КІБЕРНЕТИКА та КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ

УДК 004.8:616.1 DOI:10.34229/2707-451X.24.4.7

М.О. ПРАЗДНІКОВА

ПРОГНОЗУВАННЯ І ОЦІНКА РИЗИКУ ІНФАРКТУ МІОКАРДУ ЗА СУКУПНІСТЮ ТЕКСТІВ ЛІКАРСЬКИХ ВИСНОВКІВ

Запропоновано підхід до прогнозування ризику інфаркту на основі аналізу повнотекстових лікарських висновків, побудовано модель на основі байєсівського методу для оцінки ризику на засадах аналізу сукупності медичних заключень.

Ключові слова: *інфаркт міокарда, прогнозування ризику, машинне навчання, обробка природної мови, база медичних даних.*

Вступ. За даними Всесвітньої організації охорони здоров'я, серцеві захворювання щорічно стають причиною смерті понад 17 мільйонів людей у всьому світі [1]. Найпоширеніша причина серцевого нападу – це інфаркт міокарда (ІМ) внаслідок утворення тромбу в коронарній артерії, що призводить до її часткової або повної закупорки. Відомі фактори ризику, які найбільше впливають на розвиток цієї хвороби: куріння, вік, стать, сімейний анамнез, високий артеріальний тиск, високий рівень холестерину, діабет, ожиріння, малорухливий спосіб життя та стрес. Проте їх недостатньо для чисельної оцінки ризику інфаркту в короткостроковій, чи середньостроковій перспективі, що дало б можливість застосувати ефективні профілактично-лікувальні заходи.

В роботі розглянуто застосування машинного навчання для прогнозування ризику інфаркту та запропонована байєсівська модель оцінки ризику на засадах аналізу сукупності текстів медичних заключень, що побудована на даних реєстру медичної інформації «Ескулап» Державного управління справами України (ДУС) та проходить випробування у Державній установі «Центр Інноваційних технологій в охороні здоров'я» ДУС (ДУ ЦІТОЗ ДУС).

1. Пов'язані роботи

Задача визначення ризику інфаркту та інших серцево-судинних захворювань відома давно, і до неї застосовувались різноманітні підходи. Наприклад, у дослідженні [2] були розроблені моделі прогнозування ішемічної хвороби серця на основі машинного навчання за методами гауссового наївного Байєсу та випадкового лісу, які застосовувалися до набору даних, що складається з 303 записів і 13 обраних атрибутів. Результати дослідження показали перевагу першого метода щодо точності, прецизійності, *F*-міри та відгуку.

В роботі [3] порівнюються 4 алгоритми машинного навчання для прогнозування серцевих захворювань: SVM, логістична регресія, наївний Байєс та XGBoost. Найкращих результатів у застосуванні машинного навчання для 13 ознак у наборі даних досяг ансамблевий

метод XGBoost. Продуктивність методів машинного навчання були покращені завдяки попередній обробці даних за допомогою PCA.

У статті [4] досліджується застосування штучного інтелекту для аналізу електрокардіограм (ЕКГ) з метою прогнозування обструктивної хвороби коронарних артерій у пацієнтів зі стабільною стенокардією. Результати показують, що глибоке навчання значно підвищує точність прогнозування в порівнянні з традиційними методами машинного навчання, що робить його перспективним інструментом для ранньої діагностики серцево-судинних захворювань.

Новизна даної роботи у порівнянні з наведеними та багатьма іншими подібними дослідженнями полягає не у виборі методів, а у виборі даних. На відміну від дослідження результатів ЕКГ та інших спеціалізованих кардіологічних досліджень, що добре вміють робити лікарі, тут розглядається ширший та попередньо оброблений матеріал – сукупність всієї відомої медичної інформації про пацієнта, яка наявна у слабко структурованій формі текстів лікарських висновків за результатами огляду пацієнта з їх стійкою термінологією і одноманітним стилем подання матеріалу. Вони містять вичерпну узагальнену інформацію про анамнез, основний та побічні діагнози, заходи лікування тощо. Модель, що описана далі, це модель обробки таких текстових даних. При прогнозуванні ми на даному етапі ігнорували часовий аспект розгортання хвороби, залишивши побудову моделей прогнозування ризиків на певний період для наступної роботи. Довжина прогнозного періоду не впливає на метод прогнозування і характер моделі, а лише обмежує за часом сукупність даних для навчальної вибірки.

2. Матеріали

Деперсоналізований фрагмент бази даних відвідувань для цього дослідження був підготовлений ДНУ ЦІТОЗ ДУС та переданий Інституту кібернетики імені В.М. Глушкова НАН України у рамках договору про співробітництво. Цей фрагмент бази містить лікарські записи за даними амбулаторних відвідувань пацієнтів за 2013 – 2023 роки (773164 окремих відвідування) з текстами результатів огляду, ідентифікатором пацієнта та кодами діагнозу за МКХ (міжнародна класифікація хвороб) (від 1 до 3 кодів).

