DOI 10.34185/1562-9945-3-146-2023-12
УДК 004.4:681.5

К.Ю. Островська, А.С. Мінаєнко

ДОСЛІДЖЕННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ РІШЕННЯ ЗАДАЧ МЕДИЧНОГО ПРОФІЛЮ

Анотація. Робота присвячена дослідженню методів машинного навчання для рішення задач медичного профілю. Мета роботи - аналіз методів машинного навчання для підвищення точності та скорочення часу діагностики захворювань сечостатевої системи у дітей. Предмет дослідження – класифікатор захворювань сечостатевої системи пацієнтів Дніпропетровської обласної дитячої клінічної лікарні "Дніпропетровської обласної ради". В результаті дослідження вирішено такі завдання: зроблено аналіз літератури щодо застосування методів машинного навчання до захворювань сечостатевої системи; розроблено програму для вилучення в напівавтоматичному режимі необхідної інформації з виписок; проаналізовано бібліотеки мови Python та частину методів машинного навчання; проведено первинний аналіз та передобробка даних; застосовано методи класифікації, відбору ознак та заповнення пропущених значень; проаналізовано одержані результати та виконано обґрунтування результатів дослідження у предметній галузі.
Ключові слова: алгоритми, класифікатор, машинне навчання, діагностика захворювань, випадковий ліс, метод к найближчого сусіда, багатошаровий перцептрон, логістична регресія, градієнтний бустинг, дерево рішень.

Вступ. Аналіз літератури показав актуальність досліджень у галузі застосування методів машинного навчання до діагностики захворювань сечостатевої системи. Найбільш поширеними методами були: дерево рішень, випадковий ліс, наївний Байєсовський класифікатор, багатошаровий перцептрон та мережа радіально-базових функцій. Як ознаки зазвичай виступали деякі показники загального та біохімічного аналізу крові, а також антропометричні показники. У цьому дослідженні враховані результати аналізованих робіт, і навіть взято інші методи машинного навчання, розширено безліч ознак відповідно до наявними даними й у ролі цільової змінної використовувалися клінічні стану.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Мета роботи - аналіз методів машинного навчання для підвищення точності та скорочення часу діагностики захворювань сечостатевої системи у дітей.

Предмет дослідження – класифікатор захворювань сечостатевої системи пацієнтів Дніпропетровської обласної дитячої клінічної лікарні "Дніпропетровської обласної ради".

Актуальність роботи зумовлюється необхідністю підвищення точності та скорочення часу діагностики захворювань сечостатевої системи у дітей.

Практична значимість дослідження полягає у підборі та описі таких методів машинного навчання, які допоможуть лікарям проводити діагностику захворювань сечостатевої системи у дітей.

У дослідженні використовувалася база даних анамнезів (виписок) пацієнтів дитячого віку з різними захворюваннями сечостатевої системи Дніпропетровської обласної дитячій клінічній лікарні у період з 2014 по 2021 роки.

У виписках міститься важлива інформація про перебіг захворювань, що проводяться аналізах та лікуванні. Розмір бази даних – 3773 файли. Кожен файл формату *.docx був частково або повністю заповнений шаблон анамнезу пацієнта. Необхідно було написати програму для переведення всієї бази даних у стандартний табличний формат за допомогою парсингу.

Більшість виписки – це показники проведених обстежень, які були представлені у табличному вигляді (13 груп): загальний та біохімічний аналізи крові та сечі, коагулограма, протеїнограма, ліпідограма тощо.

Відповідно до рекомендацій лікаря було обрано лише перший тимчасовий шар та наступні групи для вилучення (всього 33 показника, крім клінічного стану):

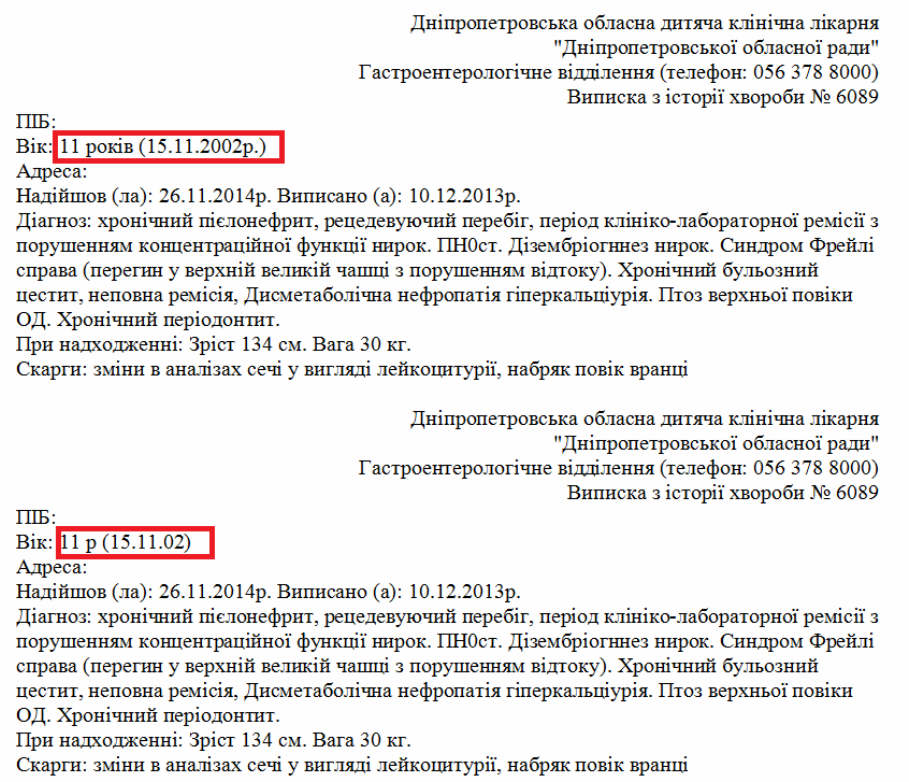
1. Вікова група: рік народження.
2. Загальний аналіз крові.
3. Біохімічний аналіз крові.
4. Загальний аналіз сечі.
5. Клінічні стани: пієлонефрит, гломерулонефрит, тубулоінтерстиціальний нефрит.

