*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)

**Вступ**

У наш час наплив інформації завдяки можливостям її зберігання та збирання став, водночас, винятковою цінністю і великою складністю. Велика кількість інформації, можливість її доступного використання відкрили багато нових можливостей та 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 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 та прикладів.

**Хід роботи:**

Для виконання дослідження необхідно було спочатку ознайомитись 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 їх дослідження; угрупування вихідних понять і розташування їх 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 - правила, а листові вузли - результат алгоритму [3].

Для аналізу такої структура, я обрала 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 обрізкою 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 мною було використано 4 області визначення

1. Класична тестова вибірка для алгоритмів сортування про вибір людини в залежності від погодніх 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, яка визначається співвідношенням правильно розпізнаних даних до загальної кількості елементів вибірки.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 20/5 | 100/20 | 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 |

(табл. 1)

Наступним показником я взяла час роботи алгоритму (а саме побудови дерева), одиниця виміру якого – секунда.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 20/5 | 100/20 | 4 000/600 |
| ID3 |  |  | 12.505325317382812 |
| C45 |  |  | 11.001523733139038 |
| CART |  |  | 87.96786594390869 |
| Random forest |  |  | 7.14400577545166 |

(табл. 2)

**Висновок**

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

Дерева рішень, з іншого боку, 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 добре підходять для вирішення задач класифікації зі складною структурою даних і можуть бути ефективними для виявлення важливих ознак. Однак, вони можуть схильні до перенавчання, особливо при побудові глибоких дерев, і вони можуть бути вразливі до шуму та змін в даних.

Якщо аналізувати результати проведені на прикладі моїх вибірок, то можна виокремити,

# Джерела

1. <https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset?resource=download> -
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