База даних містить 193116 лікарських записи про відвідування пацієнтами, що перенесли інфаркт (коди I21–I24 за МКХ): 63334 до ІМ і 129782 – після ІМ. У базі даних також є 580038 лікарських записів відвідувань пацієнтами, у яких відсутній код МКХ I21–I24. Проблема бази це те, що лікарі не завжди вносили коди діагнозів, тобто певна частка кодів I21–I24 була пропущена. Це ускладнило підготовку навчальної множини, вимагаючи відфільтровування підозрілих пацієнтів з навчальної множини.

3. Підготовка даних

В основу аналізу даних і наступної побудови моделі було покладено класичний підхід на основі векторів слів у формі словників ключ-значення, де ключем є слово, а значенням – кількість його екземплярів у тексті.

Для ототожнення слів у різних граматичних формах використовувався алгоритм Левенштейна [5]. Далі всі операції над множинами слів неявно передбачають ототожнення словоформ за алгоритмом Левенштейна, тобто всі слова у множині є унікальними в тому сенсі, що жодний член множини не може бути приблизно відповідним до іншого члена цієї множини. (Для однозначності, у випадку ототожнення як основне обирається довше з двох слів, а при однаковій довжині – те, що передує за алфавітним порядком). Перетин двох множин включатиме слова, вибрані в результаті ототожнення слів з різних множин, об’єднання множин передбачає на останок виключення тих, що не є унікальними.

За кодами діагнозу хвороби системи кровообігу (ХСК) пацієнтів було поділено дві основні множини: ті, хто перенесли ІМ, і ті, у кого ІМ ймовірно не було. За датою візиту лікарські висновки пацієнтів з визначеним ІМ було поділено на дві підмножини: до першої появи коду I21–I24 і починаючи з неї.

Нехай лікарські висновки i-го пацієнту, у якого на певному етапі в базу записувався код ІМ, об'єднані у два тексти: до першої появи кодів ІМ (позначимо його b_i), та починаючи з моменту визначення ІМ (позначимо його r_i). Для пацієнтів без жодного коду ІМ в історії відвідувань ми маємо лише один об'єднаний текст, що відповідає b_i, інший – пустий.

Позначимо функцію, що повертає множину унікальних слів у тексті t через $K(t)$, множину всіх слів у тексті t через $A(t)$, а функцію, що повертає кількість екземплярів слова s у тексті t через $N(t,s)$. Тоді вектор слів тексту t можна представити як множину кортежів $\{\langle s, N(t,s)\rangle | s \in K(t)\}$. Позначимо кількість елементів множини A через $\|A\|$.

Алгоритм підготовки даних для навчання складається з наступних кроків.

Крок 1. Побудова початкових множин характерних слів з висновків пацієнтів з відкиданням випадкових слів та помилок друку на основі мінімальної кількості N_{min} та з відкиданням службових слів на основі мінімальної частоти $0 < C_{max} < 1$:

$$B_i = \{s \mid s \in K(b_i) \text{ та } N(b_i, s) \geq N_{min}, \text{ та } N(b_i, s) < \|A(b_i)\| \times \mathrm{C}_{\max}\}; \quad (1)$$

$$R_i = \{s \mid s \in K(r_i) \text{ та } N(r_i, s) \geq N_{min}, \text{ та } N(r_i, s) < \|A(r_i)\| \times \mathrm{C}_{\max}\}. \quad (2)$$

Крок 2. Побудова об'єднаної множини слів-ознак можливого інфаркту:

$$Y = \sum\{R_i / B_i\}. \quad (3)$$

Крок 3. Побудова множини ідентифікаторів пацієнтів для навчання:
множина ідентифікаторів пацієнтів з визначеним ІМ:

$$T = \sum\{i | R_i \neq \emptyset\}, \quad (4)$$

множина ідентифікаторів пацієнтів без ознак ІМ:

$$F = \sum\{i | R_i \neq \emptyset \text{ та } B_i \cap Y = \emptyset\}. \quad (5)$$

Крок 4. Побудова об'єднаної множини потенційних слів-ознак можливого ІМ:

$$W = \sum\{B_i | i \in T\} \cap \sum\{B_i | i \in F\}. \quad (6)$$

Крок 5. Обмеження мінімальної підтримки S_{min} та максимальної підтримки $0 < C_{max} < 1$ потенційних слів-ознак можливого ІМ в обох множинах:

$$\mathrm{X} = \sum \left\{ w \;\middle|\; \begin{array}{c} w \in W \text{ та } \|\{i | i \in T \text{ та } w \in B_i\}\| \geq S_{min} \text{ та } \|\{i | i \in F \text{ та } w \in B_i\}\| \geq S_{min} \text{ та} \\ \|\{i | i \in T \text{ та } w \in B_i\}\| < \|T\| \times C_{max} \text{ та } \|\{i | i \in F \text{ та } w \in B_i\}\| < \|T\| \times C_{max} \end{array} \right\} \quad (7)$$

4. Аналіз даних

Після виключення з аналізу «підозрілих» пацієнтів, заключення яких містили слова, що в інших пацієнтів з'явились лише після перенесеного ІМ (крок 3). Розмір вибірки пацієнтів без ІМ зменшився до $\|F\|$ = 16873 осіб. Розмір вибірки інфарктників не змінився: $\|T\|$ = 893.