Усього 2881 пацієнтів із даними клінічними станами, які згодом використовувалися для аналізу.

Неможливість автоматичного вилучення необхідної інформації з анамнезів обумовлюється кількома проблемами.

У виписках присутня велика кількість явних та неявних друкарських помилок.

На рисунку 1 показаний інший приклад неявної друкарської помилки, яку доводиться виявляти та обробляти вручну: дві майже ідентичні виписки з різним віком в однієї й тієї самої людини.



Дніпропетровська обласна дитяча клінічна лікарня
"Дніпропетровської обласної ради"
Гастроентерологічне відділення (телефон: 056 378 8000)
Виписка з історії хвороби № 6089

ПІБ:
Вік: 11 років (15.11.2002р.)
Адреса:
Надійшов (ла): 26.11.2014р. Виписано (а): 10.12.2013р.
Діагноз: хронічний пієлонефрит, рецедевуючий перебіг, період клініко-лабораторної ремісії з порушенням концентраційної функції нирок. ПН0ст. Дізембріогінез нирок. Синдром Фрейлі справа (перегин у верхній великій чашці з порушенням відтоку). Хронічний бульозний цестит, неповна ремісія, Дисметаболічна нефропатія гіперкальціурія. Птоз верхньої повіки ОД. Хронічний періодонтит.
При надходженні: Зріст 134 см. Вага 30 кг.
Скарги: зміни в аналізах сечі у вигляді лейкоцитурії, набряк повік вранці

Дніпропетровська обласна дитяча клінічна лікарня
"Дніпропетровської обласної ради"
Гастроентерологічне відділення (телефон: 056 378 8000)
Виписка з історії хвороби № 6089

ПІБ:
Вік: 11 р (15.11.02)
Адреса:
Надійшов (ла): 26.11.2014р. Виписано (а): 10.12.2013р.
Діагноз: хронічний пієлонефрит, рецедевуючий перебіг, період клініко-лабораторної ремісії з порушенням концентраційної функції нирок. ПН0ст. Дізембріогінез нирок. Синдром Фрейлі справа (перегин у верхній великій чашці з порушенням відтоку). Хронічний бульозний цестит, неповна ремісія, Дисметаболічна нефропатія гіперкальціурія. Птоз верхньої повіки ОД. Хронічний періодонтит.
При надходженні: Зріст 134 см. Вага 30 кг.
Скарги: зміни в аналізах сечі у вигляді лейкоцитурії, набряк повік вранці

Рисунок 1– Приклад неявної друкарської помилки

Іншою проблемою є наявність великої кількості відсутніх значень у показниках, що може бути результатом непотрібності даного показника у конкретного пацієнта, так і людським фактором.

Також анамнези були написані різними лікарями, кожен із яких має свій власний стиль написання, тому в програму складно та недоцільно включати всі варіанти.

Приклад перших п'яти оброблених документів показано на рисунку 2.

	Diag	Age	Hb	L	Er	Ht	Tr	ESR	e	b	...	P	AlPh	Amy	G-GT	Color	SpGr	Prot	L(u)	Er(u)	F_ep
0	11.0	2010.0	129.0	5.0	4.35	NaN	190.0	4.0	1.0	0.0	...	1.15	623.73	NaN	NaN	1.0	1010.0	0.0	1.0	0.0	0.0
1	12.0	2008.0	113.0	7.7	4.28	NaN	293.0	10.0	1.0	0.0	...	1.62	652.88	NaN	NaN	1.0	1019.0	0.0	1.0	0.0	1.0
2	12.0	2001.0	152.0	6.1	5.58	NaN	343.0	10.0	4.0	0.0	...	1.18	446.10	56.9	NaN	3.0	1002.0	0.0	0.0	0.0	0.0
3	7.0	2002.0	138.0	6.6	NaN	NaN	NaN	7.0	0.0	0.0	...	1.16	53.00	NaN	NaN	0.0	1005.0	0.0	10.0	0.0	0.0
4	7.0	2007.0	133.0	9.0	5.60	47.0	176.0	2.0	3.0	3.0	...	1.50	536.00	NaN	NaN	1.0	1020.0	0.0	1.0	0.0	1.0

5 rows × 34 columns

Рисунок 2 – Перші п'ять рядків таблиці даних

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Відсутні значення позначені як NaN, показник "колір" ("Color") закодований цифрами від 0 до 6, а "діагноз" ("Diag") цифрами 7, 11, 12 відповідно до порядку вихідних діагнозів.

Завдання отримання показників з анамнезів пацієнтів може бути вирішене напівавтоматично: програма виробляє парсинг документів і видає помилки, які слід обробити вручну. Фрагмент коду, що показує загальну логіку роботи програми:

```
int number_err = 0;
for (int i = 0; i < files.size(); i++) {
List row = getRow(files.get(i));
// спроба отримати показники
if (row.size() == 1) {
// вивести докладну помилку, якщо спроба невдала
      number_err++;
      System.out.println(row.get(0) + files.get(i));13
      ...
}
else
      data.add(row);
}
```

В рамках дослідження було вирішено проблеми некоректного подання даних у випсках пацієнтів. Також з метою виконання поставленої задачі з вилучення ознак була написана програма, що витягує необхідну інформацію з анамнезів пацієнтів у напівавтоматичний режим. Результатом її роботи стала таблиця у форматі *.csv, що містить значення необхідних показників кожного пацієнта.

Перед застосуванням методів машинного навчання слід виконати первинний аналіз даних. Це дозволить зрозуміти деякі аспекти структури даних та виявити видимі закономірності у них.

По розподілу захворювань пацієнтів (рисунок 3) видно сильний дисбаланс, що слід враховувати під час виборів метрики якості і під час підбору гіперпараметрів і параметрів моделей.

На рисунку 4 видно, що у даних міститься велика кількість пропущених значень. Видалення об'єктів, у яких існує хоча б один недостатній показник, може призвести до втрати інформації та зниження якості класифікації. Для вирішення цієї проблеми можна використати методи заповнення пропущених значень. Перевіривши якість класифікації на даних без пропущених значень та із заповненими перепустками, можна оцінити доцільність використання методу заповнення пропущених значень.

