ВІДКРИТІ МЕДИЧНІ ДАНІ ЯК ОСНОВА ДЛЯ СТВОРЕННЯ КЛАСИФІКАЦІЙНИХ МОДЕЛЕЙ: ПРИНЦИПИ FAIR І ПРАКТИЧНА РЕАЛІЗАЦІЯ

Карпінська О.Д.

ДУ «Інститут патології хребта та суглобів ім. проф. М.І. Ситенка НАМН України», Харків, Україна

***Ключові слова**: FAIR, класифікаційна модель, тазовий баланс*

Дані є пальним сучасної науки, зокрема медицини. Медична наука відстає від інших галузей у сфері інформаційних технологій, обміну даними та взаємодії. Набори даних необхідні для створення прогностичних і класифікаційних систем, моделювання складних кореляційних і причинно-наслідкових зв'язків, розробки програмного забезпечення, тестування та клінічного навчання. Ефективність цих процесів суттєво залежить від якості, доступності та повторного використання даних, тобто розвитку концепції FAIR (Findable, Accessible, Interoperable, Reusable). FAIR – це набір принципів, що перетворюють хаотичні масиви інформації на структуровані, надійні та довготривалі ресурси [1].

Одним із потужних методів дотримання FAIR-принципів, таких як конфіденційність та повторне використання, є створення синтетичних наборів даних на основі реальних. Цей підхід використовує генеративні моделі для відтворення статистичних характеристик, кореляцій і розподілів вихідного набору без збереження жодного реального запису. Сьогодні глибока анонімізація підтримується численними програмними засобами – як готовими системами, так і бібліотеками/пакетами програмування.

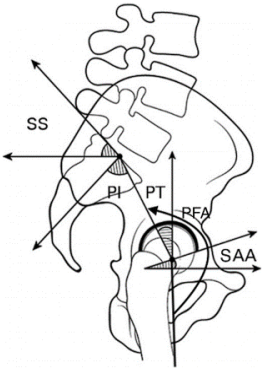
Більшість реальних наборів даних залишаються закритими для широкого кола користувачів, тоді як синтетичних – поки обмежена кількість [2]. Якщо у галузях кардіології, онкології, ендокринології, а останнім часом і COVID-досліджень спостерігається активне відкриття даних, то в ортопедії подібні ресурси трапляються значно рідше. Саме тому кожен відкритий набір даних у цій галузі має особливу наукову та практичну цінність.

Мета. На основі відкритого набору даних сагітального тазового балансу показати метод створення моделі класифікації захворювань хребта по даним сагітального тазового балансу.

Матеріали і методи. В основу для моделювання взято набір даних Vertebral Column (8/8/2011) [3]. Набір біомедичних даних, створений доктором Dr. Henrique da Mota в Центрі медико-хірургічної реадаптації масажних захворювань (Ліон, Франція). Дані були організовані для класифікації пацієнтів за однією з трьох категорій: нормальний (100 пацієнтів), грижа диска (60 пацієнтів) або спондилолістез (150 пацієнтів).

Кожен пацієнт представлений у наборі даних шістьма біомеханічними атрибутами, отриманими з форми та орієнтації тазу й поперекового відділу хребта: кут нахилу тазу, кут нахилу тазу, кут поперекового лордозу, нахил крижової кістки, променева кістка тазу та ступінь спондилолістезу.

PI (Pelvic incidence) – інцидентність тазу або кут нахилу тазу
PT (Pelvic tilt) – нахил тазу, орієнтацію тазу у сагітальній площині).
LLA (Lumbar lordosis angle) – кут поперекового лордозу (
SS (Sacral slope) – нахил крижів або кут нахилу крижової кістки
PR (pelvis radius) – відстань у міліметрах між вісь стегнових голівок та задньо-верхнім кутом замикальної пластини S1
DS (Degree of Spondylolisthesis) (рівень зміщення хребця відносно сусіднього).