Початкова кількість унікальних слів відібраних з висновків пацієнтів без ІМ (F) становила понад 17 мільйонів, а у пацієнтів до ІМ (T) було визначено 72597 унікальних слів. Після знайдення перетину тих двох наборів W, та обмеження підтримки (S_{min} = 6 пацієнтів) був сформований остаточний набір X з 4731 слова, що склав множину атрибутів для наступного навчання.

Автоматично відібрані слова демонстрували суттєву відмінність у розподілі частот вживання між групами пацієнтів T та F (рис. 1) та у межах кожної окремої групи (рис. 2).

Показані на рис. 1 кругові діаграми відображають співвідношення частотності характерних слів "холестерин", "підвищеної", "вузли", "патологічні" серед двох груп пацієнтів: з інфарктом та без нього. (Для наочності вони нормовані до 1, чи повного кола в сумі). Можна побачити, що певні слова частіше зустрічаються серед пацієнтів з інфарктом, і це відповідає інтуїтивним очікуванням. Це порівняння свідчить про те, що певні терміни можуть бути індикаторами серцево-судинних ризиків, і

їх частота може використовуватися як непрямий показник стану здоров'я пацієнтів або як додатковий засіб для діагностики.

На рис. 3 – 5 показано, що побудувати на їх основі розділення двох класів – задача не проста: ані безпосереднє використання їх як атрибутів, а ні застосування перетворень простору (PCA [6], *t*-SNE [7], UMAP [8]) не призводять до помітного структурного розділення класів навіть за кращими атрибутами.

Отже особливості даних у даному дослідженні – відносно невелика вибірка, велика кількість атрибутів, неможливість досягти радикального зменшення розмірності простору атрибутів.