Далі аналіз проводиться на даних, які не мають пропущених значень.

Згідно з коефіцієнтом варіації, існують кількісні ознаки, варіація яких дуже мала, що може свідчити про їхню неінформативність у даних (рисунок 5). Перевірити це можна за допомогою методу відбору ознак.

Аналогічно щодо номінальної ознаки «колір». Видно, що у більшості випадків він набуває значення «с/ж» чи «ж» (див. рисунок 6). Це можна передбачити в процесі кодування даного показника, прибравши значення, що рідко зустрічаються, тим самим зменшивши швидкість навчання моделей і витрати пам'яті.

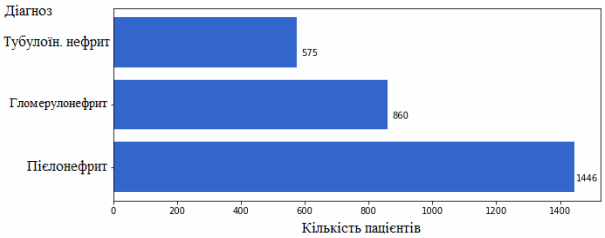


Рисунок 3 – Розподіл захворювань

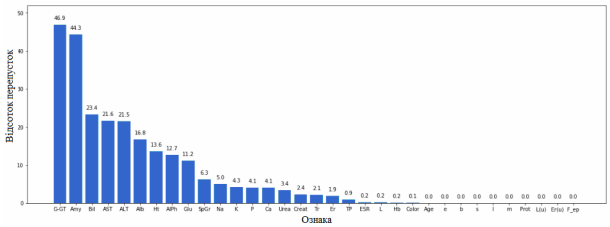


Рисунок 4 – Відсоток перепусток за кожним показником

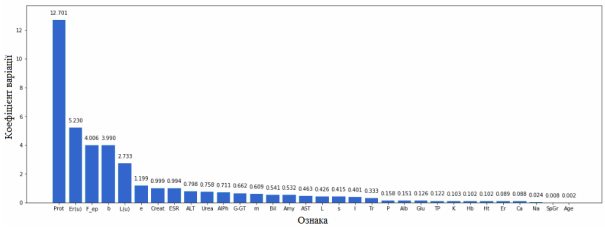


Рисунок 5 – Коефіцієнт варіації кількісних ознак

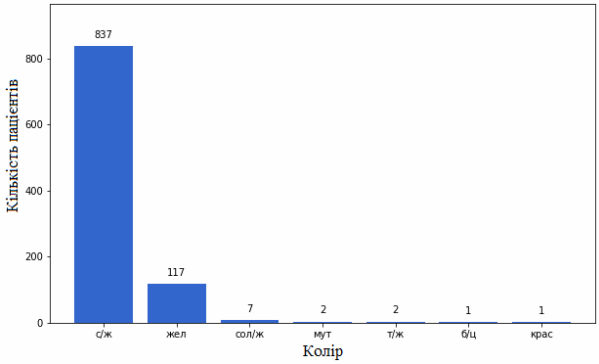


Рисунок 6 – Розподіл значень показника «колір»

Для визначення ступеня лінійного зв'язку між змінними використовувався коефіцієнт кореляції Спірмена, оскільки розподіл багатьох ознак далекий від нормального, що підтверджує статистичний тест D'Agostino-Pearson. По ній видно, що багато ознак не корелюють між собою, проте існують такі показники, коефіцієнт кореляції яких за модулем більше 0.7, що вказує на високий лінійний статистичний зв'язок.

Для визначення кількості груп у гістограмі використовувалося правило Скотта. За деякими гістограмами розподілу ознак щодо кожного класу можна побачити, що ознака загалом має здатність відокремлювати велику кількість об'єктів одного класу від іншого (рисунок 7).

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

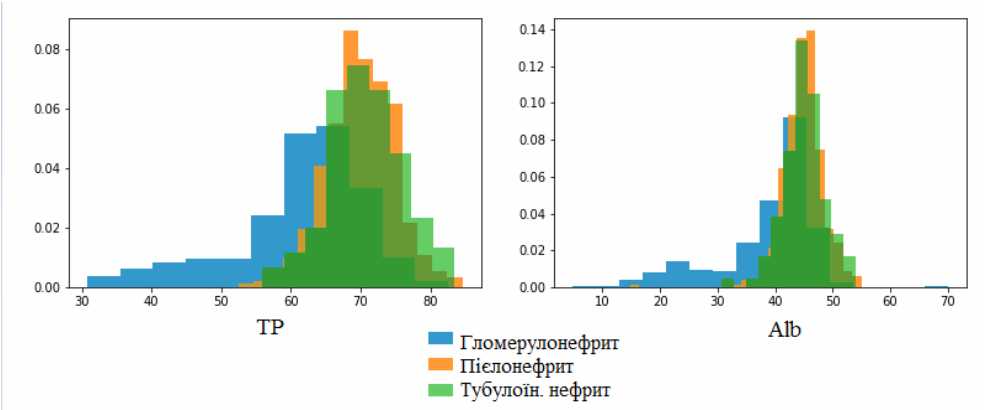


Рисунок 7 – Приклад гістограм розподілу

Аналогічно щодо графіків розсіювання попарних ознак(рисунок 8). Істинність припущення про інформативність тієї чи іншої ознаки класифікації можна оцінити, зокрема, за допомогою властивостей класифікаторів.