Специфіка цього набору полягала в тому, що дослідних при синтезі не обмежили точність генерації, тому дані були представлені з точністю до 5-6 десяткових знаків. Для вимірювання кутів це надмірна точність. Тому набір був переформатований в Excel з округленими даних до цілих чисел.

На основі даних була побудована класифікаційна система для визначення ймовірності діагнозу – грижа, спондилолістез або норма, використовуючи дані сагітального тазового балансу. Модель була побудована в R [4].

Результати. Будь-яка класифікаційна модель потребує навчальної і тестової вибірок. Як правило, набір даних ділиться у відношенні 70-80% / 20-30 %. Отже завантажений набор даних був випадковим чином поділений на 2 вибірки.

Модель побудована на основі алгоритму Random Forest (випадковий ліс), це багатомодельний алгоритм, які поєднує результати кількох моделей для підвищення точності класифікації. Результатом моделі є класифікація, у нашому випадку визначення ймовірності відношення до однієї з груп: грижа, спондилолістез, норма.

Необхідним елементом є контроль результату в процесі побудови моделі.

Реальний клас	Hernia	Normal	Spondylolisthesis	class.error
Hernia	28	14	1	0.349
Normal	12	53	4	0.232
Spondylolisthesis	0	5	100	0.048

Результат роботи моделі. Точність моделі становить 82.8 %, тобто модель правильно класифікує ≈ 83 % спостережень. Найкраща точність визначена для Spondylolisthesis (помилка ~5%), тоді як для кили – помилка біля 35 % (з 43 спостережень правильно визначено 28).

Дана модель компонентна, відповідно для кожного класу діагнозу важливими будуть свої показники.

Показник	Hernia	Normal	Spondylolisthesis
PI	4.319	8.304	11.033
PT	5.127	13.634	5.993
LLA	19.108	-4.646	9.847
SS	21.256	-2.499	8.398
PR	3.233	24.002	10.316
DS	28.611	50.630	54.129

Найбільшу вагу в діагностиці патологіє має показник DS (зміщення хребця), причому для класу Spondylolisthesis – він максимальний (54,123). Для класу Hernia важливими є показники LLA і SS.

В ході створення моделі можна контролювати здатність компонент моделі класифікувати захворювання (рис. 1).

На представлених графіках видно, що Spondylolisthesis класифікується краще, ніж Norma ↔ Hernia. Можливо для покращення діагностики потрібні додаткові показники, або інший алгоритм класифікації.

Для якісного аналізу вибірка в 310 спостережень достатня, але її верифікація потребує додаткових спостережень.

