УДК 004.9:616.379-008.64
Кіт І. - ст. гр. СНаз-22
Тернопільський національний технічний університет імені Івана Пулюя

ІНТЕЛЕКТУАЛЬНА СИСТЕМА ПРОГНОЗУВАННЯ РИЗИКУ ДІАБЕТУ 2 ТИПУ НА ОСНОВІ МЕДИЧНИХ ДАНИХ

Науковий керівник: д.е.н., професор Матійчук Л.П.
Kit I.
Ternopil Ivan Puluj National Technical University

INTELLIGENT SYSTEM FOR PREDICTING TYPE 2 DIABETES RISK BASED ON MEDICAL DATA

Supervisor: Doctor of Economic Sciences, Professor Matiichuk L.P.

Ключові слова: прогнозування ризику захворювання, персоналізована медицина, машинне навчання.
Keywords: disease risk prediction, personalized medicine, machine learning.

Одним із ключових завдань персоналізованої медицини є своєчасне виявлення пацієнтів із підвищеним ризиком виникнення захворювання на основі аналізу доступних історичних даних – електронних медичних записів (EHR) та, за наявності, генетичного профілю. Цукровий діабет 2 типу (T2D) є одним із найпоширеніших хронічних захворювань, що охоплює понад 500 мільйонів людей у світі. Його поступовий розвиток протягом багатьох років створює можливості для раннього втручання та профілактики. Для прогнозування ризикузастосовуються сучасні комп'ютерні методи обробки великих масивів інформації – електронних медичних записів (EHR) та, за наявності, генетичних профілів. Використання алгоритмів машинного навчання та інтелектуальних систем дозволяє створювати прикладні моделі прогнозування, що можуть бути інтегровані у медичні інформаційні системи та використовуватися для розробки персоналізованих програм профілактики.

Мета дослідження полягає у розробці алгоритму прогнозування 10-річного ризику розвитку цукрового діабету 2 типу на основі інтеграції історичних клінічних даних та поліморфізмів TCF7L2 (rs7903146) і KCNJ11 (rs5219), що дозволить формувати індивідуальні профілактичні рекомендації для пацієнтів групи підвищеного ризику.

Задача формулюється як бінарна класифікація: на основі базових ознак пацієнта (14 предикторів) визначити ймовірність P(T2D | X) розвитку захворювання протягом 10-річного періоду. Простір ознак X інтегрує три класи інформації:

- демографічні та антропометричні показники: вік, стать, індекс маси тіла (BMI), співвідношення талія/стегно;
- базові лабораторні аналізи з EHR: HbA1c, глюкоза натще, ліпопротеїди високої щільності (ЛПВЩ), тригліцериди, систолічний артеріальний тиск;
- анамнез та спосіб життя: сімейний анамнез діабету, статус куріння, фізична активність;
- генетичні маркери: кількість ризикових алелів TCF7L2 (rs7903146) та KCNJ11 (rs5219), що впливають на функцію β-клітин підшлункової залози.

Основна модель – регуляризована логістична регресія, що оцінює ймовірність розвитку T2D через сигмоїду від лінійної комбінації стандартизованих ознак (1):

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta \cdot X)}} \quad (1)$$

де β - вектор вагових коефіцієнтів для 14 ознак (клінічних та генетичних), β_0 - зсув. Для підвищення стійкості моделі використано L2-регуляризацію з параметром λ = 0.001.

Додатково проведено порівняльний аналіз ефективності градієнтного бустингу та комбінованих ансамблевих методів, що дозволило оцінити їхню продуктивність у задачі прогнозування ризику розвитку цукрового діабету 2 типу. Такий підхід забезпечує практичну перевірку різних алгоритмічних стратегій та сприяє вибору оптимальної моделі для інтеграції у прикладні медичні інформаційні системи. (2):

$$P_{ensemble}(y = 1 \mid X) = 0.5 \cdot P_{LR}(y = 1 \mid X) + 0.5 \cdot P_{GB}(y = 1 \mid X) \quad (2)$$

Для подолання класового дисбалансу (частка випадків ~11%) у gradient boosting використано зважування класів за принципом balanced (3):

$$w_{pos} = \frac{N}{2\,N_{pos}}, \quad w_{neg} = \frac{N}{2\,N_{neg}} \quad (3)$$

Розроблено регуляризовану логістичну регресію для прогнозування 10-річного ризику розвитку цукрового діабету 2 типу з показником AUC-ROC = 0.785 (95% ДІ: 0.760–0.810) та Brier score 0.084, що відповідає рівню сучасних моделей на реальних клінічних даних. Підтверджено градієнтний зв'язок між кумулятивним генетичним ризиком (TCF7L2 + KCNJ11) та частотою розвитку захворювання (7.6% → 20.0%), що формує стратифікаційне ядро дисертаційної концепції персоналізованої медицини. Алгоритм слугує першим етапом двоетапної системи: спочатку - ідентифікація високоризикових пацієнтів для профілактичних втручань, далі - адаптивний підбір терапії для тих, хто захворів. Подальші дослідження будуть спрямовані на валідацію моделі на реальних даних UK Biobank (понад 40 000 випадків T2D з генотипними даними), розширення простору генетичних ознак до повного поліген-ризик-скору (PRS) та інтеграцію методів Explainable AI (SHAP) для формування індивідуалізованих профілактичних рекомендацій.

Література

1. Sudlow C., Gallacher J., Allen N., et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Medicine, 2015. Vol. 12, No. 3. e1001779.
2. Grant S. F. A., Thorleifsson G., Reynisdottir I., et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nature Genetics, 2006. Vol. 38. P. 320–323.
3. Gloyn A. L., Pearson E. R., Antcliff J. F., et al. Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. New England Journal of Medicine, 2004. Vol. 350. P. 1838–1849.
4. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. P. 785–794.
5. Dolezalova N., Reed A. B., Despotovic A., et al. Development of a dynamic type 2 diabetes risk prediction tool: a UK Biobank study. arXiv:2104.10108, 2021.
6. Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 2017. P. 4765–4774.