*AMICon-2024*

**ПОРІВНЯННЯ ОПТИМАЛЬНОСТІ АЛГОРИТМІВ КЛАСИФІКАЦІЇ ДАНИХ НА ПРИКЛАДІ ДЕРЕВ РІШЕНЬ**

*Бень Христина, Квасниця Галина*

*Львiвський нацiональний унiверситет iменi Iвана Франка*

*Факультет прикладної математики та iнформатики*

[*khrystyna200303@gmail.com*](mailto:khrystyna200303@gmail.com)

У наш час наплив інформації, завдяки можливостям її зберігання та збирання, став, водночас, винятковою цінністю та надзвичайною складністю. Велика кількість інформації та можливість її доступного використання відкрили багато нових можливостей і дали поштовх для розвитку в різних сферах. Проте, це створило таку проблему, як невміння користуватись даними правильно. Такі обсяги інформації часто видаються складними для інтерпретації. Щоб їх коректно застосовувати, потрібно вміти систематизувати дані для подальшого аналізу. Тому, необхідно вивчати різні способи їхнього впорядкування.

Актуальність теми з практичної точки зору полягає в тому, щоби досліджувати можливості класифікації даних та застосовувати для різних цілей і прикладів.

Метою даної роботи є дослідження оптимальності роботи алгоритмів для побудови дерева рішень та їхнє порівняння за точністю та швидкодією.

Для виконання дослідження необхідно, спочатку, ознайомитися із теоретичними даними теми класифікації, а саме дерева рішень.

**Класифікація** –системний розподіл досліджуваних предметів, явищ, процесів за родами, видами, типами, з якими-небудь істотними ознаками для зручності їхнього дослідження; групування вихідних понять і їхнє розташування у певному порядку, що відбиває ступінь цієї схожості [2].

**Дерево рішень** – це універсальний контрольований алгоритм машинного навчання, який використовується як для задач класифікації, так і для регресії. Дерево рішень – це структура дерева, подібна до блок-схеми, де кожен внутрішній вузол позначає функцію, гілки – це правила, а листкові вузли – результат роботи алгоритму [3].

Для аналізу такої структури, було обрано найвідоміші алгоритми побудови, а саме:

* **ID3** – алгоритм**,** що базується на виборі атрибутів для розгалуження, які мають найбільший приріст інформації, що рахується на основі ентропії [4].
* **C45** –покращена версія алгоритму ID3 з обрізанням дерева, підтримкою пропущених даних та використанням інформаційного приросту для побудови оптимального дерева рішень.
* **CART** – це тип алгоритму класифікації, який необхідний для побудови дерева рішень на основі критерію Джині [5].
* **Random Forest** –працює шляхом створення кількох дерев рішень під час фази навчання. Кожне дерево будується з використанням випадкової підмножини набору даних для вимірювання випадкової підмножини ознак у кожному розділі. Ця випадковість вводить варіативність між окремими деревами, зменшуючи ризик перенавчання та покращуючи загальну ефективність прогнозування [6].

Для аналізу результатів було використано три області визначення подані у таблиці 1 і таблиці 2, де з другого по четвертий стовпець показано співвідношення вибірки у форматі тренувальні/тестові дані:

1. класична тестова вибірка для алгоритмів сортування про вибір людини в залежності від погодних умов;
2. теоретична вибірка успішності студентів із врахуванням даних про їхні бали, додаткові активності, відвідуваність для прийняття рішення про надання стипендії;
3. вибірка із платформи Kaggle про працівників, включаючи їхню освіту, історію роботи, демографічні дані та фактори, пов’язані з роботою [1].

Важливим показником для оцінки алгоритму є точність – **Accuracy Точність** – дає загальну точність моделі, тобто частку правильно класифікованих від загальної кількості зразків. Для розрахунку використовують таку формулу:

AC = (TP+TN)/(TP+TN+FP+FN), де

* TP (True Positives) – правильно класифіковані зразки (істинно-позитивні випадки);
* TN (True Negatives) — правильно класифіковані негативні приклади (істинно-негативні випадки);
* FN (False Negatives) – позитивні зразки, класифіковані як негативні. Це помилка 1-го роду (хибно-негативні зразки);
* FP (False Positives) – негативні зразки, класифіковані як позитивні. Це помилка 2-го роду (хибно-позитивні випадки) [7].

Табл. 1. Порівняння алгоритмів за точністю

|  |  |  |  |
| --- | --- | --- | --- |
|  | 20/5 | 115/15 | 4 000/600 |
| ID3 | 1.0 | 1.0 | 0.5770992 |
| C45 | 1.0 | 1.0 | 0.5770992 |
| CART | 1.0 | 1.0 | 0.8045802 |
| Random Forest | 1.0 | 1.0 | 0.8274809 |

Наступним показником було обрано час роботи алгоритму, а саме – побудови дерева, одиниця виміру якого – секунда.

Табл. 2. Порівняння алгоритмів за часом

|  |  |  |  |
| --- | --- | --- | --- |
|  | 20/5 | 115/15 | 4 000/600 |
| ID3 | 0.50800633 | 0.161302328 | 12.50532531738281 |
| C45 | 0.47323369 | 0.135096073 | 11.00152373313904 |
| CART | 0.01834154 | 0.031525135 | 87.96786594390869 |
| Random Forest | 0.10077023 | 0.030347825 | 7.14400577545166 |

**Висновок**

У цій роботі було проведено дослідження та порівняння алгоритмів побудови дерева рішень.

Дерева рішень, є гнучкими моделями, які використовують набір правил для прийняття рішень. Вони можуть обробляти як числові, так і категоріальні ознаки, і мають високу інтерпретованість. Дерева рішень є зручними для вирішення задач класифікації зі складною структурою даних і можуть бути ефективними для виявлення важливих ознак. Однак, вони можуть схильними до перенавчання, особливо при побудові глибоких дерев, і можуть бути вразливі до шуму та змін у даних.

Якщо аналізувати результати проведені на прикладі наших вибірок, то можна виокремити такі спостереження:

* на невеликих вибірках усі алгоритми мають хорошу точність, проте CART та Random Forest швидше виконують задачу;
* на вибірках із більшою кількістю даних CART та Random Forest мають набагато кращу точність, проте CART займає суттєво довший час виконання, натомість як Random Forest є суттєво швидший за всі алгоритми.

# Список використаних джерел

1. <https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset?resource=download> – employee dataset
2. <https://moodle.znu.edu.ua/pluginfile.php/486129/mod_resource/content/1/%D0%9B%D0%B5%D0%BA%D1%86%D1%96%D1%8F%206.pdf> – Задачі Data Mining. Класифікація та кластеризація
3. <https://www.paloaltonetworks.com/cyberpedia/data-classification> – What Is Data Classification?
4. <https://medium.com/@ashirbadpradhan8115/decision-tree-id3-algorithm-machine-learning-4120d8ba013b> – Decision Tree ID3 Algorithm
5. <https://www.analyticssteps.com/blogs/classification-and-regression-tree-cart-algorithm> – A Classification and Regression Tree (CART) Algorithm
6. <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/> – Random Forest Algorithm in Machine Learning
7. <https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826> - Confusion Matrix for Your Multi-Class Machine Learning Model