

АВТОМАТИЗАЦІЯ, ІНТЕРНЕТ РЕЧЕЙ, РОБОТОТЕХНІКА ТА ІНФОРМАЦІЙНО-ВИМІРЮВАЛЬНІ СИСТЕМИ

<https://doi.org/10.31649/1997-9266-2026-187-4-178-187>

УДК 004.85:616.379-008.64

О. В. Коротун¹
О. М. Свінцицька¹
Д. К. Марчук¹
М. Р. Штиль¹

ЗАСТОСУВАННЯ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЗНАЧЕННЯ НАЯВНОСТІ ЦУКРОВОГО ДІАБЕТУ У ПАЦІЄНТІВ

¹Державний університет «Житомирська політехніка»

Досліджено та розглянуто практичне застосування сучасних методів машинного навчання для розв'язання проблеми виявлення наявності цукрового діабету у пацієнтів. На початковому етапі робота зосереджена на попередній обробці даних: виявленню порожніх та аномальних значень, вивченню змінних набору даних тощо. Для узагальнення, організації та представлення основних характеристик набору даних використана описова статистика. Для числових змінних отримано ключові показники: середнє, медіана, стандартне відхилення, мінімальне та максимальне значення, а також квартилі. Це дозволило оцінити центральну тенденцію даних, їхній розкид, діапазон та виявити нерівномірності у розподілі. Для категоріальних змінних визначено кількість унікальних значень, найпоширенішу категорію та її частоту. Описано використання алгоритмів машинного навчання для визначення цукрового діабету у пацієнтів. Поставлено завдання передбачити наявність (результат = 1) або відсутність (результат = 0) діабету у пацієнтки на основі низки діагностичних показників, зокрема рівня глюкози, артеріального тиску, товщини шкірної складки, рівня інсуліну, індексу маси тіла, віку тощо. У межах дослідження здійснено аналіз вхідних даних з метою виявлення закономірностей, взаємозв'язків між діагностичними параметрами, аналізу розподілу ознак, а також виявлення потенційних викидів. Для побудови моделей розглянуто та застосовано кілька алгоритмів класифікації, зокрема логістична регресія, метод k-найближчих сусідів та випадковий ліс. Кінцевим етапом дослідження стало порівняння ефективності різних алгоритмів. Найкращі результати продемонструвала модель логістичної регресії з адаптованим порогом класифікації. У роботі застосовано зміщення порогу прийняття рішення від стандартного значення 0,5 до діапазону 0,25...0,3, що дозволило мінімізувати помилки другого роду (хибнонегативні результати). Це важливо для ранньої діагностики діабету, оскільки дозволяє виявити більшу кількість хворих пацієнтів, які могли бути пропущені у разі стандартних налаштувань моделі.

Ключові слова: аналіз даних, машинне навчання, прогнозування, логістична регресія, k-найближчих сусідів, випадковий ліс.

Вступ

У сучасному світі цифрових технологій та швидко зростаючих обсягів інформації, аналіз даних набуває важливого значення. Він здатний виокремлювати цінні знання, закономірності та інсайти з великих масивів даних. Саме перетворення «сирих» даних на стратегічно важливі є критично важливим для багатьох сфер людського життя, не виключенням є і медична сфера. Використання методів машинного навчання стає ключовим потужних інструментів прогнозування та класифікації даних.

Одним з найпоширеніших хронічних захворювань у світі є цукровий діабет. Його своєчасна діагностика є важливою для запобігання ускладненням та проведення ефективного лікування. Традиційні методи діагностики засновані на дослідженні окремих медичних показників. Відповідно до цього постає необхідність створення інтелектуальних систем, які зможуть швидко та точно перед-



© О. В. Коротун, О. М. Свінцицька, Д. К. Марчук, М. Р. Штиль, 2026. Ця стаття відкритого доступу та поширюється на умовах ліцензії [Creative Commons Attribution License 4.0](https://creativecommons.org/licenses/by/4.0/)

бачати наявність діабету на основі доступних медичних даних. Розвиток технологій обробки даних та методів машинного навчання (МН) надає можливість побудувати ефективні моделі прогнозування, що враховують множини ознак і виявляють приховані закономірності у великих масивах медичної інформації. Вони можуть бути використані як допоміжний інструмент у прийнятті медичних рішень.

Метою роботи є застосування алгоритмів машинного навчання для визначення наявності цукрового діабету у пацієнтів.

Сьогодні зростає увага дослідників до використання методів МН у медичній діагностиці та прогнозуванні, зокрема у такому поширеному та соціально значущому захворюванні, як діабет. Наразі він є глобальною проблемою у сфері охорони здоров'я, його раннє прогнозування наявності або ризику розвитку стає критично важливим для своєчасного втручання та запобігання ускладненням. Висвітливо праці науковців щодо застосування алгоритмів МН для аналізу даних, зокрема клінічних даних пацієнтів задля підвищення точності та ефективності діагностичного прогнозування хвороб.

Результати дослідження щодо перспективних програмних інструментів для аналізу даних, їх характеристики, переваги, недоліки та рекомендації щодо впровадження для використання у вітчизняному бізнесі подані у статті [1]. Описані інструменти є потужним засобом для покращення ефективності компанії, прийняття обґрунтованих управлінських рішень, покращення процесів і досягнення бізнес-цілей шляхом аналізу наборів даних, зокрема Big Data.

Про впровадження підходів до аналізу даних, де штучний інтелект (ШІ) посідає провідне місце, описано у роботі [2]. Визначаються ролі та напрями використання ШІ в оптимізації бізнес-стратегій та процесі прогнозування, охоплюючи широкий спектр застосувань, таких як прогнозування енергоспоживання, попиту, продажів, ланцюгів постачання тощо.

У дослідженні [3] виконано огляд застосування базових інструментів сучасного топологічного аналізу даних, зокрема, стійких гомологій, що включають діаграми стійкості, числа та криві Бетті. Як зазначають дослідники, розглянуті інструменти для аналізу та порівняння даних типу хмара точок набули широкого застосування протягом останнього десятиліття в різноманітних сферах, таких як комп'ютерний зір, біомедичний аналіз тощо.

