рення деревоподібної її структури. Подальші дослідження доцільно провести у напрямку автоматизації планування та виконання комп'ютерного моделювання технічного об'єкта або його обраної частини.

Література:

1. IDEF Family of Methods. URL: http://www.idef.com (дата звернення: 15.09.2020).
2. Serifi V. et al. Functional and information modeling of production using IDEF methods. *Strojniški vestnik*. 2009. Т. 55. – №. 2. – С. 131–140.
3. Jeong K. Y., Wu L., Hong J. D. IDEF method-based simulation model design and development. *Journal of Industrial Engineering and Management*. 2009. – Т. 2. – №. 2. – С. 337–359.
4. Antypenko V. et al. Functional Modeling of the Means for Heat Consumption Monitoring During Its Design Using the Information. *International Conference «New Technologies, Development and Applications». Springer, Cham*. 2020. – С. 701–708.

DOI https://doi.org/10.30525/978-9934-588-79-2-1.2

ІНТЕРПРЕТАЦІЯ РЕЗУЛЬТАТІВ ПРОГНОЗУВАННЯ ОТРИМАНИХ ВИПАДКОВИМ ЛІСОМ

Бабенко В. О.
бакалавр комп'ютерних наук,
магістр кафедри біомедичної кібернетики
Національного технічного університету України
«Київський політехнічний інститут імені Ігоря Сікорського»

Носовець О. К.
кандидат технічних наук,
доцент кафедри біомедичної кібернетики
Національного технічного університету України
«Київський політехнічний інститут імені Ігоря Сікорського»

Тищенко Б. В.
бакалавр комп'ютерних наук,
магістр кафедри теорії та технології програмування
Київського національного університету імені Тараса Шевченка
м. Київ, Україна

Вступ. Машинне навчання є доволі відомим підрозділом штучного інтелекту [1], який вивчає алгоритми, що здатні навчатися самостійно. Дані алгоритми застосовуються в різноманітних задачах, пов'язаних з обробкою даних. Однією із таких задач є задача класифікації. В ній необхідно налаштувати алгоритм таким чином, щоб він зумів самостійно розбити задані об'єкти на класи. Наприклад, алгоритм для клієнтів банку (об'єктів) може визначити за їх характеристиками, чи варто давати їм кредит, чи ні (класом в даному випадку є характеристика надійності клієнта). Як правило, класи попередньо визначені (даний підхід називається «навчанням з вчителем»). Алгоритм навчається на історичних даних, після чого він здатен класифікувати навіть невідомі йому об'єкти, які можуть з'явитися в майбутньому. З розвитком машинного навчання з'являлися нові алгоритми (та їх модифікації) для вирішення подібних задач. Доволі популярним та надійним в сучасному світі Data Science вважають алгоритм випадкового лісу [2]. Тим не менш, незважаючи на те, що в більшості задач даний алгоритм показував високу точність прогнозування, залишається невирішеною проблема того, яким чином інтерпретувати результат випадкового лісу, щоб кожна людина зуміла зрозуміти причини такого результату і сформувати правильні висновки. Це важливо, оскільки в більшості випадків, де застосовується алгоритм машинного навчання, юридичну відповідальність несе саме людина, а не машина, яку вона використовує.

Постановка задачі. Цілю даного дослідження було визначити підхід для правильної інтерпретації результатів прогнозування, отриманий за допомогою випадкового лісу. Для її досягнення були поставлені наступні задачі:

1. Обрати експериментальні дані для дослідження, на базисі яких буде вирішуватися задача класифікації.

2. Побудувати випадковий ліс для обраних даних.

3. Інтерпретувати результати лісу в зручному для людини вигляді.

Основна частина. Для дослідження були обрані клінічні дані пацієнтів з вродженими даними серця, надані *Національним інститутом серцево-судинної хірургії імені М.М. Амосова*. База даних налічує 128 пацієнтів віком від 3 до 27 років, кожен з яких описується наступними параметрами: *наявність стенозу гілок легеневої артерії* (1 – ні, 2 – так); *індекс Наката* (відношення площі поперечного перерізу гілок легеневої артерії до площі поверхні тіла); *діаметр правої легеневої артерії*; *стать* (1 – чоловіча, 2 – жіноча); *вік (в місяцях)*; *кількість тромбоцитів в крові*; *кількість лейкоцитів*; *співвідношення*

об'ємних кровотоків (легеневої до артеріальної); *градієнт тиску на легеневу артерію*; *тиск в легеневій артерії*; *опір легеневих судин*.

В ролі класу об'єкта (пацієнта) виступала наявність стенозу гілок легеневої артерії (у 100 пацієнтів наявний стеноз, у всіх інший він був відсутній), тобто задача полягала в побудові випадкового лісу класифікації, який з максимально високою точністю спрогнозує цей клас на основі вхідних даних пацієнта. Незважаючи на те, що в реальному житті процедура діагностування даного стенозу не є надто дорогою та важкою, вона має певні ризики, і може бути небезпечною для пацієнтів (особливо тих, які мають вроджені вади серця). Саме тому в даному випадку задача класифікації є доволі актуальною.

