DOI 10.34185/1562-9945-1-156-2025-05
УДК 004.89

Т.В. Хом’як, К.В. Сидоренко, А.В. Малієнко, О.С. Мінєєв

ПРОГНОЗУВАННЯ ПРИЧИН ВИЯВЛЕННЯ ЦУКРОВОГО ДІАБЕТУ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Анотація. Цукровий діабет - одне з найпоширеніших хронічних захворювань у світі, яким страждає близько 530 мільйонів людей. Основна причина появи включає генетичну схильність, ожиріння, неправильну харчову поведінку, інсулінорезистентність та шкідливі звички. Вчасне виявлення захворювання може запобігти його розвитку. Багато симптомів цукрового діабету, таких як сухість у роті, часті сечовипускання, погіршення зору, втрата ваги, постійне відчуття голоду, не завжди відразу розглядаються як ознаки захворювання. Але ці симптоми можуть бути ранніми показниками високого рівня глюкози у крові. В роботі проведено аналіз факторів та причин, які впливають на ризик розвитку цукрового діабету та зроблено прогнозування методом машинного навчання Decision Tree, Random Forest, k-NN та Ada Boost. Поведено аналіз отриманих результатів, оцінено точність використаних методів. Отримані результати нададуть змогу виявляти значно більшу кількість випадків діабету до його появи, раннє та ефективне лікування, зменшення витрат на медичне обслуговування.
Ключові слова: цукровий діабет, прогнозування, машинне навчання, Decision Tree, Random Forest, Ada Boost, k-NN.

Постановка проблеми. Одним з головних факторів, які можуть сприяти виникнення цукрового діабету є відсутність своєчасного виявлення захворювання у людини. Зазвичай, люди, які мають такі симптоми, як сухість у роті, часті сечовипускання, погіршення зору, постійну втому, підвищення чи втрату ваги, постійне відчуття голоду, не звертаються до лікаря. Оскільки вони не знають, що дані симптоми можуть вказувати на високий рівень глюкози у крові. Профілактика та раннє виявлення діабету є критично важливими для глобального розвитку здоров'я.

Метою роботи є проведення аналізу факторів та причин, які впливають на появу цукрового діабету, а також прогноз виявлення захворювання до його появи аби запобігти його розвитку в організмі людини, використовуючи показники та звички людини. Оскільки, діабет може призвести до серйозних ускладнень, включаючи пошкодження нирок та серцево-судинних захворювань, що загрожує життю. У разі несвоєчасного виявлення захворювання можливі важкі стани, такі як ампутація кінцівок та втрата зору.

Аналіз останніх досліджень і публікацій. На сьогоднішній день існує велика кількість публікацій, пов’язана з дослідженням хвороби. Так, в роботі [1] наведено кла-

сифікацію діабету, описано причини та наслідки хвороби. Також є робіти в сфері прогнозування виникнення діабету за допомогою методів машинного та глибокого навчання [2-4].

В роботі [4] зазначено, що до 2034 року кількість пацієнтів із захворюванням на цукровий діабет перевищить 592 мільйони. Для передбачення і діагностики діабету запропоновано комп'ютеризовану систему, що заснована на поєднанні алгоритмів Vector Machine, Decision Tree, Naive Bayes і ANN. Але для навчання взята невелика вибірка в розмірі 768 екзепмлярів.

В дослідженні [5] виявлено проблеми з класифікацією і запропоновано зменшити дані для досягнення вищої та ефективнішої точності. Для прогнозування були використані алгоритм побудови дерева рішень та модифікования Fuzzy SLIQ.

Таким чином, кожен алгоритм має власний набір переваг і недоліків в залежності від початкових даних. Тож використаємо прогнозування методами машинного навчання Decision Tree, Random Forest, k-NN, Ada Boost і порівняємо їх точність для задачі прогнозування виявлення захворювання до його появи

Викладення основного матеріалу дослідження. Початкові дані для аналізу та прогнозування виявлення цукрового діабету взято з сайту «Центр контролю та профілактики захворювань», які подано у форматі csv-файлу із 22 колонок (показники та параметри людей) та 253689 рядків (значення показників та параметрів) для кожної людини [6] (табл. 1).

