ДОСЛІДЖЕННЯ ТА АНАЛІЗ МЕТОДІВ ПРОГНОЗУВАННЯ. ДЕРЕВА РІШЕНЬ. ГРАДІЄНТНИЙ БУСТИНГ.

Чуприна А. С., Смикова А. Ю.
Харківський національний університет радіоелектроніки, Харків, Україна

Останні роки набирає популярність та розширює свої можливості наука про дані. За допомогою її підходів, методів та архітектури були спроектовані та реалізовані такі рішення, які ще 10 років тому не можливо було уявити. Наразі існують такі програмні рішення, які дозволяють оброблювати та аналізувати текст натуральною мовою[1], виконувати пошук за зображеннями[2] та розпізнавати образи і обличчя[3]. Наука про дані стала популярною темою, яка привертає велику увагу дослідницької спільноти. Розвиток технологій аналізу даних і доступні наукові датасети дозволяють досліджувати прогнозування на основі даних, яке відіграє ключову роль у пошуку тенденцій наукового впливу. Основними методами прогнозування є класифікація, кластеризація, пошук К-середніх, та інші. Розглянемо класифікаційні методи прогнозування більш детально.

Метою доповіді є аналіз і порівняння різних методів вирішення задачі прогнозування за допомогою підходу класифікації. До цих методів відносяться дерева рішень (tree decision), випадковий ліс (random forest) та градієнтний бустинг (gradient boosted machine). Порівняння буде виконано на основі задачі прогнозування результатів спортивних матчів. Для цього буде виконано навчання на основі відкритого датасету всіх матчів з тенісу за період з першого Вімбелдону. Прогнозування має бути з точністю не менш 85 відсотків. В доповіді наводяться результати порівняння класифікаційних методів прогнозування, наведених вище, за наступними критеріями: час навчання, точність прогнозування, схильність до перенавчання, коефіцієнт помилок, частота помилок. Отримані результати показують, що метод дерев рішень має найвищу швидкість навчання та достатньо високий показник точності на навчальному датасеті, проте низька точність прогнозування на реальних даних. Випадковий ліс у свою чергу має меншу схильність до перенавчання та показує кращі результати на тестовій виборці даних. Градієнтний бустинг є кращим рішенням для поставленої задачі, адже він містить переваги обох цих методів, але немає вагомих недоліків.

Список літератури

1. K. Smelyakov, D. Karachevtsev, D. Kulemza, Y. Samoilenko, O. Patlan. Effectiveness of preprocessing algorithms for natural language processing applications. *IEEE* - 2020. p. 187-191.

2. K. Smelyakov, A. Chupryna, D. Sandrkin, M. Kolisnyk. Search by Image Engine for Big Data Warehouse. *IEEE – 2020. P. 1-4.*

3. K. Smelyakov, A. Chupryna, O. Bohomolov, N. Hunko. The Neural Network Models Effectiveness for Face Detection and Face Recognition. *IEEE – 2021.* p. 1-7