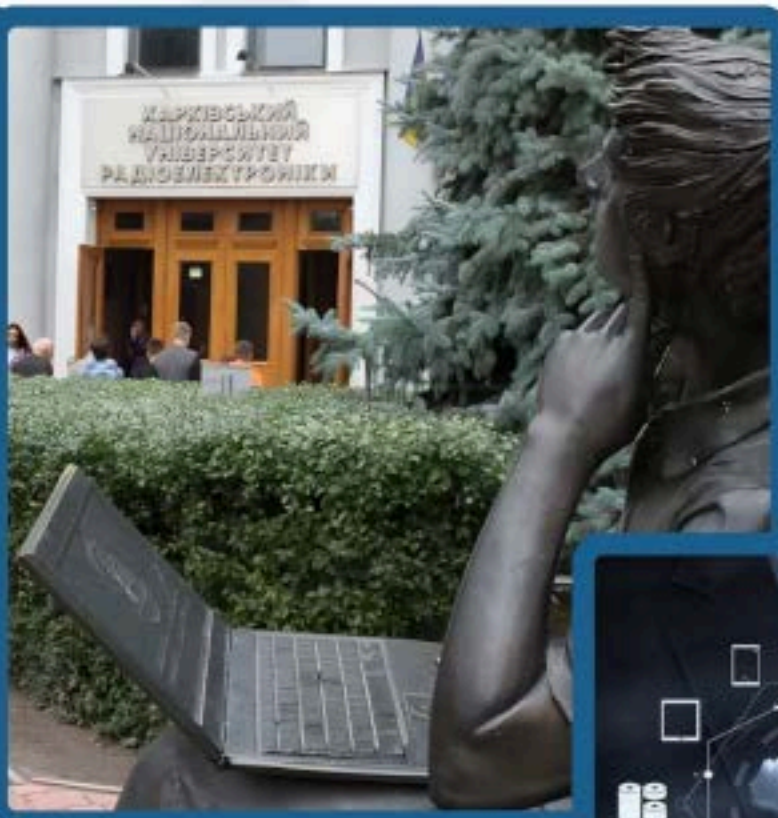
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
РАДІОЕЛЕКТРОНІКИ

«РАДІОЕЛЕКТРОНІКА ТА МОЛОДЬ У XXI СТОЛІТТІ»

Матеріали XXIX Міжнародного молодіжного форуму



2025

ТОМ 6 : «Інформаційні інтелектуальні системи»

Харків, 2025

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
РАДІОЕЛЕКТРОНІКИ

МАТЕРІАЛИ XXVIII МІЖНАРОДНОГО МОЛОДІЖНОГО ФОРУМУ

«РАДІОЕЛЕКТРОНІКА ТА МОЛОДЬ У XXI СТОЛІТТІ»

16 – 18 квітня 2025 р.

Том 6

КОНФЕРЕНЦІЯ «ІНФОРМАЦІЙНІ ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ» INFORMATION INTELLIGENT SYSTEMS

Харків 2025

УДК 004.85:616.379-008.64

ПІДВИЩЕННЯ ТОЧНОСТІ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ ПРИ ЛІКУВАННІ ЦУКРОВОГО ДІАБЕТУ НА ОСНОВІ ЗБАГАЧЕННЯ ТЕСТОВИХ ДАНИХ

Мінухін С.В., Семенець О.М.
e-mail: serhii.minukhin@hneu.net, semenets.oleksandr@hneu.net
Харківський національний економічний університет ім. С. Кузнеця
каф. інформаційних систем
м. Харків, Україна

Type 1 diabetes is a chronic condition requiring precise insulin dosing. Effective management depends on accurate predictions of insulin needs, which can be improved using machine learning techniques. This study leverages data from the HUPA-UCM Diabetes Dataset, containing detailed patient information obtained through continuous monitoring. The data undergoes pre-processing and integration to improve quality, followed by the application of machine learning models such as Linear Regression, Decision Tree, Random Forest, and K-Nearest Neighbors to predict insulin requirements.

Цукровий діабет – це хронічне захворювання, яке потребує постійного контролю рівня глюкози в крові. Важливою частиною терапії є використання базального інсуліну, який підтримує стабільний рівень глюкози протягом доби, незалежно від прийомів їжі. Завдяки методам машинного навчання з’являється можливість прогнозувати його оптимальну дозу, враховуючи індивідуальні особливості пацієнта, фізіологічні показники та дані безперервного моніторингу. В даному дослідженні пропонується підхід до підвищення ефективності прогнозування з застосуванням методів машинного навчання для точного контролю глікемії та мінімізації ризиків гіпо- і гіперглікемії на основі збагачення тестових даних. У роботі застосовано такі моделі машинного навчання [3]: Linear Regression, Decision Tree, Random Forest та K-Nearest Neighbors для прогнозування дози інсуліну на основі даних пацієнтів.

Алгоритм збагачення (Data Enrichment) тестових даних.

Крок 1. Здійснено інтеграцію 25 окремих датасетів у форматі CSV, що містять часові ряди спостережень за пацієнтами [1]. Для цього всі файли були зчитані з вказаної директорії й об’єднані в один датафрейм [2].