Таблиця 1

Структура початкових даних

Назва колонки	Тип (значення) колонки	Опис колонки
1	2	3
Diabetes_012	0 – no diabetes 1 – prediabetes 2 – diabetes	0 – пацієнт не має діабету, 1 – пацієнт схильний до діабету (переддіабет), 2 – пацієнт має діабет
HighBP	0 – no high BP 1 – high BP	0 – не має високий тиск, 1 – має високий тиск.
HighChol	0 – no high cholesterol 1 – high cholesterol	0 – не має високий холестерин, 1 – має високий холестерин.
CholCheck	0 – no cholesterol check in 5 years 1 – yes cholesterol check in 5 yearsCholCheck	0 – не було перевірки холестерину через 5 років, 1 – була перевірка холестерину через 5 років.
BMI	Число	Індекс маси тіла
HeartDiseaseorAttack	0 – no	Чи була ішемічна хвороба серця або ін-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

	1 – yes	фаркт: 0 – ні, 1 – так.
Smoker	0 – no 1 – yes	Інформація про пацієнта, чи курив як мінімум 100 цигарок за своє життя: 0 – ні, 1 – так.
Stroke	0 – no 1 – yes	Інформація про пацієнта, чи був колись інсульт: 0 – ні, 1 – так.
PhysActivity	0 – no 1 – yes	Інформація про фізичну активність пацієнта протягом останні 30 днів, окрім роботи: 0 – ні, 1 – так.
Fruits	0 – no 1 – yes	Інформація про споживання фруктів 1 або більше разів у день пацієнта: 0 – ні, 1 – так.
Veggies	0 – no 1 – yes	Інформація про споживання овочів 1 або більше разів у день пацієнта: 0 – ні, 1 – так.
HvyAlcoholConsump	0 – no 1 – yes	Інформація про споживання алкоголю (для чоловіків більше 14 напоїв на тиждень, для жінок – 7): 0 – ні, 1 – так.
AnyHealthcare	0 – no 1 – yes	Інформація про наявність медичного страхування у пацієнта: 0 – ні, 1 – так.
NoDocbcCost	0 – no 1 – yes	Інформація про похід пацієнта до будь-якого лікаря протягом останніх 12 місяців: 0 – ні, 1 – так.
GenHlth	1 – excellent, 2 – very good, 3 – good, 4 – fair, 5 – poor.	Інформація про здоров’я пацієнта на його думку за шкалою: 1 – відмінна, 2 – дуже хороша, 3 – нормальна, 4 – задовільна, 5 – погана.
MenthHlth	Число (1-30)	Інформація про психічне здоров’я, де пацієнт мав стрес, депресію та проблеми з емоціями протягом останніх 30 днів.
PhysHlth	Число (1-30)	Інформація про фізичне здоров’я, де пацієнт мав фізичні захворювання та травми протягом останніх 30 днів (скільки днів пацієнт мав такий стан)
DiffWalk	0 – no 1 – yes	Інформація про пацієнта чи є серйозні труднощі при ходьбі чи підйомі по сходах: 0 – ні, 1 – так.
Sex	0 – female,	Стать пацієнта:

	1 - male	0 – жінка, 1 – чоловік.
Age	Число	Вік, де поділено на 13 категорій: 1 – 18-24, 13 – 80 і більше років.
Education	1 – never attended school 2 – elementary grades 3 – some high school graduate 4 – high school graduate 5 – college year to 3 years 6 – college 4 years or more	Інформація про освіту пацієнта: 1 – ніколи не відвідував школу, 2 – 1-8 класи, 3 – 9-11 класи, 4 – 12 класів, 5 – коледж до 3 років, 6 – закінчив коледж.
Income	Число (1-8)	Інформація про заробітну плату, яка розподілена на 8 груп, де 1 – менше ніж 10000$, 8 – більше, ніж 75000$.

Для аналізу та прогнозування методами машинного навчання дані було завантажено в Google Colab для подальшої обробки мовою Python. На рис. 1 наведено початковий дата сет.

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0	18.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	3.0	0.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	5.0	30.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	0.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	3.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
253675	0.0	1.0	1.0	1.0	45.0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0	3.0	0.0
253676	2.0	1.0	1.0	1.0	18.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	4.0	0.0
253677	0.0	0.0	0.0	1.0	28.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	1.0	0.0
253678	0.0	1.0	0.0	1.0	23.0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0	3.0	0.0
253679	2.0	1.0	1.0	1.0	25.0	0.0	0.0	1.0	1.0	1.0	...	1.0	0.0	2.0	0.0

253680 rows × 22 columns

I	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	3.0	0.0	0.0	0.0	0.0	7.0	6.0	1.0
0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	5.0	30.0	30.0	1.0	0.0	9.0	4.0	8.0
0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	0.0	0.0	0.0	0.0	11.0	3.0	6.0
0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	3.0	0.0	0.0	0.0	11.0	5.0	4.0
..	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0	3.0	0.0	5.0	0.0	1.0	5.0	6.0	7.0
0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	4.0	0.0	0.0	1.0	0.0	11.0	2.0	4.0
0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0	5.0	2.0
0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0	3.0	0.0	0.0	0.0	1.0	7.0	5.0	1.0
0	0.0	0.0	1.0	1.0	1.0	...	1.0	0.0	2.0	0.0	0.0	0.0	0.0	9.0	6.0	2.0