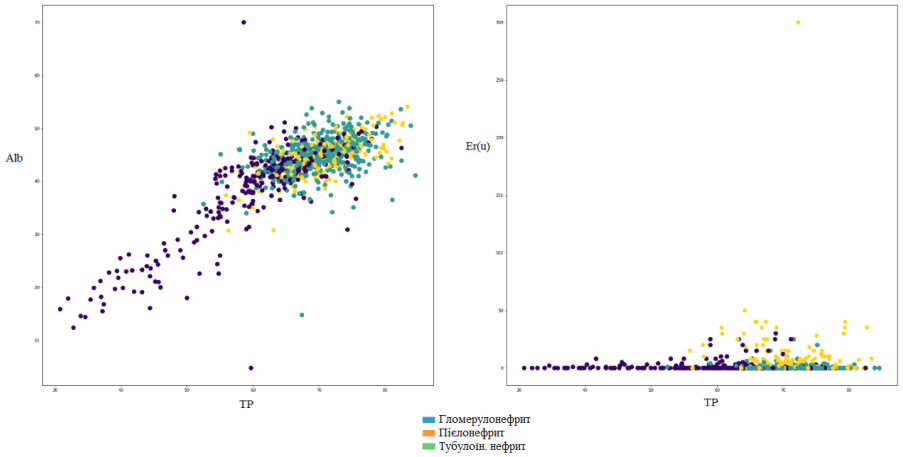


Рисунок 8 – Приклад графіків розсіювання

Виконавши зниження розмірності з метою візуалізації шляхом основних компонентів, можна побачити складну структуру даних (рисунок 9).

Пояснені дисперсії кожної компоненти становлять 12.1%, 8.3%, 7.5%, що у сумі дорівнює 27.9% поясненої дисперсії першими трьома компонентами. Це говорить про те, що 72.1% інформації залишається поза аналізом даного методу та якісно апроксимувати наявні дані лінійними різноманіттями розмірності 3 не вийде.

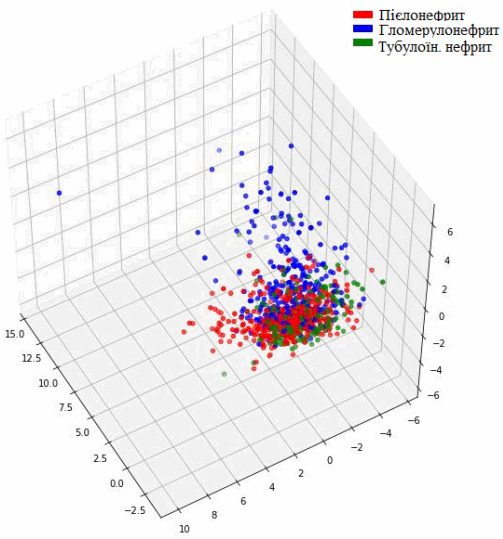


Рисунок 9 – Візуалізація методом головних компонент

Інший метод візуалізації - картка Кохонена, що самоорганізується, - здійснює нелінійне, впорядковане, гладке відображення даних на двовимірну решітку нейронів (карту). По побудованій карті видно, що є невелика кількість ознак та їх комбінацій, за якими можна оцінити чітку кластерну структуру даних, і навіть різні залежності між ознаками (рисунок 10). Ознаки, за картами яких видно рівномірність значень і кореляції коїться з іншими ознаками, були представлені на рисунку 10.

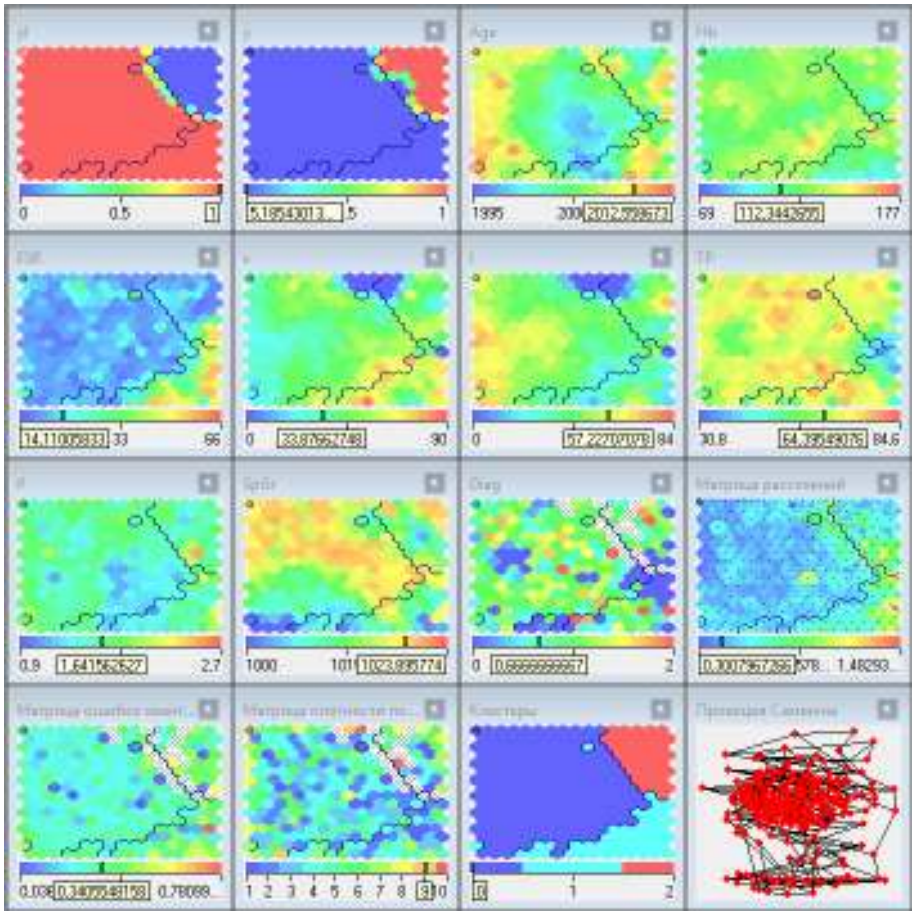


Рисунок 10 – Картка Кохонена, що самоорганізується

Для приведення числових ознак до єдиного масштабу використовувалася функція MinMaxScaler бібліотеки Scikit-Learn, яка перекладає задану множину значень ознак до певного діапазону. У цій роботі як діапазон використовувався відрізок [0, 1].

Номінальна ознака «колір» закодована за допомогою функції OneHotEncoder бібліотеки Scikit-Learn з використанням тільки перших двох значень, що його найчастіше зустрічаються: «с/ж», «жов».