Про актуальність та перспективи використання машинного навчання в обробці медичних даних йдеться у статті [4]. Ключовими аспектами визначені опис основних алгоритмів, а також попередня обробка та підготовка даних для забезпечення високої точності, надійності та етичності моделей. Наголошується на потребі мультидисциплінарного підходу, що об'єднує медиків, технічних фахівців та експертів з конфіденційності.

Порівняння точності діагностування захворювань легень нейромережею з діагнозами лікарів наводять дослідники з [5]. Вони протестували метод обстеження на здорових людях, використовуючи для цього комплекс теоретичних, діагностичних, емпіричних методів, математичної статистики та глибокого машинного навчання.

Застосування методів машинного навчання, зокрема лінійної регресії, для розв'язання задач прогнозування рівня щастя населення різних країн на основі їхніх соціально-економічних характеристик досліджується у [6]. У статті описано процес побудови моделі, зокрема попередню обробку даних, навчання та оцінку за допомогою метрик MSE, RMSE, MAE та R^2 , використовуючи мову Python. Результати підкреслюють, методи масштабування значно підвищують продуктивність моделі, демонструючи її здатність до узагальнення.

Аналіз методик використання технологій МН у медичній діагностиці висвітлено авторами [7], які наголошують на їхній важливості для зміни підходів до діагностики, лікування та оптимізації медичних послуг. Розглянуто ключові парадигми МН, проаналізовано основні типи медичних даних для навчання моделей, а також описано конкретні наукові підходи та алгоритми.

У дослідженні [8] йдеться про застосування методів МН для прогнозування аеродинамічних властивостей матеріалів. Запропоновано методологію класифікації, що включає етапи збирання та аналізу даних, моделювання, а також оцінювання та підвищення ефективності моделей. Для оцінки точності прогнозів визначені та проаналізовані такі метрики: точність, відгук, F-міра тощо.

Застосування моделей навчання нейронних мереж на медичних даних з метою створення дієвого інструменту для раннього та прискореного виявлення ішемічних захворювань подано у статті [9]. Це має потенційно значний вплив на поліпшення діагностики, прогнозування та лікування цієї хвороби. Результатом проведеного дослідження стала побудована нейронна мережа, яка продемонструвала високу точність у виявленні хвороби на ранніх етапах.

Використання логістичної регресії та дерева рішень, посилені алгоритмом аналізу головних

компонентів (РСА), для виявлення та прогнозування раку досліджено у [10]. У результаті отримано, алгоритм RSA, який значно підвищив точність набору даних з великою кількістю ознак. Дослідження демонструє, що зазначені методи здатні поліпшити процес діагностики онкології, сприяючи зменшенню кількості помилкових класифікацій та ранньому виявленню захворювання.

Науковцем [11] описано моделі та методи обробки даних системи віддаленого моніторингу стану пацієнтів з цукровим діабетом.

Відтак, об'єктом дослідження є побудова моделей для виявлення цукрового діабету, а також визначення їх точності щодо виявлення цього захворювання у пацієнтів. Побудова таких моделей здійснюватиметься з використанням низки діагностичних показників пацієнтів (рівень глюкози, артеріальний тиск, товщина шкірної складки, рівень інсуліну, ІМТ та вік). У дослідженні виконано аналіз вхідних даних і порівняння різних методів класифікації (логістична регресія, *k*-ближчих сусідів, випадковий ліс) для вибору найкращого підходу.

Результати дослідження

Для побудови моделі для діагностики цукрового діабету на основі аналізу медичних даних виконаємо класифікацію стану пацієнтів набору діагностичних показників (мають діабет позначимо як 1, не мають — 0). Відповідно до цього потрібно підготувати дані (очистка від пропущених значень, аномалій, нормалізація даних), проаналізувати змінні набору даних (розподіл ознак, виявлення кореляцій між змінними тощо), розробити та навчити моделі класифікації (логістична регресія, метод *K*-ближчих сусідів, випадковий ліс), оцінити ефективність побудованих моделей (порівняти точності, повноти та інші метрики для визначення найкращої моделі) та описати отримані результати для визначення найважливіших ознак для прийняття рішень, формулювання висновків щодо доцільності використання моделі в медичній практиці. Отже, задача зводиться до розробки та оцінювання моделей класифікації, здатних з високою точністю передбачати наявність діабету на основі медичних даних пацієнтів.

З платформи Kaggle вибрано набір даних з такими медичними показниками (Diabetes, <https://www.kaggle.com/datasets/sefakocakalay/diabetes>): Pregnancies (кількість вагітностей), Glucose (рівень глюкози в крові), BloodPressure (діастолічний артеріальний тиск), SkinThickness (товщина шкіри), Insulin (рівень інсуліну в крові), BMI (індекс маси тіла), DiabetesPedigreeFunction (відсоток діабету/функція діабетичної спадковості), Age (вік), Outcome (наявний діабет — 1, немає діабету — 0). Для проведення дослідження використано мову програмування Python. На етапі підготовки даних виконано перевірку на наявність пропущених даних, які ідентифіковані як нульові записи (0) у відповідному наборі даних. Для ефективної обробки виявлених пропущених значень застосована техніка імпутації. З метою збереження форми розподілу даних та цілісності взаємозв'язків між різними

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105466	20.536458	79.799479	31.982578	0.471376	33.240685	0.348958
std	3.369578	31.972618	19.356807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Рис. 1. Описова статистика змінних набору даних

ознаками, ухвалено рішення замінити кожне пропущене значення на медіанне значення відповідного стовпця. Оскільки вона є стійкою до викидів і менш схильна спотворювати загальний розподіл даних порівняно із середнім значенням.

Після проведеного етапу підготовки даних, переглянемо описову статистику змінних набору даних (рис. 1).

Для детального вивчення стру-

тури даних побудовано гістограми розподілу для всіх змінних набору даних (рис. 2).

Кількість вагітностей демонструє правосторонній асиметричний розподіл з максимальною частотою для нульових та одиничних значень (близько 250 спостережень). Зі збільшенням кількості вагітностей частота різко зменшується. Рівень глюкози в плазмі крові характеризується приблизно нормальним розподілом з центром у діапазоні 90...120 мг/дл. Спостерігається незначна правостороння асиметрія, що може вказувати на наявність підвищених значень глюкози у частини обстежених.