Рисунок 1 – Початкові дані в Google Colab
(перші та останні 5 рідків початкових даних)

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Набір даних є незбалансованим, значення «0» (не має діабету) зустрічається 213703 рази, «1» (перед діабет) − 35346 раз та «2» (має діабет) − 4631 раз. Цей фактор враховано в процесі попереднього опрацювання даних. Наведемо основні кроки попереднього опрацювання даних:

– збір та завантаження даних;

– видалення даних, які мають дублікати, кількість знайдених дублікатів дорівнює 15548, даних зі значенням NaN або NULL не було знайдено;

– аналіз та візуалізація даних для отримання розподілу та взаємозв'язків між ознаками;

– приведення числових ознак до одного і того ж масштабу або стандартної одиниці вимірювання;

– видалення викидів (outliers) за допомогою методу Isolation Forest, де було знайдено та видалено 114071 аномальних значень, тобто після видалення аномалій дата сет має 114071 рядків значень показників пацієнтів;

– відбір ознак (Feature Selection): для даного набору даних при першому аналізі кожної колонки даних було виявлено, що колонки Education, Income, AnyHealthcare, є неінформативними для виявлення цукрового діабету у людини;

– розділення даних (Data Splitting): для даного набору даних було обрано, що 70% даних – навчальна вибірка та 30% – тестова;

– балансування класів (Class Balancing): для даного набору даних було обрано метод SMOTEENN, після застосування якого дані були збалансовані, де класи мали наступні значення «0» – 27948 значення, «1» – 39597, «2» – 35628 значень.

– масштабування атрибутів методом Standard Scaler.

Після опрацювання початкових даних було побудовано діаграму співвідношення трьох можливих стадій діабету (рис. 3). Виявлено, що 82,7% - не мають діабет, 15,3 % - мають діабет та 2% - перед діабет.

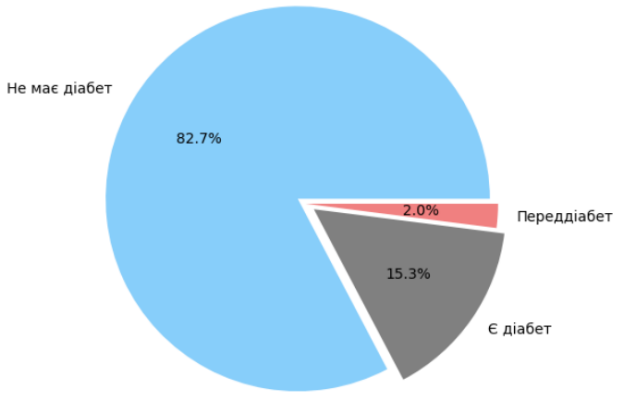

Рисунок 3 – Діаграма співвідношення можливих стадій діабету

Для аналізу ступеня впливу на виявлення діабету побудовано матрицю кореляції (рис. 4). Кореляцію близько 0,2 мають показники холестерин, індекс маси тіла, ішемічна хвороба серця, фізичне здоров’я, труднощі з ходьбою та вік, тобто важливі фактори на вплив появи та розвитку діабету в організмі людини. Менший вплив мають наступні

фактори, як куріння, інсульт, наявність медстрахування, відвідування лікаря, стать та психічне здоров'я (кореляція у проміжку від 0 до 0,1). Від'ємний коефіцієнт кореляції спостерігається у наступних факторів: фізична активність, вживання фруктів та овочів, алкогольна залежність, освіта та дохід.