Використані дані з медичних виписок пацієнтів мають складну багатовимірну структуру, значний дисбаланс у захворюваннях, велика кількість

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

пропущених значень, що необхідно враховувати під час застосування методів машинного навчання. Також багато методів вимагають від ознак певних властивостей, тому важливим етапом є передобробка даних, що запобігатиме уповільненню збіжності алгоритмів і зниження якості класифікації.

Після первинного аналізу та передобробки даних слідує етап застосування методів машинного навчання. У рамках цієї роботи використовувалися такі методи: дерево рішень (DT), випадковий ліс (RF), градієнтний бустинг (GB), логістична регресія (LR), метод К найближчих сусідів (KNN) та багатошаровий перцептрон (MLP).

Для кожного класифікатора представлена сітка з багатьох гіперпараметрів. Вона використовувалася для всіх підходів підвищення якості, і по ній було здійснено пошук оптимального набору гіперпараметрів, що забезпечує максимальне середнє збалансовану метрику якості mcc (Matthews correlation coefficient) за 5-блоковою перехресною перевіркою.

У таблиці 1 зазначено якість методів по 5-блоковій перехресній перевірці (mcc cv) після знаходження оптимальних гіперпараметрів.

Перехресна перевірка показує адекватність методу в цілому стосовно до даних.

Таблиця 1

Якість без відбору ознак та заповнення перепусток

	DT	RF	GB	LR	KNN	MLP
mcc cv	0.417	0.594	0.6	0.499	0.409	0.532

З метою підвищення якості класифікації можна використати метод рекурсивного відбору ознак RFE бібліотеки Scikit-Learn на базі результату роботи кожної моделі Він дозволить виявити такий набір інформативних ознак, який міг би спростити модель та підвищити її якість.

До існуючої сітки множин гіперпараметрів додалися потужності відбираються множин ознак: 2, 3, ..., n-1, де n = 34 після кодування номінального показника «колір».

На відміну від таблиці 1, у таблиці 2 зазначено якість методів відібраних інформативних ознак.

Таблиця 2

Якість з відбором ознак і без заповнення перепусток

	DT	RF	GB	LR	KNN	MLP
mcc cv	0.451	0.602	0.6	0.515	0.501	0.575

Дані містять велику кількість пропущених значень. Видаливши об'єкти з перепустками, можна втратити значну частину інформації в даних і погіршити якість класифікації.

З метою запобігання цьому наслідку в цій роботі використовувався метод К найближчих сусідів для заповнення пропущених значень.

У таблиці 3 зазначено якість методів на даних без відбору ознак і після заповнення пропущених значень методом К найближчих сусідів при К = 5 і евклідовою відстанню.

Таблиця 3

Якість без відбору ознак та із заповненням перепусток

	DT	RF	GB	LR	KNN	MLP
mcc cv	0.358	0.541	0.554	0.453	0.364	0.503

Спроба підвищення якості класифікації можна застосувати метод рекурсивного відбору ознак цього разу після заповнення перепусток. Таблиця 4 показує якість методів після відбору ознак та заповнення перепусток.

Таблиця 4

Якість з відбором ознак та заповненням перепусток

	DT	RF	GB	LR	KNN	MLP
mcc cv	0.384	0.546	0.564	0.461	0.451	0.514

За графіками пошуку оптимальних гіперпараметрів моделей (без заповнення припусків) (рисунки 11 – 22) можна оцінити їхню стійкість.