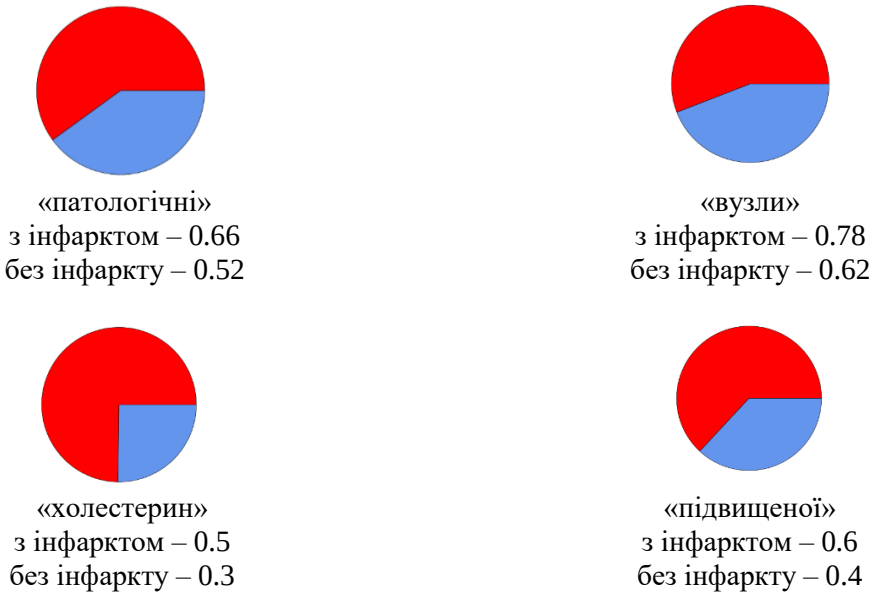

РИС. 1. Частоти появи слів у лікарських висновках залежно від перспективи ІМ

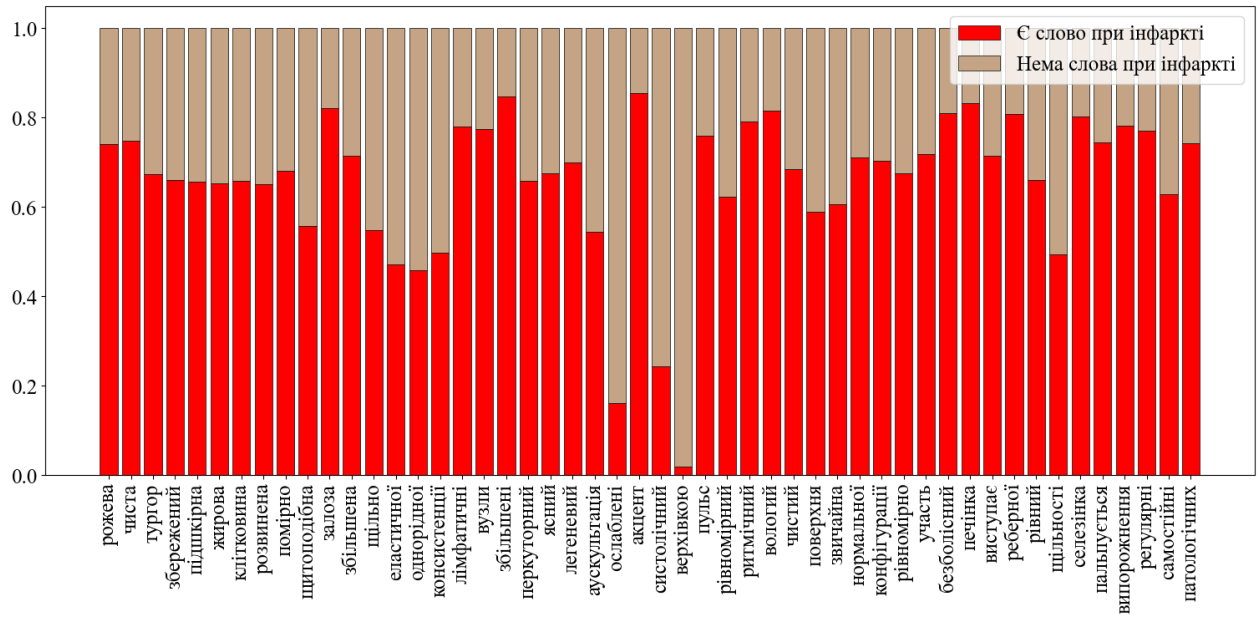

РИС. 2. Частота появи окремих слів у лікарських висновках за перспективи перенесення ІМ

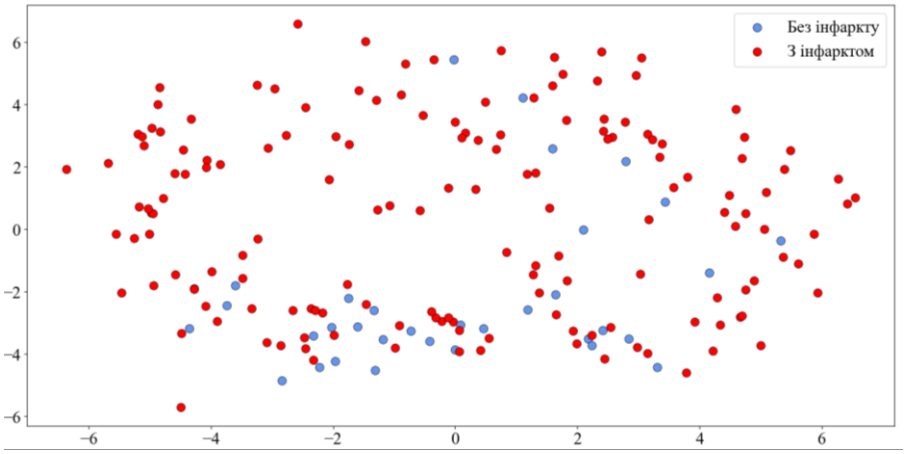

РИС. 3. Метод головних компонент (PCA)

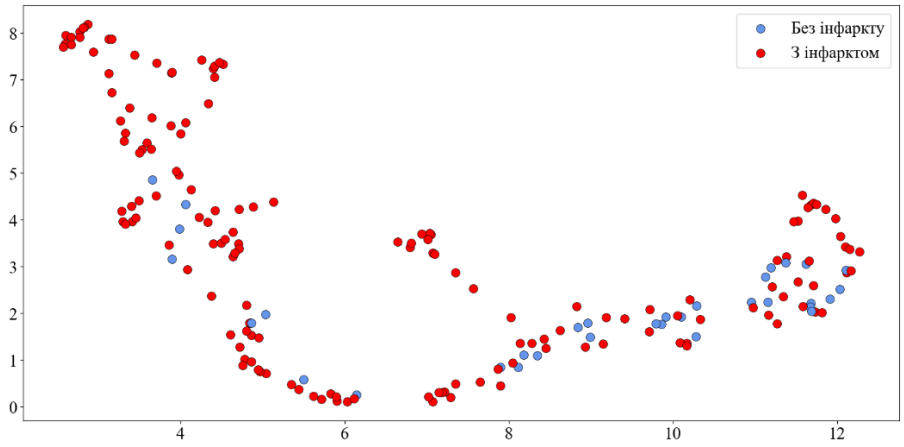

РИС. 4. Уніформне наближення і проєкція багатовимірних даних (UMAP)

