір даних, можна відкинути деякі ознаки пасажирів, які не несуть ніякої корисної інформації про ймовірність їхнього виживання під час катастрофи. Такими ознаками є: імена пасажирів, номери їхніх кабін і квитків, ідентифікатори пасажирів. Також суттєвою проблемою в наборі даних є відсутніх значень, тобто інформації про вік деяких пасажирів. Це питання було виправлено заповненням відсутніх даних середнім значенням віку пасажирів. Набір даних після їх очищення наведено на рис. 1.

	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

Рисунок 1. Набір даних після їх очищення

Після того, як дані приведено в належну форму і немає відсутніх значень, можна переходити до побудови дерева рішень для передбачення цільової ознаки.

Дерево рішень є одним із загальновживаних методів розв'язання задач класифікації та прогнозування [3]. Даний метод являє собою ієрархічну структуру і починається з кореневого вузла, який не має жодних вхідних гілок. Вихідні гілки з кореневого вузла потім потрапляють у внутрішні вузли, також відомі як вузли прийняття рішень. На основі наявних ознак обидва типи вузлів проводять оцінки для формування однорідних підмножин, які позначаються листовими вузлами або термінальними вузлами. Листові вузли представляють всі можливі результати в наборі даних. Дерева прийняття рішень мають такі переваги над іншими алгоритмами машинного навчання, як: легка інтерпретованість, здатність до роботи з якісними та кількісними ознаками, відсутність потреби в попередній нормалізації даних.

Дерево прийняття рішень для кінцевого набору даних побудовано за допомогою бібліотеки scikit-learn (sklearn), яка надає потужні інструменти для машинного навчання в мові програмування Python. Максимальна глибина побудованого дерева дорівнює трьом. Його ефективність вимірювалась за допомогою метрики точності й становить 77%. Графічне представлення дерева рішень наведено на рис. 2.

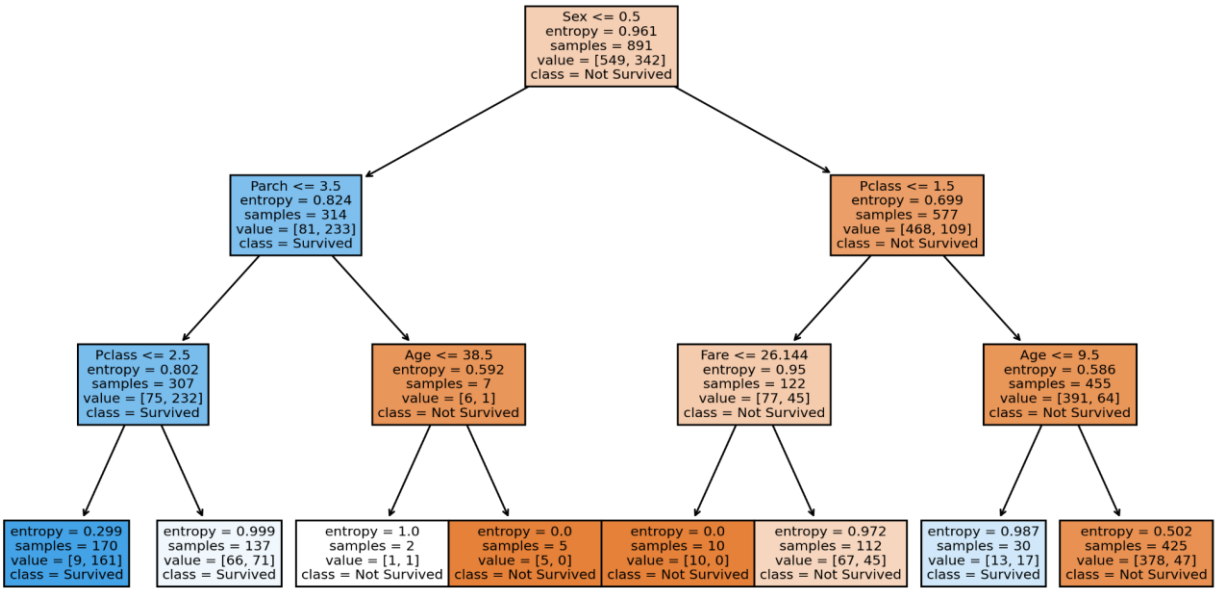

Рисунок 2. Дерево прийняття рішень для набору даних

Таким чином, дерева рішень добре підходять для розв'язування задач класифікації, а легкість інтерпретації дерев рішень робить їх пріоритетним вибором серед інших алгоритмів машинного навчання, коли необхідне чітке розуміння, як саме приймається певне рішення.

Література

1. Titanic – Machine Learning from Disaster. URL: https://www.kaggle.com/c/titanic/data
2. Цей день в історії: 14 квітня 1912: Загибель «Титаніка». URL: https://www.jnsm.com.ua/h/0414M/
3. Yan-yan Song, Ying Lu, 2015. Decision tree methods: applications for classification and prediction. Shanghai Archives of Psychiatry, Vol. 27(2), 130-135.