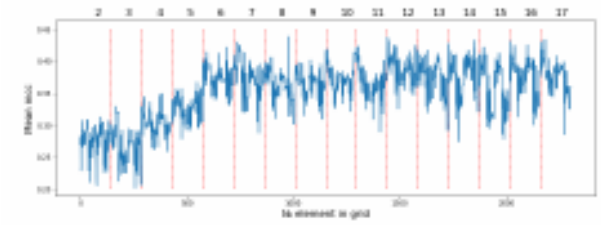


Рисунок 11 - Дерево рішень

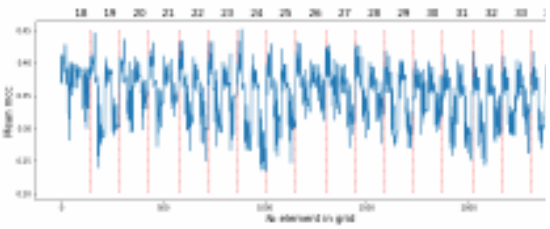


Рисунок 12 - Дерево рішень

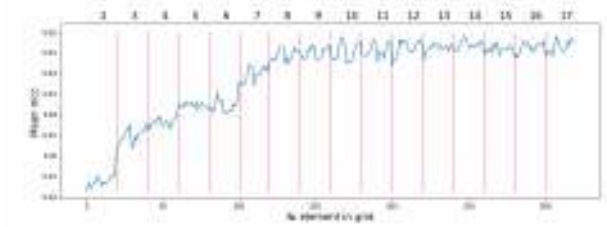


Рисунок 13 - Випадковий ліс

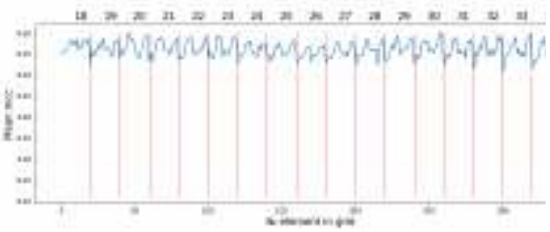


Рисунок 14 - Випадковий ліс

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

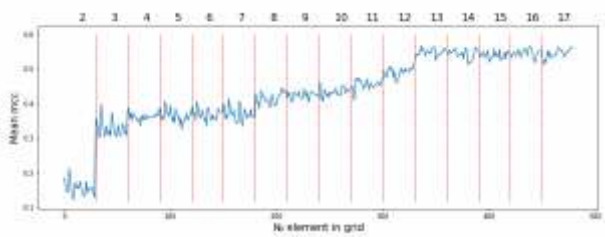


Рисунок 15 - Градієнтний бустинг

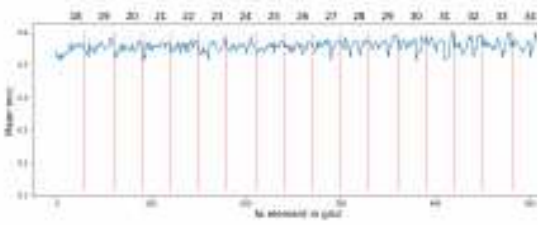


Рисунок 16- Градієнтний бустинг

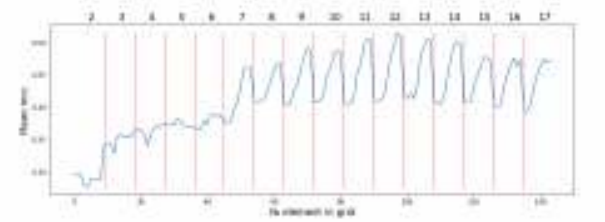


Рисунок 17 - Логістична регресія

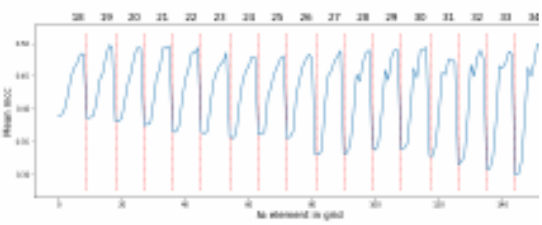


Рисунок 18 - Логістична регресія

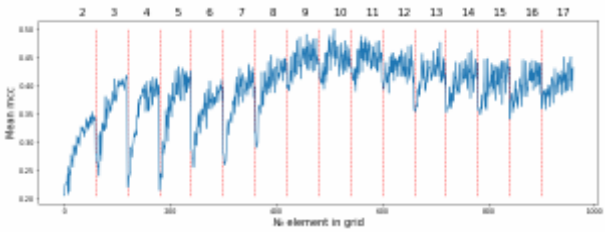


Рисунок 19 - Метод
К найближчих сусідів

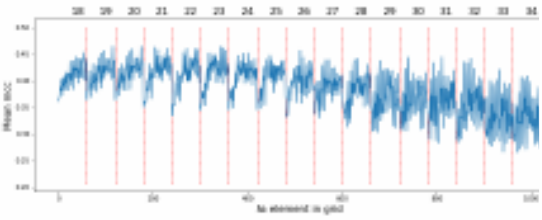


Рисунок 20 - Метод
К найближчих сусідів

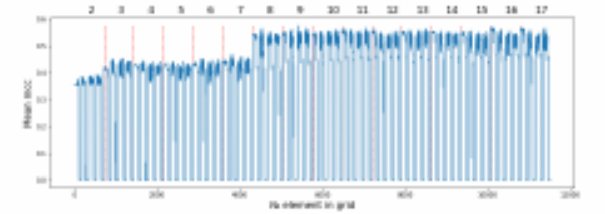


Рисунок 21 - Багатошаровий
перцептрон

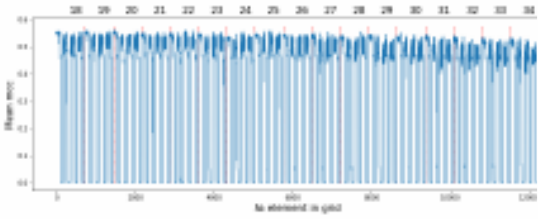


Рисунок 22 - Багатошаровий
перцептрон

З рисунків 11 – 22 видно, що випадковий ліс і градієнтний бустинг мають більшу в порівнянні з іншими моделями стійкість щодо гіперпараметрів: їх мале зміна веде до малої зміни якості по перехресній перевірці як всередині певної множини ознак, і глобально, починаючи з певної потужності. Це правильно до і після заповнення перепусток.

Розглянуті методи машинного навчання показали різні результати наявних даних. Зокрема, випадковий ліс та градієнтний бустинг дали кращу в порівнянні з іншими методами якість по перехресній перевірці та були більш стійкими щодо гіперпараметрів. Все це є показником їх адекватності стосовно

даних. Метод відбору ознак значно підвищив адекватність моделей, а метод заповнення перепусток лише погіршив її.

Після попередніх етапів необхідно перевірити вже навчені моделі на тестовому наборі даних: такому наборі, який брав участь у пошуку гіперпараметрів і параметрів моделей.

Така перевірка дозволить оцінити передбачувану здатність моделі та зрозуміти, наскільки добре вона може працювати на практиці.

Для оцінки якості навчених моделей використовувалися матриця неточностей класифікації (CM) та mcc на тестовому наборі даних (mcc).

У зведеній таблиці 5 зазначено якість методів на тестовому наборі з різними підходами:

(1) – без відбору ознак, без заповнення перепусток;

(2) - з відбором ознак, без заповнення перепусток;

(3) – без відбору ознак, із заповненням перепусток;

(4) – з відбором ознак, із заповненням пропусків.

Таблиця 5

Якість на тестовому наборі

		DT	RF	GB	LR	KNN	MLP
(1)	CM	58 19 7 19 66 24 7 13 29	59 24 1 8 97 4 5 27 17	69 12 3 8 90 11 9 19 21	60 17 7 11 74 24 7 16 26	53 30 1 6 93 10 6 36 7	59 22 3 11 91 7 7 25 17
	mcc	0.433	0.548	0.591	0.472	0.408	0.503
(2)	CM	52 24 8 20 63 26 9 10 30	59 24 1 9 95 5 6 26 17	67 15 2 9 90 10 8 19 22	61 14 9 11 78 20 7 12 30	60 20 4 10 83 16 8 28 13	62 18 4 12 89 8 7 22 20
	mcc	0.385	0.532	0.584	0.533	0.429	0.53
(3)	CM	135 45 35 65 202 95 35 40 69	155 53 7 21 331 10 18 91 35	163 43 9 31 316 15 24 75 45	160 36 19 57 230 75 25 47 72	111 92 12 19 335 8 12 125 7	146 59 10 34 307 21 23 90 31
	mcc	0.33	0.54	0.547	0.438	0.369	0.448
(4)	CM	143 37 35 65 189 108 29 51 64	150 60 5 18 335 9 15 93 36	164 42 9 31 306 25 21 70 53	157 35 23 56 228 78 26 49 69	140 67 8 27 322 13 16 114 14	145 62 8 35 315 12 26 90 28
	mcc	0.314	0.542	0.545	0.421	0.427	0.458

За таблицею 5 видно, що градієнтний бустинг має велике значення метрики якості на тестовому наборі щодо всіх підходів, що говорить про кращу передбачувальну здатність. Що стосується самих підходів, то в цілому метод відбору ознак покращив передбачувальну здатність тільки метод К найближчих

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

сусідів: до і після заповнення перепусток його якість підвищилася. Інші випадки є суворо однозначними. Метод заповнення перепусток загалом лише погіршив якість більшості методів на тестовому наборі, і навіть знизив оцінку якості методів по перехресній перевірці, що ставить під його застосування практично.