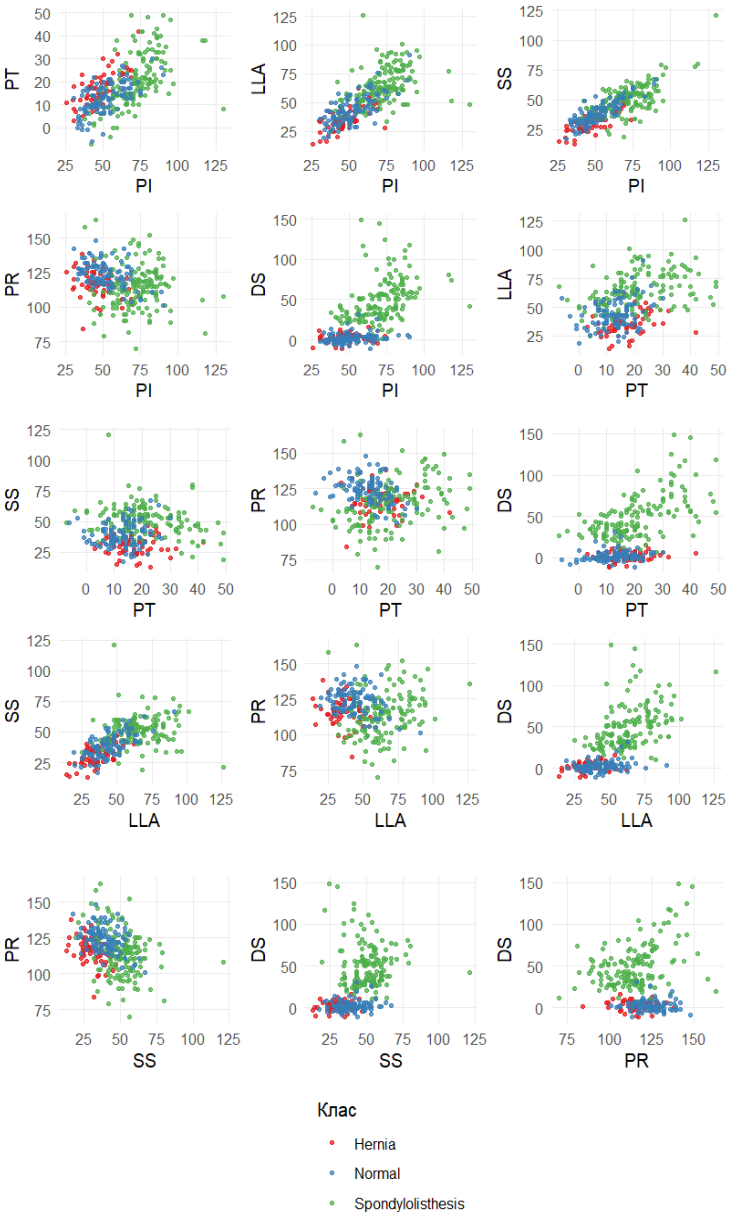
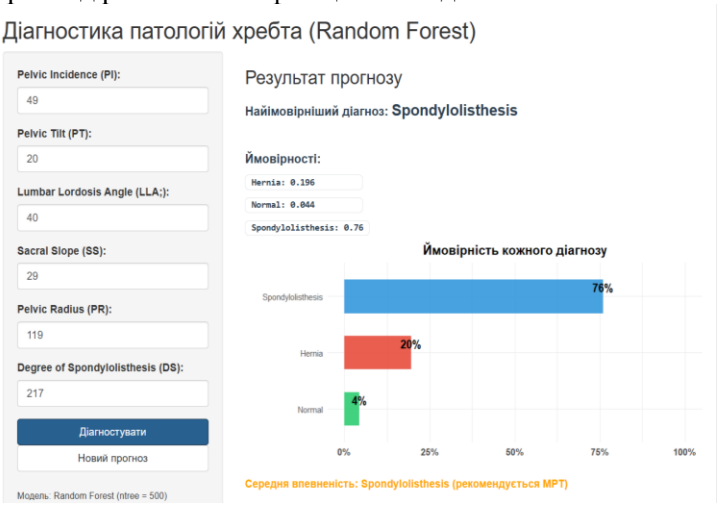


Рис. 1.

Приклад роботи класифікаційної моделі.



https://m0ygxz-0-0.shinyapps.io/pelvisbalanse/

Висновок. Наведений приклад не претендує на роль остаточної або універсальної класифікаційної моделі. Метою роботи було продемонструвати можливість застосування відкритих наборів даних та підкреслити важливість їх доступності для відтворюваності й повторного використання у наукових дослідженнях. Подальше накопичення нових спостережень і їх ретельна верифікація дозволять підвищити точність, надійність, прогностичну та класифікаційну спроможність подібних моделей у майбутньому.

Література

1. FAIR Principles. https://www.go-fair.org/fair-principles/
2. Walonoski J., Kramer M., Nichols J., Quina A. et al. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, *Journal of the American Medical Informatics Association*, V. 25 (3): 230–238 https://doi.org/10.1093/jamia/ocx079
3. Barreto, G. & Neto, A. (2005). Vertebral Column [Dataset]. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5K89B
4. R Core Team. (2021). R: A language and environment for statistical computing [Computer software].
5. Roussouly, P., & Pinheiro-Franco, J.L. (2011). Biomechanical analysis of the spino-pelvic organization and adaptation in pathology. *European spine journal: official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society*, *20 Suppl 5* (Suppl 5), 609–618. https://doi.org/10.1007/s00586-011-1928-x