Випадковий ліс об'єднує в собі дві головні ідеї: *метод бегінга* Браймана [3] і *метод випадкових підмножин* [4] (який був придуманий Тін Кам Хо). Бегінг полягає в тому, що із вибірки, яка використовується для навчання, утворюється певна кількість нових навчальних вибірок, на кожній із яких навчаються свої класифікатори. Метод випадкових підмножин діє схожим чином, але підхід застосовується для множини ознак. Тобто, випадковий ліс будує множину некорельованих між собою дерев прийняття рішень, які формують так званий ансамбль, після чого клас об'єкта визначається шляхом голосування (чим більше дерев вказують на певний клас, тим більша ймовірність даного класу). В більшості випадків таких дерев може бути велика кількість з різною глибиною та різною кількістю «листків». Саме тому при прогнозуванні важко зробити висновок, чому ліс визначив саме цей клас.

Для обраної вибірки експериментальних даних був побудований випадковий ліс класифікації за допомогою мови програмування *Python* [5], з використанням пакетів *pandas*, *numpy* і *scikit-learn*. Попередньо була розбита загальна вибірка (яка містила 128 спостережень) на навчальну (80%) та тестову (20%). Оскільки спостережень не надто багато, максимальна можлива кількість побудованих дерев складала 5.

Отриманий ліс оцінювався за критеріями точності, чутливості та специфічності, які представлені в табл. 1.

Таблиця 1

Оцінка отриманого випадкового лісу класифікації

Вибірка	Точність (%)	Чутливість	Специфічність
Навчальна (80%)	99.4	0.974	1
Тестова (20%)	93.5	0.824	0.967
Загальна (100%)	97.7	0.929	0.99

Для спроби інтерпретувати результат випадкового лісу було взято перевірочного пацієнта. Він мав наступні характеристики: індекс Наката – 193.27; діаметр правої легеневої артерії – 14.5; стать – *жіноча* (2); вік – 104 місяці; кількість тромбоцитів – 175; кількість лейкоцитів – 6.1; співвідношення об'ємних кровотоків – 0.8; градієнт тиску на легеневу артерію – 74; тиск в легеневій артерії – 20; опір легеневих судин – 2.1. Випадковий ліс спрогнозував, що у даної пацієнтки є стеноз гілок легеневої артерії (що є правдою) з вірогідністю 0.6 (тобто 60% дерев в лісі дали такий результат). На рис. 1 показані результати дерев, які дали правильний клас.

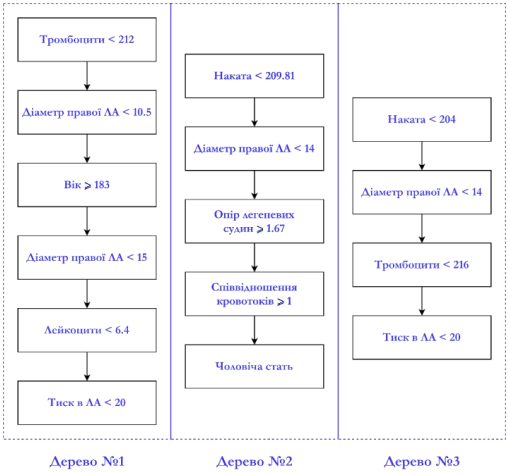

Рис. Результати дерев

Виконання умови в кожному блоці дає перший клас, тобто відсутність стенозу. Оскільки дереву під номером 3 знадобилося менше перевірок умов для визначення класу, можна зробити висновок, що воно є більш надійним. Індекс Наката у пацієнтки менше 204, діаметр правої легеневої артерії трохи більше 14, кількість тромбоцитів менше 216, проте тиск в легеневій артерії дорівнює 20, що свідчить про наявність стенозу. Таким чином, лікар може сказати, що підвищення тиску є однією з основних причин появи стенозу.

Висновки. Були успішно виконані задачі, поставлені в даному дослідження. Це дозволило досягти головної мети, а саме інтерпретувати результати прогнозування випадкового лісу класифікації максима-

льно зручно. Дана спроба є першим кроком для створення повноцінної системи підтримки рішень. В майбутньому планується автоматизувати даний процес, щоб машина окрім результату та його вірогідності показувала причини даного прогнозу, що допоможе експерту при прийнятті рішення.

Література:

1. Mehrotra D. Basics of Artificial Intelligence & Machine Learning. *Notion Press*. 2020. 80 p.
2. Biau G. Analysis of a Random Forests Model. *Journal of Machine Learning Research*. 2012. Vol. 13. P. 1063-1095.
3. Bühlmann P. Bagging, Boosting and Ensemble Methods. *Handbook of Computational Statistics*. 2011. P. 985–1022. DOI: 10.1007/978-3-642-21551-3_33
4. Дьяконов А.Г. Методы ансамблирования обучающихся алгоритмов. 2015. 41 ст.
5. Campesato O. Python 3 for Machine Learning. *Stylus Publishing, LLC*. 2020. 364 p.

DOI https://doi.org/10.30525/978-9934-588-79-2-1.3

ПРОГРАМНИЙ СПОСІБ ПІДГОТОВКИ ТЕКСТОВИХ ДАНИХ ДЛЯ ЇХ АПАРАТНОЇ ОБРОБКИ З ВИКОРИСТАННЯМ ПЛІС

Голуб Т. В.
аспірант кафедри комп'ютерних систем та мереж
Національного університету «Запорізька політехніка»

Зеленьова І. Я.
кандидат технічних наук,
доцент кафедри комп'ютерних систем та мереж
Національного університету «Запорізька політехніка»

Грушко С. С.
кандидат технічних наук,
доцент кафедри комп'ютерних систем та мереж
Національного університету «Запорізька політехніка»
м. Запоріжжя, Україна

Вступ. Інформація у вигляді текстових даних, зокрема представлених в природомовному виді, використовується у більшості напрямків