Також класифікатори зазвичай гірше розпізнають тубулоінтерстиціальний нефрит, ніж інші захворювання. Зокрема, градієнтний бустинг без відбору ознак та заповнення перепусток розпізнав 21 випадок даного діагнозу з 49, 19 пацієнтів відніс до тих, хто має пієлонефрит.

В результаті перевірки передбачуваної спроможності методів градієнтний бустинг показав більшу якість класифікації по всім підходам. Серед підходів методи відбору ознак та заповнення пропущених значень не покращили якість градієнтного бустингу на тестовій множині.

Загалом за всіма класифікаторами метод відбору ознак значно не покращив і не погіршив результати, а метод заповнення перепусток лише погіршив їх, тому перевірка доцільності використання його на практиці вимагає подальшого дослідження у цьому напрямі.

Отримані в цій роботі результати можуть мати практичну значимість, пов'язану з отриманням необхідної інформації з виписок пацієнтів для подальшого аналізу, а також теоретичну значущість, яка обумовлюється вибором градієнтного бустингу як методу підвищення точності і скорочення часу діагностики захворювань сечостатевої системи в дітей. Як інтерпретацію отриманих результатів та вибору градієнтного бустингу використовувалася специфіка даного методу, і навіть його математичні характеристики.

Висновки. Дослідження в галузі застосування методів машинного навчання до діагностики захворювань сечостатевої системи у дітей є актуальними на сьогоднішній день і проводяться як у нашій країні, так і за кордоном.

В рамках даної роботи були вирішені різні труднощі роботи з виписками пацієнтів і була написана програма, яка одержує необхідну інформацію з них у напівавтоматичному режимі.

Також були проаналізовані різні методи машинного навчання та підходи до підвищення їх якості стосовно вилучених даних. В результаті градієнтний бустинг показав більшу в порівнянні з іншими методами адекватність і передбачувальну здатність за всіма підходами. Метод рекурсивного відбору ознак загалом значно покращив і погіршив якість класифікаторів, а метод заповнення пропущених значень у більшості випадків сильно погіршив його.

Результати роботи показали необхідність розвитку цієї проблематики. Подальші дослідження в галузі застосування на практиці методів машинного навчання до діагностики захворювань сечостатевої системи у дітей можуть допомогти лікарю не тільки у вилученні інформації з масиву виписок та його аналізі, а й у скороченні часу постановки діагнозу та збільшенні його точності.

ЛІТЕРАТУРА

1. Мінцер О.П. Оброблення клінічних і експериментальних даних у медицині: навч. посібник / О.П. Мінцер, Ю.В. Вороненко, В.В. Власов - К.: Вища шк., 2003. - 350 с.
2. Кобзарь А.И. Прикладная математическая статистика. – М.: Физматлит, 2006. – 626–628 с.
3. Кохонен Т. Самоорганизующиеся карты / пер. 3-го англ. изд. – М.: БИНОМ. Лаборатория знаний, 2014. – 655 с.
4. Голованова І.А. Основи медичної статистики: навч. посіб. для аспірантів та клінічних ординаторів / І.А. Голованова, І.В. Бєлікова, Н.О. Ляхова. – Полтава, 2017. – 113 с.
5. Фадеев П.А. Болезни почек. Пиелонефрит. – М.: Мир и Образование, 2011. – 180 с.
6. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А.А. Слинкина. – М.: ДМК Пресс, 2015. – 400 с.
7. Хайкин С. Нейронные сети: Полный курс / пер. с англ. Н.Н. Куссуль, А.Ю. Шелестова. – 2-е изд., испр. – М.: Издательский дом Вильямс, 2008. – 1103 с.
8. Bauer E. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants / E. Bauer, R. Kohavi // Machine Learning. – 1999. – P. 105–139.
9. Boughorbel S. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric / S. Boughorbel, F. Jarray, M. El-Anbari // PLoS ONE 12(6). – 2017. – 17 p.
10. Breiman L. Bagging Predictors / L. Breiman // Machine Learning. – 1996. – P. 123–140.
11. D'Agostino R.B. An omnibus test of normality for moderate and large sample size / R.B. D'Agostino // Biometrika. – 1971. – Vol. 58, No. 2. – P. 341–348.
12. Gopika S. Machine learning Approach of Chronic Kidney Disease Prediction using Clustering Technique / S. Gopika, Dr.M. Vanitha // International Journal of Innovative Research in Science, Engineering and Technology. – 2017. – Vol. 6, No. 7. – P. 14488–14496.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

13. Hornik K. Approximation Capabilities of Muitilayer Feedforward Networks / K. Hornik // Neural Networks. – 1990. – Vol. 4. – P. 251–257.
14. Kazemi Y. A novel method for predicting kidney stone type using ensemble learning / Y. Kazemi, S.A. Mirroshandel // Artificial Intelligence in Medicine. – 2017. – Vol. 79, No. 3. – P. 1696–1707.
15. Lambodar J. Distributed Data Mining Classification Algorithms for Prediction of Chronic Kidney Disease / J. Lambodar, K. Narendra // International Journal of Emerging Research in Management and Technology. – 2015. – Vol. 4, No. 11. – P. 110–180.
16. Ramya S. Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms / S. Ramya, N. Radha // International Journal of Innovative Research in Computer and Communication Engineering. – 2016. – Vol. 4, No. 1. – P. 812–820.
17. Scott D.W. On Optimal and Data-Based Histograms / D.W. Scott // Biometrika. – 1979. – Vol. 66, No. 3. – P. 605–610.
18. United States Patent № US 7,657,521 B2, 02.02.2010. System and method for parsing medical data [text] / Fred E. Masarie, Stuart Lopez, Michael I. Lieberman // United States Patent № US 7657521 B2. 2010.
19. Yoruk U. Automatic Renal Segmentation for MR Urography Using 3D-GrabCut and Random Forests / U. Yoruk, B.A. Hargreaves, S.S. Vasanawala // International Society for Magnetic Resonance in Medicine. – 2017. – Vol. 79, No. 3. – P. 1696–1707.