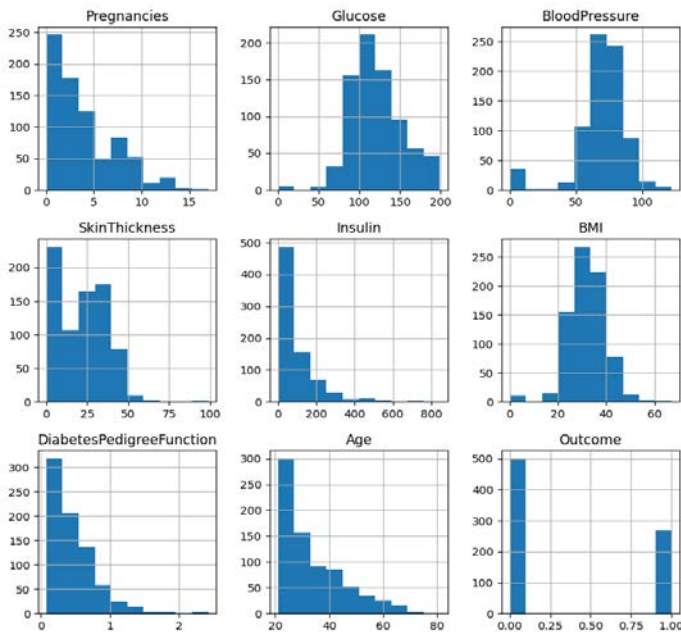


Рис. 2. Гістограми розподілу для всіх змінних набору даних

Діастолічний артеріальний тиск має основну концентрацію значень у діапазоні 60...80 мм рт.ст., що відповідає нормальним фізіологічним показникам. Проте виявлено аномальні нульові значення, які можуть свідчити про пропущені дані або помилки вимірювання. Для товщини шкірної складки спостерігається розподіл переважно в діапазоні 15...45 мм.

Рівень інсуліну демонструє експоненціальний розподіл з критично високою часткою нульових значень (близько 500 спостережень), може свідчити про технічні проблеми збору даних або використання нуля як позначення відсутніх вимірювань. Індекс маси тіла має розподіл, близький до нормального, з піком у діапазоні 30...35 кг/м², відображає поширеність надмірної ваги в обстеженій групі.

Функція діабетичної спадковості показує експоненціальний розподіл з максимальною концентрацією значень близько нуля. Більшість спостережень не перевищує 1, що вказує на відносно низький генетичний ризик діабету в більшості випадків. Вік учасників демонструє правосторонню асиметрію з піком у молодшому віковому діапазоні (20—30 років) та поступовим зменшенням частоти зі збільшенням віку до 80 років.

Цільова змінна (діагноз діабету) є бінарною з розподілом приблизно 2:1 на користь негативних випадків (близько 500 проти 270), що відображає природну поширеність захворювання в популяції та створює помірний дисбаланс класів для подальшого моделювання.

Для виявлення аномальних значень та оцінки симетричності розподілів побудовано діаграми розмаху всіх змінних набору даних та отримано такі результати (рис. 3).

– *кількість вагітностей* демонструє розподіл з медіаною близько 1—2 вагітностей, міжквартильний розмах є відносно невеликим, проте спостерігаються поодинокі викиди з високими значеннями (понад 10 вагітностей);

– *рівень глюкози* характеризується симетричним розподілом з медіаною в діапазоні нормальних значень (близько 110...120 мг/дл), спостерігаються поодинокі викиди як у бік занижених, так і завищених значень, що може вказувати на наявність як гіпо-, так і гіперглікемічних станів;

– *діастолічний артеріальний тиск* показує концентрацію основної маси даних у фізіологічно нормальному діапазоні з медіаною близько 70 мм рт.ст., виявляються численні викиди у верхній частині розподілу, що може свідчити про наявність артеріальної гіпертензії в пацієнтів;

– *товщина шкірної складки* демонструє значну концентрацію даних у нижній частині розподілу з медіаною близько 20 мм, відмічена наявність поодиноких викидів у верхній частині, може вказувати на випадки значного збільшення підшкірно-жирової клітковини;

– *рівень інсуліну* характеризується вкрай асиметричним розподілом з медіаною близько нуля, що підтверджує велику кількість нульових або низьких значень, спостерігаються численні викиди в діапазоні

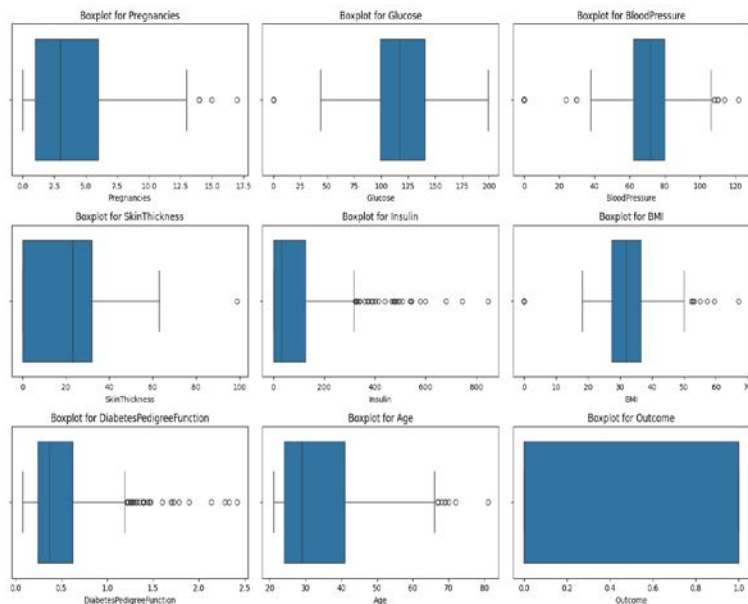


Рис. 3. Діаграми розмаху всіх змінних набору даних

– *товщина шкірної складки* демонструє значну концентрацію даних у нижній частині розподілу з медіаною близько 20 мм, відмічена наявність поодиноких викидів у верхній частині, може вказувати на випадки значного збільшення підшкірно-жирової клітковини;