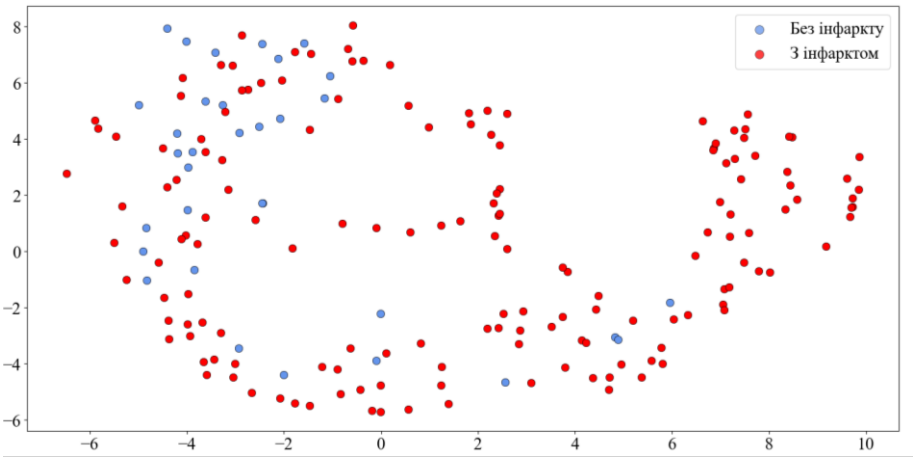

РИС. 5. *t*-розподілене стохастичне вбудовування сусідів (*t*-SNE)

5. Метод і модель

В основу рішення покладено наївний метод Байєса [8] з невеликими модифікаціями. Цей метод дозволяє працювати з великою кількістю незалежних атрибутів і добре обробляє асиметричні навчальні вибірки (із значною різницею у кількості представників різних класів). Натомість він втрачає ефективність при складній геометрії роздільної границі.

У даній задачі незалежність атрибутів (слів у висновках) абсолютно точно відсутня. Цей факт навіть не потребує статистичної перевірки, оскільки це природний зв'язок між діагнозом, ліками, та аналізами, між головним та побічними захворюваннями. Тут всі атрибути корелюють з усіма. Відомі підходи до боротьби з цією проблемою полягають в ортогоналізації (за допомогою вказаних методів PCA, *t*-SNE, UMAP), чи якнайменше у визначені та виключені залежних атрибутів шляхом попарного кореляційного аналізу. Але це радикально зменшить, чи знищить вплив більшості атрибутів, тому ми пішли іншим шляхом, залишивши всі слова для перевірки. Гіпотеза, покладена нами в основу цього рішення базується на розумінні характеру впливу кореляції на результат. Застосування 100 % корельованого (того ж самого) атрибута підносить відповідне відношення ймовірностей у квадрат. І взагалі, кореляція працює як ступінь. Але при великій кількості атрибутів вплив кореляції на обидва класи (чи на чисельник і знаменник відношення умовних ймовірностей) має за законом великих чисел схилятись до середнього (нівелюватись). Отже за рахунок ігнорування кореляції наш метод є гібридом методу Байєса з методом Парето, а висока ступінь групового впливу корельованих параметрів робить класифікацію однозначною. (Це вливає на оцінку ризику, див. розділ 6).

Модель складається з множини кортежів, що безпосередньо обчислюються за навчальною вибіркою:

$$\boldsymbol{M} = \{\langle w, P_T(w), Q_T(w), P_F(w), Q_F(w)\rangle | w \in X\},$$

де

$$P_T(w) = \|\{i | i \in T \text{ та } w \in B_i\}\| / \|T\| \quad (8)$$

– оцінка ймовірності наявності слова w у групі T,

$$Q_T(w) = 1 - P_T(w) \quad (9)$$

– оцінка ймовірності відсутності слова w у групі T,

$$P_F(w) = \|\{i | i \in F \text{ та } w \in B_i\}\| / \|F\| \quad (10)$$

– оцінка ймовірності наявності слова w у групі F,

$$Q_F(w) = 1 - P_F(w) \quad (11)$$

– оцінка ймовірності відсутності слова w у групі F, тобто до кожного слова-атрибута додаються оцінки 4-х умовних ймовірностей.

Застосування моделі $\boldsymbol{M}$ до об'єднаного тексту t_j висновків за всіма відвідуваннями пацієнта j полягає у обчислені

$$R(t_i) = R_0 \times \prod \{r(w, t_i) | w \in X\},$$

де R_0 – відношення апріорних імовірностей майбутнього ІМ та його відсутності, може бути оцінена за всією базою даних (не тільки навчальною вибіркою) як

$$R_0 = \|T\| / \|F\|;$$
$$r(w,t) = P_T(w)/P_F(w), \text{ якщо } w \in K(t),$$
$$r(w,t) = Q_T(w)/Q_F(w), \text{ інакше.} \quad (12)$$

Якщо $R(t_j) > 1$, то приймається гіпотеза про високий ризик інфаркту, а якщо менше, то про низький ризик інфаркту.