REFERENCES

1. Mintser O.P. Obroblennia klinichnykh i eksperymentalnykh danykh u medytsyni: navch. posibnyk / O.P. Mintser, Yu.V. Voronenko, V.V. Vlasov - K.: Vyshcha shk., 2003. - 350 s.
2. Kobzar A.Y. Prykladnaia matematycheskaia statystyka. – M.: Fyzmatlyt, 2006. – 626–628 s.
3. Kokhonen T. Samoorhanyzuiushchyesia karti / per. 3-ho anhl. yzd. – M.: BYNOM. Laboratoryia znanyi, 2014. – 655 s.
4. Holovanova I.A. Osnovy medychnoi statystyky: navch. posib. dlia aspirantiv ta klinichnykh ordynatoriv / I.A. Holovanova, I.V. Bielikova, N.O. Liakhova. – Poltava, 2017. – 113 s.
5. Fadeev P.A. Bolezny pochek. Pyelonefryt. – M.: Myr y Obrazovanye, 2011. – 180 s.
6. Flakh P. Mashynnoe obuchenye. Nauka y yskusstvo postroenyia alhorytmov, kotorыe yzvlekaiut znanyia yz dannыkh / per. s anhl. A.A. Slynkyna. – M.: DMK Press, 2015. – 400 s.
7. Khaikyn S. Neironnie sety: Polnii kurs / per. s anhl. N.N. Kussul, A.Iu. She-lestova. – 2-e yzd., yspr. – M.: Yzdatelskyi dom Vyliams, 2008. – 1103 s.

8. Bauer E. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants / E. Bauer, R. Kohavi // Machine Learning. – 1999. – P. 105–139.
9. Boughorbel S. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric / S. Boughorbel, F. Jarray, M. El-Anbari // PLoS ONE 12(6). – 2017. – 17 p.
10. Breiman L. Bagging Predictors / L. Breiman // Machine Learning. – 1996. – P. 123–140.
11. D'Agostino R.B. An omnibus test of normality for moderate and large sample size / R.B. D'Agostino // Biometrika. – 1971. – Vol. 58, No. 2. – P. 341–348.
12. Gopika S. Machine learning Approach of Chronic Kidney Disease Prediction using Clustering Technique / S. Gopika, Dr.M. Vanitha // International Journal of Innovative Research in Science, Engineering and Technology. – 2017. – Vol. 6, No. 7. – P. 14488–14496.
13. Hornik K. Approximation Capabilities of Muitilayer Feedforward Networks / K. Hornik // Neural Networks. – 1990. – Vol. 4. – P. 251–257.
14. Kazemi Y. A novel method for predicting kidney stone type using ensemble learning / Y. Kazemi, S.A. Mirroshandel // Artificial Intelligence in Medicine. – 2017. – Vol. 79, No. 3. – P. 1696–1707.
15. Lambodar J. Distributed Data Mining Classification Algorithms for Prediction of Chronic Kidney Disease / J. Lambodar, K. Narendra // International Journal of Emerging Research in Management and Technology. – 2015. – Vol. 4, No. 11. – P. 110–180.
16. Ramya S. Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms / S. Ramya, N. Radha // International Journal of Innovative Research in Computer and Communication Engineering. – 2016. – Vol. 4, No. 1. – P. 812–820.
17. Scott D.W. On Optimal and Data-Based Histograms / D.W. Scott // Biometrika. – 1979. – Vol. 66, No. 3. – P. 605–610.
18. United States Patent № US 7,657,521 B2, 02.02.2010. System and method for parsing medical data [text] / Fred E. Masarie, Stuart Lopez, Michael I. Lieberman // United States Patent № US 7657521 B2. 2010.
19. Yoruk U. Automatic Renal Segmentation for MR Urography Using 3D-GrabCut and Random Forests / U. Yoruk, B.A. Hargreaves, S.S. Vasanawala // International Society for Magnetic Resonance in Medicine. – 2017. – Vol. 79, No. 3. – P. 1696–1707.

Received 17.04.2023.
Accepted 20.04.2023.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Research in machine learning methods
for solving problems of the medical profile

The work is devoted to the study of machine learning methods for solving medical problems. The aim of the work is to analyze machine learning methods to improve the accuracy and reduce the time for diagnosing diseases of the genitourinary system in children. The object of research is machine learning methods. The subject of the study is a classifier of diseases of the genitourinary system of patients of the Dnipropetrovsk Regional Children's Clinical Hospital "Dnepropetrovsk Regional Council". As a result of the study, the following tasks were solved: an analysis of the literature on the application of machine learning methods to diseases of the genitourinary system was made; a program was developed to extract the necessary information on statements in a semi-automatic mode; Python libraries and part of machine learning methods were analyzed; primary analysis and processing of data was carried out; applied methods of classification, feature selection and filling in missing values; the obtained results were analyzed and the substantiation of the research results in the subject area was made.

Островська Катерина Юріївна – к.т.н., доцент, доцент кафедри інформаційних технологій і систем НІІ ІПБТ УДУНТ.
Мінаєнко Анна Сергіївна – магістр кафедри інформаційних технологій і систем НІІ ІПБТ УДУНТ.

Ostrovska Kateryna Yuriyivna - Ph.D., associate professor, associate professor of the department of information technologies and systems of Ukrainian State University of Science and Technology.
Minayenko Anna Serhiyivna - Master of the Department of Information Technologies and Systems of Ukrainian State University of Science and Technology.