– *рівень інсуліну* характеризується вкрай асиметричним розподілом з медіаною близько нуля, що підтверджує велику кількість нульових або низьких значень, спостерігаються численні викиди в діапазоні

високих значень, відображає стани інсулінорезистентності або компенсаторної гіперінсулінемії;

– *індекс маси тіла* пацієнта має відносно симетричний розподіл з медіаною у діапазоні надмірної ваги (близько 30...32 кг/м²), присутні поодинокі викиди у верхній частині розподілу вказують на наявність випадків ожиріння;

– *функція діабетичної спадковості (цільова змінна)* демонструє правосторонню асиметрію з медіаною близько 0,3...0,4, виявлені численні викиди у верхній частині розподілу свідчать про наявність осіб з високим генетичним ризиком розвитку діабету;

– *вік* пацієнтів характеризується помірною правосторонньою асиметрією з медіаною близько 30 років, його міжквартильний розмах охоплює діапазон від молодого до середнього віку, з поодинокими викидами у старшій віковій групі;

– *діагноз діабету* подано у вигляді бінарного розподілу, що відображає співвідношення позитивних та негативних випадків захворювання.

Для дослідження взаємозв'язків між змінними побудовано матрицю діаграм розсіювання з розподілами, опишемо їх. Підтверджено виявлені особливості щодо наявності правосторонньої асиметрії для кількості вагітностей, спостерігається нормальний розподіл глюкози та артеріального тиску, експоненціальні розподіли для інсуліну та функції діабетичної спадковості. Взаємозв'язки з цільовою змінною демонструють чіткі відмінності між групами: особи з діабетом (помаранчеві точки) характеризуються вищими рівнями глюкози, що є очікуваним діагностичним критерієм; спостерігається тенденція до вищих значень артеріального тиску у групі з діабетом; пацієнти з діабетом мають дещо вищі показники ІМТ та товщини шкірної складки; функція діабетичної спадковості показує помітно вищі значення у групі з позитивним діагнозом.

Міжзмінні кореляції виявляють декілька важливих зв'язків — це помірна позитивна кореляція між ВМІ та товщиною шкірної складки, що відображає загальну тенденцію до накопичення жирової тканини; слабкий зв'язок між віком та кількістю вагітностей, що є природним для жіночої популяції; відсутність сильних лінійних кореляцій між більшістю предикторів, що свідчить про їхню відносну незалежність; нелінійні взаємозв'язки спостерігаються у декількох парах змінних, що вказує на можливість використання нелінійних методів моделювання для підвищення точності прогнозування.

Викиди та аномальні значення помітні у змінних «Інсулін», «Функція діабетичної спадковості» та «Кількість вагітностей», потребує додаткової уваги під час побудови моделей.

Виявлені взаємозв'язки підтверджують клінічну обґрунтованість вибраних предикторів для діагностики діабету та вказують на необхідність комплексного підходу до аналізу, що враховує лінійні та нелінійні залежності між змінними. Чітке розмежування груп за цільовою змінною свідчить про високий потенціал для побудови ефективних класифікаційних моделей.

Діаграма розмаху порівняння вікових характеристик між групами пацієнок без діабету (0) та з діабетом (1) показана на рис. 4. Медіанний вік у групі без діабету становить приблизно 27 років, тоді як у групі з діабетом — близько 36 років, свідчить про тенденцію до збільшення ризику розвитку діабету з віком. Міжквартильний розмах у групі без діабету охоплює діапазон від 22 до 37 років. Той факт, що 75 % здорових осіб мають вік до 37 років (Q3), свідчить про відносно молодий віковий розподіл у цій групі. Міжквартильний розмах віку у групі з діабетом охоплює діапазон від 28 до 44 років. Це вказує на статистичний зсув вікової групи у бік старших вікових категорій. Розподіл викидів показує наявність поодиноких випадків діабету в молодому віці (близько 20 років) у групі з позитивним діагнозом. У групі без діабету спостерігаються викиди у старшому віці (понад 60 років), що вказує на наявність здорових осіб похилого віку.

Різниця в медіанному віці між групами (близько 9 років) є статистично та клінічно значущою. Прогностична цінність віку як предиктора діабету підтверджується чітким розмежуванням розпо-

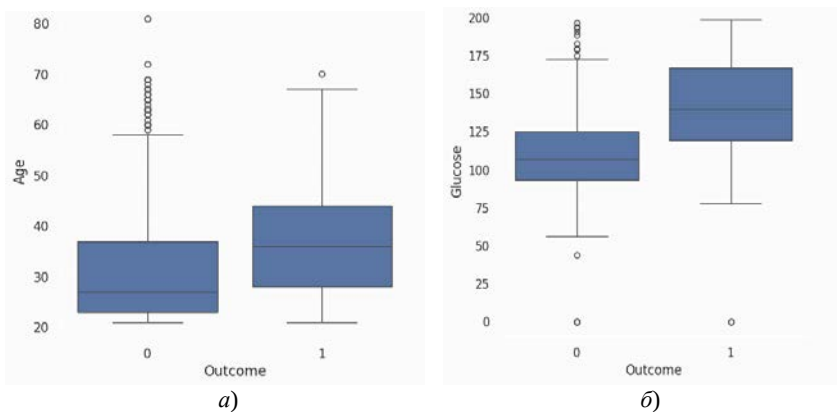


Рис. 4. Діаграми розмаху порівняння *a* — вікових характеристик між групами пацієнтів без діабету та з діабетом; *б* — глюкози між групами пацієнтів без діабету та з діабетом

ділів між групами, хоча спостерігається певне перекриття діапазонів, що вказує на необхідність використання додаткових діагностичних критеріїв. Виявлені вікові особливості мають важливе значення для розробки стратегій скринінгу та профілактики діабету, підкреслюючи необхідність ретельнішого моніторингу метаболічних показників у осіб старшого віку.

Побудована діаграма розмаху демонструє кардинальні відмінності у рівнях глюкози між групами пацієнтів без діабету (0) та з діабетом (1), що підтверджує ключову роль цього показника для виявлення цього захворювання.