6. Експериментальні результати

Розраховану модель було перевірено на збалансованій випадковій тестовій вибірці з 900 пацієнтів без очікуваного ІМ і 893 пацієнтів з очікуваним ІМ [9]. Модель змогла правдиво визначити великий ризик ІМ для 714 осіб (80,0 %) і правдиво визначити низький ризик ІМ для 734 осіб (81,6 %). Помилки першого і другого роду становили 20,0 % і 18,4 % відповідно (таблиця). Близька до 1 площа під ROC-кривою 0,898 свідчить про високу надійність даного методу – його стійкість до змін порогового значення (рис. 6).

З технічного завдання та практичного використання корисна – плавна (чи бальна) оцінка відносного ризику ІМ, зокрема з обчисленням (12) для випадку $R_0 = 1$, тобто для порівняння ризику пацієнту з середнім. Позначимо ризик ІМ пацієнта j, як p_j:

$$R(t_j) = p_j/(1 - p_j), \text{ звідки } p_j = R(t_j)/(1 + R(t_j)). \quad (13)$$

Проте експерименти показали, що формула (13) не дає проміжних значень ризику, а лише значення дуже близькі до 0 та 1 внаслідок впливу сильної кореляції атрибутів.

ТАБЛИЦЯ. Результати тестування демонструють ефективність побудованої моделі

	Прогноз	
Факт	Інфаркт	Ні
Інфаркт	TP = 80.0 %	FN = 20.0 %
Ні	FP = 18.4 %	TN = 81.6 %

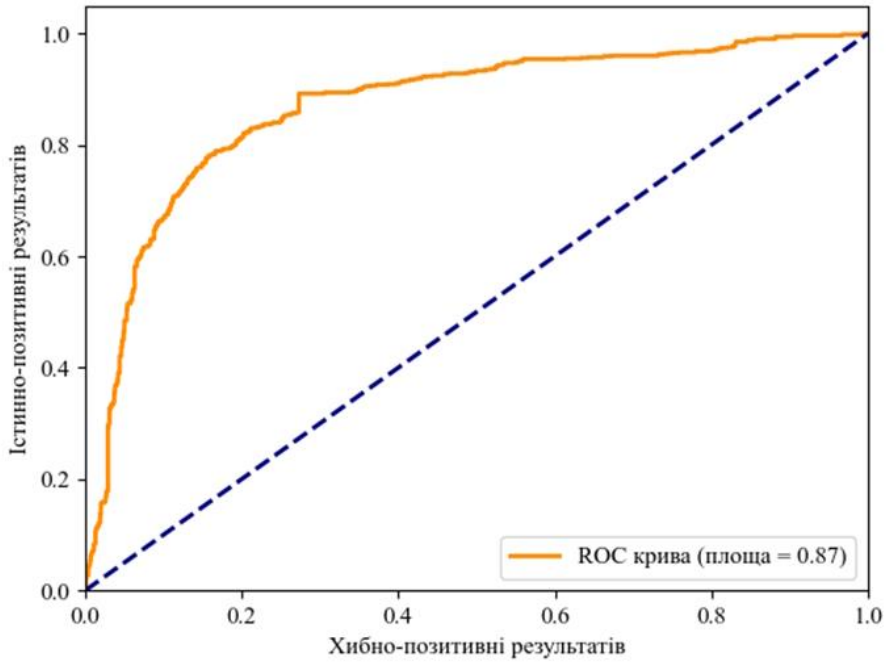

РИС. 6. ROC-крива демонструє високу надійність побудованої байєсівської моделі

Враховуючи степеневий характер впливу кореляції, формулу (13) було перероблено в напрямку наближення до міри Парето шляхом переходу від значення відношення $R(t_j)$ до його логарифму. Експериментально підібрана евристична формула ризику:

$$p_j = R_j^*/\left(1 + \left|R_j^*\right|\right) + 1, \text{ де } R_j^* = lgR(t_j)/100\,. \quad (14)$$

За формулою ризику (14) на основі послідовного поділу медіаною встановлено такі границі груп ризику:

- $p_j \le 1$ – ризик низький,
- $1 < p_j \le 1{,}3$ – ризик помірний,
- $1{,}3 < p_j \le 1{,}5$ – ризик високий,
- $p_j > 1{,}5$ – ризик дуже високий (гострий).

Висновки

Машинне навчання, що використовує статистичні методи та штучний інтелект, представляє собою перспективний додатковий інструмент для медичних спеціалістів. Його головна цінність полягає не в заміні традиційних методів діагностики, а у наданні відповіді на складні питання, на які традиційна медицина може не завжди надати чіткі відповіді.

Зокрема, машинне навчання може бути використане для оцінки таких ризиків:

- ймовірність загострення або ускладнень: це включає прогнозування ризику виникнення загострення хвороби або ускладнень протягом різних часових проміжків, таких як місяць, квартал або рік.
- агресивність хвороби: оцінка ймовірності того, що хвороба буде прогресувати в агресивній формі, може суттєво вплинути на вибір стратегії лікування.
- ризик смерті пацієнта: прогнозування ймовірності смерті пацієнта протягом визначеного періоду часу є критично важливим для прийняття рішень щодо подальшого лікування.

