УДК 004.89
Субботін С.О.[1], Хохлова В.С. [2]
[1] проф. ЗНТУ
[2] студ. гр. КНТ-415 ЗНТУ

ДОСЛІДЖЕННЯ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ВИРІШЕННЯ ЗАДАЧІ МЕДИЧНОЇ ДІАГНОСТИКИ

Машинне навчання – великий підрозділ штучного інтелекту, що вивчає методи побудови алгоритмів, здатних навчатися. Для машинного навчання використовують різні технології та алгоритми. Зокрема, можуть застосовуватися дискримінантний аналіз, байєсовські класифікатори, штучні нейронні мережи та багато інших математичних методів [1].

Машинне навчання з кожним днем займає все більше місце в нашому житті з огляду на величезного спектру його застосувань. Починаючи від аналізу пробок і закінчуючи самоврядними автомобілями, все більше завдань перекладається на машини. Алгоритми машинного навчання можна вважати найпотужнішим інструментом, орієнтованим на використання великих обсягів даних для прогнозування та прийняття рішень [2].

Метою даної роботи є дослідження та порівняння алгоритмів машинного навчання для вирішення задачі медичної діагностики.

У процесі виконання роботи були розглянуті та реалізовані наступні методи: лінійна регресія, метод k-найближчих сусідів, метод опорних векторів, ядрова регуляризація Тихонова (Kernel ridge regression), багатошаровий перцептрон (MLP regressor), random forest (випадковий ліс), дерева рішень.

Методи порівнювалися за такими параметрами: час побудови моделі (мс), об'єм пам'яті, витрачений на побудову моделі (МіВ), об'єм пам'яті, який займає модель (КіВ), кількість параметрів моделі, які можна налаштувати (ваги), помилка моделі для навчальної вибірки (%), помилка моделі для тестової вибірки (%).

Усі методи були реалізовані на мові програмування Python [3].

Аналізуючи результати, можемо зробити висновок, що найгірші результати дали метод опорних векторів та ядрова регуляризація Тихонова. Хоча метод опорних векторів витрачає дуже малий об'єм пам'яті на побудову моделі, він дав найбільшу помилку при тестуванні (~10%). Ядрова регуляризація Тихонова показала помилки на окремих значеннях.

Велику кількість неправильних рішень під час використання методу опорних векторів спостерігалася на графіку залежності спрогнозованих значень від реальних.

Лінійна регресія та метод k-найближчих сусідів показують гарні та середні результати по усім параметрам, мають середній відсоток помилок (~5%). Метод k-найближчих сусідів показує відносно поганий результат при тестуванні на навчальній вибірці, але порівняння з іншими методами він дає однаковий відсоток помилок на обох вибірках, тож загалом відсоток помилок є задовільним. Графік цього методу є досить лінійними.

Дерева рішень показали гарні показники майже по усім параметрам, але зробили значний відсоток помилки на тестовій виборці – один з найгірших (~8%).

Спірні результати показали багатошаровий перцептрон та random forest. По усім параметрам, окрім помилок моделі, вони поступаються іншим методам, але ці два метода показали найменший відсоток помилок як для навчальної вибірки, так і для тестової (~2,5%). Підтвердженням невеликої кількості помилок є лінійність графіку залежності спрогнозованих значень від реальних.

Таким чином, якщо ми маємо обмеження в часі побудови моделі та пам'яті, то найкраще буде обрати метод k-найближчих сусідів. Результати моделювання показуватимуть середній відсоток помилок (~5%).

У випадку достатніх ресурсів пам'яті та часу найкращим методом з протестованих для вирішення поставленої задачі є random forest, який показує найменший відсоток помилок серед усіх методів (~2,5%).

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Machine Learning – Машинне навчання [Електронний ресурс]. – Режим доступу: https://www.it.ua/knowledge-base/technology-innovation/machine-learning.

2. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных [Текст] / П. Флах. – М.: ДМК Пресс, 2015. – 775с.

3. Gorelick M. High Performance Python / M. Gorelick, I. Ozsvald. – Севастополь : ДМК Пресс, 2014. – 351 с.