Медіанні значення показують суттєву різницю, а саме: у групі без діабету медіанний рівень глюкози становить приблизно 107 мг/дл, що відповідає верхній межі нормальних значень, тоді як у групі з діабетом медіана досягає 140 мг/дл, що значно перевищує діагностичний поріг для діабету. Міжквартильний розмах у групі без діабету охоплює діапазон від 95 до 125 мг/дл, переважно в межах нормальних та пограничних значень. У групі з діабетом міжквартильний розмах становить приблизно від 125 до 170 мг/дл, що чітко відповідає діабетичному діапазону концентрацій глюкози. Розподіл викидів у групі без діабету представлений поодинокими випадками підвищених значень (до 180...200 мг/дл), що може свідчити про наявність недіагностованого преддіабету або ранніх стадій захворювання. Критично низькі значення (близько 0 мг/дл) в обох групах, швидше за все, відображають технічні помилки вимірювання.

Діагностична значущість рівня глюкози підтверджується практично повною відсутністю перекриття між основними розподілами груп. Завдяки чому рівень глюкози є найнадійнішим одиночним предиктором діабету у цьому наборі даних.

Клінічна інтерпретація результатів відповідає встановленим діагностичним критеріям — рівень глюкози ≥ 126 мг/дл або ≥ 200 мг/дл в довільний час, що є діагностичним для цукрового діабету. Спостережувані медіанні значення точно відображають зазначену діагностичну межу.

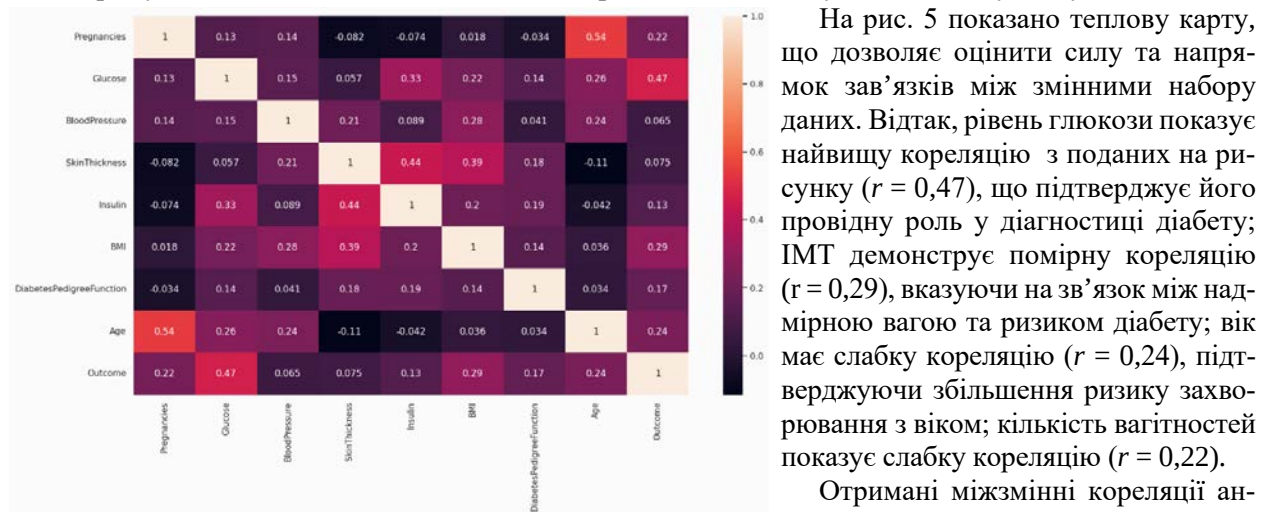


Рис. 5. Теплова карта набору даних

ною шкірної складки, що відображає загальну тенденцію накопичення жирової тканини; вік та кількість вагітностей мають взаємну кореляцію, що є природним для жіночої популяції.

Метаболічні взаємозв'язки підтверджують, інсулін демонструє помірну кореляцію з глюкозою, що відображає компенсаторну відповідь на гіперглікемію, також інсулін корелює з ІМТ, підтверджуючи зв'язок між ожирінням та інсулінорезистентністю.

Кореляції ІМТ, віку та інсуліну з цільовою змінною демонструють багатофакторну природу захворювання.

Помірна сила більшості кореляцій ($r < 0,5$) вказує на відсутність мультиколінеарності між предикторами, що є сприятливим для побудови машинних моделей навчання. Комбінація декількох помірно корелюючих змінних може забезпечити кращу прогностичну здатність, ніж окремі сильні предиктори. Розподіл кореляцій свідчить про доцільність використання всіх змінних у моделюванні, оскільки кожна з них вносить унікальну інформацію про ризик діабету, не дублюючи повністю інформацію інших предикторів.

Побудуємо модель логістичної регресії для класифікації пацієнток щодо наявності або відсутності у них діабету. Оцінювання моделі виконаємо за допомогою обчислення метрик на тренувальному та тестовому наборах даних, визначення яких є важливим для розуміння продуктивності

моделі та подальшої класифікації. Обчислимо п'ять ключових метрик, які допомагають комплексно оцінити якість класифікатора: Accuracy, Precision, Recall, F1-score та ROC-AUC (рис. 6). Для цієї моделі метрики Accuracy та AUC-ROC близькі за значеннями, тому можемо стверджувати, що модель добре узагальнює, а також про відсутнє значне перенавчання. На тренувальному наборі спостерігається помітний дисбаланс між Precision (0,71) та Recall (0,56). Модель є «обережнішою» (високий Precision), пропускає значну частину позитивних випадків (низький Recall). На тестовому наборі Recall та Precision значно вирівнялися. Це означає, що, хоча модель стала трохи частіше помилково позначати негативні випадки як позитивні (зниження Precision), вона водночас набагато краще ідентифікує справжні позитивні випадки (зростання Recall). Для підвищення діагностичної ефективності в роботі впроваджено поняття адаптивного порогу прийняття рішення та зміщено його до рівня 0,25...0,3. Таке коригування дозволяє перерозподілити баланс між помилками першого та другого роду на користь мінімізації останніх. У медичній практиці виявлення цукрового діабету мінімізація хибнонегативних результатів (пропуску хвороби) має пріоритетне значення, навіть за умови незначного зростання кількості хибнопозитивних спрацювань.