Інтеграція таких прогнозів у клінічну практику може допомогти лікарям у виборі найбільш ефективної тактики лікування, будь то консервативне або хірургічне втручання. Машинне навчання, обробляючи великі обсяги невизначених даних, пропонує рішення, які наближаються до інтуїтивного підходу, але з перевагою обґрунтованості, достовірності та строгості.

Таким чином, хоча методи машинного навчання можуть бути схожими на інтуїтивні рішення, вони забезпечують науково обґрунтовані прогнози, які допомагають зменшити невизначеність і покращити якість медичного обслуговування. Це дозволяє лікарям приймати більш інформовані рішення, що можуть значно вплинути на результат лікування і загальний стан пацієнтів.

Що можна покращити. Подальші дослідження передбачають розвиток запропонованого у даній роботі підходу в двох напрямках. По-перше, йдеться про введення часової складової у прогнозування, зокрема, прогнозування ризику інфаркту протягом року, кварталу, можливо місяця. Це не потребуватиме зміни підходу, але вимагатиме певної переробки моделі і перерахунків. По-друге, можливо доповнення переліку атрибутів дискретизованими результатами досліджень (типу параметрів ЕКГ). Це потенційно може підняти точність прогнозу. По-третє, йдеться про покращення існуючої моделі за рахунок більш точного аналізу атрибутів моделі. Можливо доцільно позбутися частини слів, що незначно впливають на результати прогнозування (з метою пришвидшення обчислень). Нарешті, завжди залишається можливість покращити якість прогнозування за рахунок подальших експериментів з методами машинного навчання, зокрема, ансамблевими.

Наявність даних. ДНУ ЦІТОЗ ДУС є власником оригінальної бази даних «Ескулап» та її деперсоналізованого фрагмента, що був використаний в даній роботі. Договір про співробітництво, в рамках якого був отриманий доступ до даних не дає нам право розповсюджувати їх, чи викладати у відкритий доступ.

Фінансування. Автор не отримував додаткового фінансування для проведення досліджень та написання статті.

Список літератури

1. Кабінет Міністрів України. Як виявити інфаркт і що робити при серцевому нападі: коментує експерт. 2024 липня 24. https://kmu.gov.ua/news/yak-viyaviti-infarkt-i-shcho-robiti-pri-sercevomu-napadi-komentuye-ekspert (звернення: 24.07.2024)
2. Bemando C., Miranda E., Aryuni M. Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms. *International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management*. 2021. P. 232–237.
3. Nandal N., Goel L., Tanwar R. Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis. 2022. P. 5–17.
4. Park J., Kim J., Kang S.-H., Lee J., Hong Y., Chang H.-J., et al. Artificial intelligence–enhanced electrocardiography analysis as a promising tool for predicting obstructive coronary artery disease in patients with stable angina. *European Heart Journal - Digital Health*. Vol. 5, Iss. 4. P. 444–453. https://doi.org/10.1093/ehjdh/ztae038
5. Медведєв М.Г. Відстань Левенштейна та пов'язані з нею задачі. *Наукові записки НаУКМА. Комп'ютерні науки*. 2000. Т. 18. С. 33–37. https://ekmair.ukma.edu.ua/server/api/core/bitstreams/c1190351-8225-469b-9d01-3eabafb6fee4/content
6. Jain Y. Principal Component Analysis (PCA) [NLP, Python]. 2024 https://medium.com/@yashj302/principal-component-analysis-pca-nlp-python-ce9caa58bd7a (звернення: 09.09.2024)
7. Suman S. Text Data Pre-Processing Using Word2Vector and t-SNE. 2024. https://medium.com/swlh/text-data-pre-processing-using-word2vector-and-t-sne-2321fbce5b9 (звернення: 09.09.2024)
8. Becht E., Dutertre C.-A., Kwok I.W.H., Ng L.G., Ginhoux F., Newell E.W. Evaluation of UMAP as an alternative to t-SNE for single-cell data. *The preprint server of biology*. April 10, 2018. https://doi.org/10.1101/298430
9. Nigam K., McCallum A., Thrun S., Mitchell T. Learning to classify text from labeled and unlabeled documents using EM. *Machine Learning*. 2000. **39**. P. 103–134.
10. Празднікова М.О., Кравченко А.М., Тульчинський В.Г., Чайковський І.А. Використання бази даних ДНУ НЦППКМ ДУС для прогнозування ризику інфаркту міокарда: від ідеї до реалізації. Тези доповідей Міжнародної науково-практичної конференції «Організаційні та клінічні аспекти пацієнт-орієнтованого підходу до лікування та реабілітації в сучасних умовах», 29-30 травня 2024. С. 152. https://ua-medical.com/downloads/conferences/abstracts-2024.pdf