Крок 2. Здійснено обробка значень часу: часові дані було приведено до формату «datetime» за допомогою методу to_datetime() [2], що дало змогу виділити додаткові ознаки: хвилини, години, місяць і день.

Крок 3. Здійснено збагачення тестових даних. Подібний підхід до збагачення даних ефективно застосовується у різних галузях [5]: додавання супутніх ознак (наприклад, помилок прогнозу погоди) дозволяє підвищити точність моделей.

Для покращення точності прогнозування збільшено об’єм тестових даних, що отримані на основі доступних даних [1]: стать, рівень HbA1c,

вік, час із моменту діагнозу, вагу, зріст і тип лікування шляхом інтеграції клінічної інформації з часовими рядами через спільний ідентифікатор пацієнта «person_id». Це забезпечило формування повного та узгодженого набору даних для їх подальшого аналізу. У результаті об'єднання об'єм даних цього датафрейму збільшився на 30,6%.

Крок 4. Проведено перевірку аномалій в даних (табл. 1). Для кожної змінної встановлені обмеження в межах її припустимих значень:

$$variable \in [min\_value, max\_value],$$

де: variable – значення відповідної змінної; min_value – мінімальне припустиме значення змінної; max_value – максимальне припустиме значення змінної.

Таблиця 1 – Опис перевірки на аномалії

Зміна в наборі даних	Межа, діапазон
glucose (глюкоза)	$glucose \in [0,500]$
heart_rate (серцевий ритм)	$heart\_rate \in [40,200]$
minute (хвилина)	$minute \in [0,59]$
hour_of_day (година доби)	$hour\_of\_day \in [0,23]$
month (місяць)	$month \in [1,12]$
day (день)	$day \in [1, days\_in\_month]$ $days\_in\_month$ залежить від місяця
age (вік)	$age \in [0,120]$
dx_time (час діагностики)	$dx\_time \in [0, current\_time]$ Час діагностики не повинен бути в майбутньому, тому його значення має бути в межах від 0 до поточного часу
weight (вага)	$height \in [1,300]$
height (зріст)	$heart\_rate \in [50,250]$

Пропуски в даних можуть суттєво вплинути на якість моделей та точність прогнозів, тому їх було оброблено за допомогою методу fillna() з бібліотеки Pandas [2] (табл. 2).

Таблиця 2 – Методи заповнення колонок

Назва колонки	Метод заповнення
glucose	Заповнення пропусків здійснюється за допомогою функції fillna(), в яку передається середнє значення відповідного стовпця, обчислене методом mean().
calories	
heart_rate	
Інші колонки	Заповнення за допомогою функції fillna(0), де пропуски даних замінюються на значення 0.

Для колонок glucose, calories та heart_rate було обрано метод заповнення пропусків середнім значенням, а інші колонки заповнені значенням 0.

Крок 5. Кодування категоріальних змінних із клінічних даних пацієнтів. Для цього було застосовано метод One-Hot Encoding за допомогою функції get_dummies з бібліотеки Pandas [2]. Цей метод перетворює категоріальні змінні на набір бінарних стовпців, де кожна категорія стає окремим стовпцем із значеннями 0 або 1, що вказують на наявність або відсутність відповідної категорії в кожному спостереженні.

У результаті було створено 4 нові стовпці: gender_female, gender_male, treatment_CSII і treatment_MDI. Це дозволило коректно врахувати категоріальні ознаки без спотворення їхнього значення.

Підсумковий набір даних було збережено в змінній data, з якої сформовано матрицю ознак та цільову змінну. До ознак увійшли як клінічні характеристики, так і часові параметри пацієнтів, включаючи створені в процесі One-Hot Encoding стовпці для змінних gender і treatment. Після проведення процедури збільшення даних за описаним алгоритмом було використано розділення набору даних на навчальний та тестовий [4]. Результати показали, що якість прогнозування з використанням запропонованого підходу для різних методів машинного навчання збільшилася на 8%-14%, а кращим з них виявився метод Random Forest [3].

Для подальшого розвитку пропонується збільшити об’єм тестових даних та застосувати зворотний зв'язок від пацієнтів для адаптації моделей в реальному часі, що дозволить покращити точність прогнозів та забезпечить персоналізований підхід до лікування.

Список використаних джерел:

1. Hidalgo J. I., Alvarado J., Botella M., Aramendi A., Velasco J. M., Garnica, O. HUPA-UCM Diabetes Dataset // Data in Brief. 2024. Vol. 55. DOI: 10.1016/j.dib.2024.110559.

2. Pandas. URL: https://pandas.pydata.org/docs/ (дата звернення: 01.03.2025).

3. Scikit-learn. URL: https://scikit-learn.org/stable/index.html (дата звернення: 01.03.2025).

4. Smelyakov K., Hurova Y., Osiievskyi S. Analysis of the Effectiveness of Using Machine Learning Algorithms to Make Hiring Decisions // CEUR Workshop Proceedings. 2023. Vol. 3387. P. 77–92. URL: https://ceur-ws.org/Vol-3387/paper7.pdf (дата звернення: 04.03.2025).

5. Zhou Y., Ma L., Ni W., Yu C. Data Enrichment as a Method of Data Preprocessing to Enhance Short-Term Wind Power Forecasting // Energies. 2023. Vol. 16, Issue 5. DOI: 10.3390/en16052094.