Аналіз матриці плутанини (рис. 7) показав, модель логістичної регресії продемонструвала стабільну роботу без ознак перенавчання, показавши майже ідентичну точність на навчальній (0,7687) та тестовій (0,7662) вибірках. Вона стабільна, намагалася краще передбачати наявність діабету на навчальній вибірці (нижча чутливість), ніж на тестовій.

```

=== Метрики на TRAIN ===
Accuracy: 0.7687
Precision: 0.7126
Recall: 0.5587
F1-score: 0.6263
AUC-ROC: 0.8485

=== Метрики на TEST ===
Accuracy: 0.7662
Precision: 0.6792
Recall: 0.6545
F1-score: 0.6667
AUC-ROC: 0.8197

```

Рис. 6. Обчислені метрики для моделі логістичної регресії

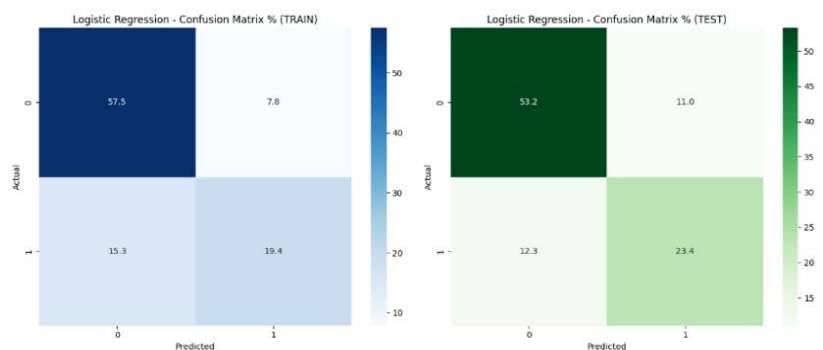


Рис. 7. Матриця плутанини для моделі логістичної регресії

Побудуємо модель на основі алгоритму k -найближчих сусідів ($k = 5$). Оцінимо ефективність побудованої моделі за допомогою вищезазначених метрик (рис. 8). Спостерігається помітний розрив між показниками на навчальній та тестовій вибірках. Особливо велика різниця у Accuracy (0,83 та 0,75) та AUC-ROC (0,90 та 0,79). Такий значний спад продуктивності під час переходу до тестових даних є класичною ознакою перенавчання. Модель k -найближчих сусідів, ймовірно, точно підлаштувалася під шум і специфічні деталі навчального набору, через що втратила здатність ефективно узагальнювати на незалежних даних. Для покращення моделі можна зменшити кількість сусідів, провести нормалізацію або стандартизацію ознак, або використати інший алгоритм класифікації. Втім, на тестовій вибірці модель все ще демонструє задовільну якість, особливо з огляду на Accuracy = 0,75 та AUC-ROC = 0,79.

```

=== K-Nearest Neighbors: TRAIN
Accuracy: 0.8339
Precision: 0.8136
Recall: 0.6761
F1-score: 0.7385
AUC-ROC: 0.9029

=== K-Nearest Neighbors: TEST
Accuracy: 0.7468
Precision: 0.6429
Recall: 0.6545
F1-score: 0.6486
AUC-ROC: 0.7906

```

Рис. 8. Обчислені метрики для моделі k -найближчих сусідів

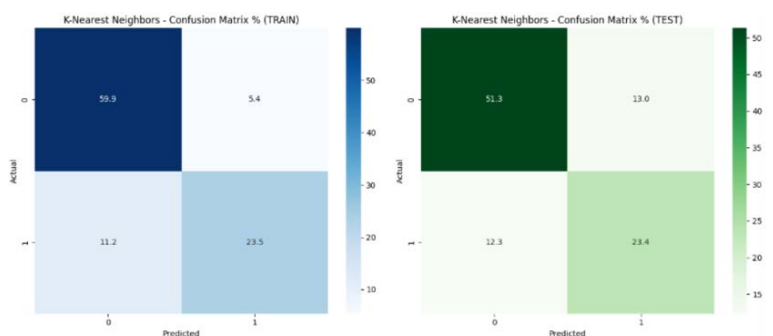


Рис. 9. Матриця плутанини для моделі k -найближчих сусідів

Хоча модель на тестовій вибірці, все ще добре виявляє справжні позитивні випадки (метрика Recall залишилася майже незмінною, оскільки хибнонегативних спрацювань стало лише трохи

більше — з 11,2 % до 12,3 %), її загальна надійність прогнозів значно погіршилася (значно гірший Precision). Це сталося через збільшення кількості хибнопозитивних помилок. Таким чином, модель дуже добре знаходить позитивні випадки (висока чутливість), але її прогнози не дуже достовірні. Щоб зробити модель надійнішою та покращити Precision, потрібно зменшити кількість хибнопозитивних випадків, наприклад, збільшивши число сусідів (k) під час налаштування гіперпараметрів.

Побудуємо модель за допомогою алгоритму випадкового лісу. Отримані оцінки ефективності побудованої моделі дають змогу провести аналіз її продуктивності та здатності до узагальнення на нових даних. Показники на навчальній вибірці найкращі та свідчать про те, що модель вивчила всі закономірності та особливості навчального набору даних, не припустившись жодної помилки класифікації, що свідчить про наявність перенавчання (рис. 10).