Одержано 24.09.2024

Празднікова Маргарита Олександрівна,
аспірантка Інституту кібернетики імені В.М. Глушкова НАН України, Київ.
prazdnikovamargarita@gmail.com

УДК 004.8:616.1

М.О. Празднікова

Прогнозування і оцінка ризику інфаркту міокарду за сукупністю текстів лікарських висновків

Інститут кібернетики імені В.М. Глушкова НАН України, Київ
ЛИСТУВАННЯ: prazdnikovamargarita@gmail.com

Вступ. Інфаркт міокарда залишається однією з провідних причин смерті у світі, викликаючи ушкодження серцевого м'яза через раптове порушення кровопостачання. В основі розвитку інфаркту лежать фактори ризику, такі як куріння, вік, стать, високий рівень холестерину, діабет та інші. Незважаючи на значний прогрес у методах діагностики та лікування, завчасне прогнозування ризику інфаркту залишається важливим завданням, яке може значно знизити смертність і покращити якість життя пацієнтів. У цій статті розглянуто підхід до прогнозування ризику інфаркту на основі аналізу текстових даних лікарських висновків за допомогою машинного навчання.

Мета роботи – розробка та впровадження ефективної моделі прогнозування ризику інфаркту міокарда шляхом аналізу великих обсягів медичних даних. Використовуючи деперсоналізовану базу даних ДНУ ЦІТОЗ ДУС, що містить лікарські записи за десятирічний період, дослідження спрямоване на виявлення ключових факторів та патернів, які можуть свідчити про підвищений ризик інфаркту. Застосування методів машинного навчання, зокрема наївного байєсівського класифікатора, дозволить оцінити ефективність таких підходів у медичній практиці та визначити їх потенціал для інтеграції у системи підтримки клінічних рішень.

Результати. Запропонована модель прогнозування показала високу ефективність у виявленні пацієнтів з підвищеним ризиком інфаркту. Аналізуючи частоту появи певних слів у медичних записах, алгоритм зміг передбачити високий ризик інфаркту для 80 % пацієнтів з очікуваним інфарктом. Це свідчить про значний потенціал використання текстових даних та методів машинного навчання для медичної діагностики. Крім того, зниження кількості помилкових прогнозів підкреслює надійність моделі та її придатність для практичного застосування.

Висновки. Використання машинного навчання для прогнозування ризику інфаркту на основі аналізу бази даних є перспективним напрямком у медичній практиці. Запропонований метод дозволяє підвищити точність діагностики та прогнозування, що може суттєво вплинути на вибір стратегії лікування та покращення результатів для пацієнтів. Інтеграція таких інструментів у клінічну практику сприятиме більш інформованому прийняттю рішень лікарями та зниженню ризиків для пацієнтів.

Ключові слова: інфаркт міокарда, прогнозування ризику, машинне навчання, база даних, наївний байєсівський класифікатор, медична аналітика.

MSC 68T50, 68T20

Margaryta Prazdnikova

Prediction and Assessment of Myocardial Infarction Risk on the Base of Medical Report Text Collection

V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv
Correspondence: prazdnikovamargarita@gmail.com

Introduction. Myocardial infarction remains one of the leading causes of death worldwide, resulting from sudden disruption of blood supply to the heart muscle. Key risk factors include smoking, age, gender, high cholesterol levels, diabetes, and others. Despite advancements in diagnostics and treatment, early detection of heart attack risk is crucial for reducing mortality and improving patient quality of life. This paper explores an approach to predicting heart attack risk based on analysis of text data of medical reports using machine learning.

The purpose of the article is to demonstrate how the application of machine learning, particularly the Naive Bayes classifier, can enhance the prediction of myocardial infarction risk through the analysis of extensive medical data. By leveraging a depersonalized database from SSO CITHC SAA, containing medical records collected during a decade of operating, this study seeks to reveal how the identification of critical patterns and factors can improve prediction accuracy. Additionally, the article explores how integrating these predictive models into clinical decision support systems can refine medical diagnostics and decision-making processes.

Results. The proposed prediction model demonstrated high efficiency in identifying patients at increased risk of heart attack. By analyzing the frequency of specific words in medical records, the algorithm successfully predicted a high risk of heart attack for 80 % of patients with an expected event. This underscores the significant potential of leveraging textual data and machine learning methods in medical diagnostics. Moreover, the reduction in false predictions highlights the model's reliability and suitability for practical application.

Conclusions. Employing machine learning for heart attack risk prediction based on medical data analysis represents a promising direction in modern medicine. The developed model showcases the possibility of enhancing diagnostic and predictive accuracy, which can substantially influence treatment strategy decisions and improve patient outcomes. Integrating such tools into clinical practice will facilitate more informed decisions by physicians and reduce patient risks.

Keywords: myocardial infarction, risk prediction, machine learning, database, Naive Bayes classifier, medical analytics.