На тестових даних, які модель раніше не бачила, результати є реалістичнішими, але помітно нижчими. Загальна точність становить 0,7727, тобто модель правильно класифікує близько 77 % пацієнток. Показник AUC-ROC (0,8379) є досить високим — це свідчить про те, що модель має гарну здатність розрізняти хворих і здорових пацієнток у цілому, навіть якщо порогові значення класифікації не є ідеальними. Precision на рівні 0,6724 показує, що серед усіх пацієнток, яким модель встановила наявність діабету, лише 67 % дійсно мають це захворювання (є певна кількість хибнопозитивних результатів). Recall становить 0,7091, що означає здатність моделі виявити приблизно 71 % реальних випадків діабету. F1-score (0,6903) підтверджує помірний баланс між цими двома метриками, що є критично важливим для медичної діагностики. Представимо матрицю плутанини моделі алгоритму випадкового лісу (рис. 11).

```

=== Random Forest: TRAIN
Accuracy: 1.0000
Precision: 1.0000
Recall: 1.0000
F1-score: 1.0000
AUC-ROC: 1.0000

=== Random Forest: TEST
Accuracy: 0.7727
Precision: 0.6724
Recall: 0.7091
F1-score: 0.6903
AUC-ROC: 0.8379

```

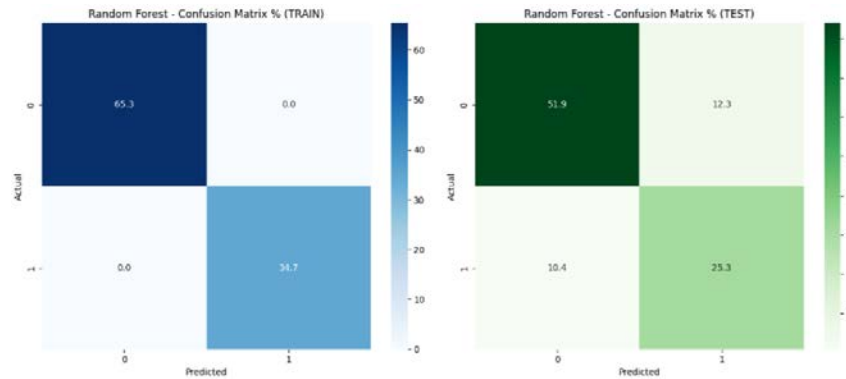


Рис. 10. Обчислені метрики для моделі випадкового лісу

Рис. 11. Матриця плутанини для моделі випадкового лісу

Матриці помилок підтверджують перенавчання моделі: на навчальних даних класифікація є бездоганною (0 % помилок), проте на тестовій вибірці з'являється значна кількість хибних результатів. Зокрема, модель правильно ідентифікувала відсутність діабету у 51,9 % випадків та його наявність у 25,3 %, але водночас припустилася помилок у 22,7 % від загальної кількості тестів. Найкритичнішим для медицини є показник у 10,4 % хибнонегативних результатів (пацієнтки з діабетом, яких модель пропустила), що свідчить про необхідність додаткового налаштування гіперпараметрів для покращення узагальнювальної здатності алгоритму. Отже, головною проблемою цієї моделі класифікації є розрив між ідеальними показниками на навчанні та реальними на тестуванні, що підтверджує перенавчання. Для того, щоб зробити модель надійною для прогнозування на нових даних, необхідно застосувати регуляризацію, наприклад, обмежити глибину дерев або збільшити мінімальну кількість об'єктів у листі.

Висновки

У ході дослідження побудовано моделі алгоритмів машинного навчання (логістичної регресії, k -найближчих сусідів та випадкового лісу) для виявлення наявності цукрового діабету у пацієнток. Отже, на основі проведеного аналізу стабільності та чутливості алгоритмів, можна стверджувати, що для задач первинного скринінгу оптимальною є модель логістичної регресії з адаптованим порогом класифікації.

Наукова новизна отриманих результатів полягає у відмові від використання стандартного порогу прийняття рішення (0,5) на користь діапазону 0,25...0,3. Це забезпечує моделі високу прогностичну здатність у виявленні пацієнтів з цукровим діабетом, зводячи до мінімуму ризик пропуску захворювання. Запропонований підхід до налаштування класифікатора підвищує надійність систем підтримки прийняття медичних рішень, через що вони є адаптованішими до реальних клінічних умов, де

ціна помилки другого роду є високою.

Модель на основі k -найближчих сусідів демонструє ознаки перенавчання. Значний розрив у точності між навчальною (0,83) та тестовою (0,75) вибірками вказує на те, що модель надто точно підлаштувалася під навчальні дані і втратила частину здатності ефективно працювати з новими, невідомими даними. На тестовій вибірці спостерігалось зростання хибнопозитивних випадків, що знижує точність прогнозів.

Модель на основі випадкового лісу, незважаючи на найвищі номінальні показники на тестовій вибірці (точність 0,77 та AUC-ROC 0,84), виявилася сильно перенавченою. Досягнення бездоганного результату (точність 1.0) на навчальних даних та значно нижчого на тестових є класичною ознакою втрати генералізаційної здатності. Критичним недоліком є 10,4 % хибнонегативних результатів на тесті, що є неприйнятним у медичній діагностиці, де пропуск реального захворювання може мати серйозні наслідки.

У рамках дослідження вдосконалено методику вибору класифікаційних моделей. Встановлено, що завдяки високій стабільності та можливості гнучкого регулювання чутливості, оптимальною є модель логістичної регресії з адаптованим порогом класифікації. Запропонований підхід забезпечує не лише математичну стійкість моделі, а й її практичну цінність для систем підтримки прийняття медичних рішень, оскільки дозволяє виявляти пацієнтів з ризиком діабету на ранніх стадіях з вищою надійністю, ніж у разі використання стандартних параметрів класифікації.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

- [1] Л. Ліщинська, і Н. Добровольська, «Перспективні програмні інструменти для аналізу даних у бізнесі», *Вісник Хмельницького національного університету*, № 1 (305), с. 78-83, 2022. <http://ir.lib.vntu.edu.ua/handle/123456789/35278>.
- [2] В. Нестеров, А. Шиш, і Т. Музиченко, «Ефективний економічний розвиток підприємства через інтелектуальний аналіз даних: використання AI для прогнозування та оптимізації стратегій бізнесу», *Економіка та суспільство*, № 59, 2024. <https://doi.org/10.32782/2524-0072/2024-59-87>.
- [3] М. Ціж, «Огляд застосування сучасних методів топологічного аналізу даних», *TSynergy*, № 2, с. 112-122, 2024, <https://doi.org/10.53920/TTS-2024-2-7>.
- [4] О. Дацюк, «Етапи обробки медичних даних для розв'язання задач алгоритмами машинного навчання», *Вісник Національного технічного університету*, т. 1. № 2(14), с. 144-156, 2025. <https://doi.org/10.20998/2411-0558.2025.02.10>.
- [5] Д. Панаскін, і Є. Білоконь, «Машинне навчання в діагностиці захворювань легеневої системи», *Технічні науки та технології*, № 2 (28), с. 76-87, 2022. [https://doi.org/10.25140/2411-5363-2022-2\(28\)-76-87](https://doi.org/10.25140/2411-5363-2022-2(28)-76-87).
- [6] Н. Бугаєць, і І. Лисенко, «Задача прогнозування в машинному навчанні», *Автоматизація технологічних та бізнес-процесів*, т. 17, вип. 2, 2025. <https://doi.org/10.15673/atbp.v17i2.3150>.
- [7] М. Лисенко, і О. Пронькін, «Застосування технологій машинного навчання для встановлення медичного діагнозу», *Зв'язок*, № 5 (171), с. 75-82, 2025. <https://doi.org/10.31673/2412-9070.2024.051397>.
- [8] І. Калініна, і О. Гожий, «Дослідження ефективності методів класифікації при прогнозуванні в задачах машинного навчання», *Управління розвитком складних систем*, вип. 46, с. 173-180, 2021. <https://doi.org/10.32347/2412-9933.2021.46.173-180>.
- [9] В. Соломатін, і Л. Машенко, «Застосування математичних моделей при обробці медичних даних діагностування ішемічної хвороби серця», *Актуальні проблеми автоматизації та інформаційних технологій*, т. 27, с. 178-187, 2023. <http://dx.doi.org/10.15421/432317>.
- [10] Н. Бойко, і В. Курило, «Алгоритм класифікації медичних даних для прогнозування онкології», *Системи та технології*, т. 66, № 2, с. 21-31, 2023. <https://doi.org/10.32782/2521-6643-2023.2-66.3>.
- [11] В. Л. Левківський, «Моделі та методи обробки даних системи віддаленого моніторингу стану пацієнтів з цукровим діабетом.» дис. д-ра філос. (PhD): 121 – Програмна інженерія. Державний університет «Житомирська політехніка», Житомир, 2023. <https://eztuir.ztu.edu.ua/handle/123456789/8375>.

Рекомендована кафедрою системного аналізу та інформаційних технологій ВНТУ

Дата надходження 25.11.2025

Дата прийняття після доопрацювання 16.12.2025

Дата прийняття до друку після рецензування 9.01.2026

Дата публікації 2.10.2026

Коротун Ольга Володимирівна — канд. пед. наук, доцент, доцент кафедри комп'ютерних наук, e-mail: korotun-o@ztu.edu.ua . <https://orcid.org/0000-0003-2240-7891>;

Свінцицька Олександра Миколаївна — канд. екон. наук, доцент, доцент кафедри комп'ютерних наук, e-mail: up_som@ztu.edu.ua . <https://orcid.org/0000-0002-2613-2437>;

Марчук Дмитро Костянтинович — старший викладач кафедри комп'ютерних наук, e-mail: klpz_mdsk@ztu.edu.ua . <https://orcid.org/0000-0001-8675-8047>;

Штиль Михайло Романович — здобувач кафедри комп'ютерних наук, e-mail: ist221_shmr@student.ztu.edu.ua . <https://orcid.org/0009-0005-8770-7985>

Державний університет «Житомирська політехніка», Житомир

O. V. Korotun¹
 O. M. Svintsytska¹
 D. K. Marchuk¹
 M. R. Shtyl¹

Use of Machine Learning Algorithms to Determine the Presence of Diabetes Mellitus

¹Zhytomyr Polytechnic State University

The article is devoted to the research and practical application of modern machine learning methods for solving the urgent problem of detecting diabetes mellitus in patients. At the initial stage, the work focuses on preliminary data processing: detecting empty and abnormal values, studying data set variables, etc. Descriptive statistics are used to summarize, organize, and present the main characteristics of the dataset. Key indicators are obtained for numerical variables: mean, median, standard deviation, minimum and maximum values, as well as quartiles. This made it possible to assess the central tendency of the data, their dispersion, range, and identify irregularities in the distribution. For categorical variables, the number of unique values, the most common category, and its frequency were determined. The use of machine learning algorithms to diagnose diabetes in patients was described. The task was to predict the presence (result = 1) or absence (result = 0) of diabetes in patients based on a number of diagnostic indicators, including glucose level, blood pressure, skinfold thickness, insulin level, body mass index, age, etc. The study analyzed input data to identify patterns and correlations between diagnostic parameters, analyze the distribution of characteristics, and identify potential emissions. Several classification algorithms were considered and applied to build models, including logistic regression, the k-nearest neighbors method, and random forest. The final stage of the study was to compare the efficiency of different algorithms. The best results were demonstrated by the logistic regression model with an adapted classification threshold. The paper justifies shifting the decision threshold from the standard value of 0.5 to the range of 0.25...0.3, which significantly minimizes type II errors (false negatives). This is important for the early diagnosis of diabetes, as it allows for the detection of a larger number of patients who might have been missed with the standard model settings.

Keywords: data analysis, machine learning, forecasting, logistic regression, k-nearest neighbors, random forest.

Korotun Olga V. — Cand. Sc. (Pedagog.), Associate Professor, Associate Professor of the Chair of Computer Science, e-mail: korotun-o@ztu.edu.ua . <https://orcid.org/0000-0003-2240-7891>;

Svintsytska Oleksandra M. — Cand. Sc. (Econom.), Associate Professor, Associate Professor of the Chair of Computer Science, e-mail: up_som@ztu.edu.ua . <https://orcid.org/0000-0002-2613-2437>;

Marchuk Dmytro K. — Senior Lecturer of the Chair of Computer Science, e-mail: kipz_mdk@ztu.edu.ua . <https://orcid.org/0000-0001-8675-8047>;

Shtyl Mykhailo R. — Researcher of the Chair of Computer Science, e-mail: ist221_shmr@student.ztu.edu.ua . <https://orcid.org/0009-0005-8770-7985>