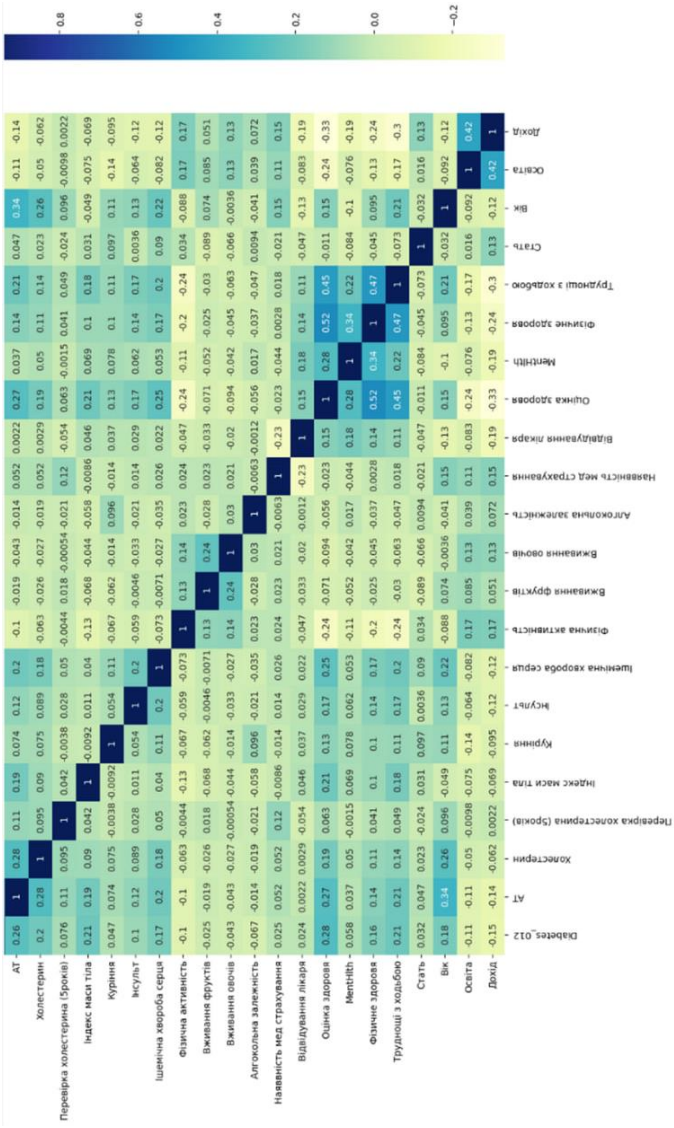

Рисунок 4 – Матриця кореляції факторів, що впливають на виявлення діабету

На основі отриманої діаграми розподілу ознаки статі в залежності від стадій діабету (рис. 4) можна зробити висновки:

- мають діабет майже однакова кількість жінок та чоловіків близько 8%;
- стан перед діабету мають близько 1%, але жінки трохи більше;
- жінок, які не мають діабета на 20000 більше, ніж чоловіків.

Таким чином, стан діабету або перед діабету мають трохи більше жінки ніж чоловіки, оскільки жінки можуть мати гормональні зміни під час життя, які можуть впливати на ризик діабету.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

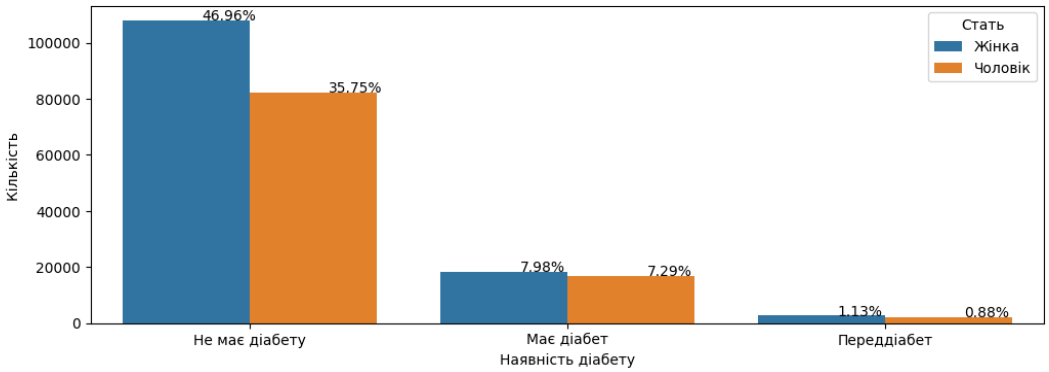

Рисунок 5 – Діаграма розподілу стадій діабету за статтю

Після проведення аналізу кожного показнику можна зробили наступні висновки:

– пацієнти, які мають ішемічну хворобу серця (ІХС), хворіють приблизно у 4 рази більше на цукровий діабет;

– із 27% пацієнтів, які палять, 8% мають цукровий діабет або перед діабет;

– жінок, які не мають діабету, на 20000 більше, ніж чоловіків;

– пацієнти, які не вживають фрукти, майже у 2 рази більше, які не хворіють на цукровий діабет;

– пацієнти з високим холестерином хворіють у два рази частіше на цукровий діабет;

– пацієнти у віці від 50 до 64 років складають найбільшу кількість тих, хто має діабет;

– кількість пацієнтів з ожирінням та цукровим діабетом складає 50%;

– частина пацієнтів, які мають інсульт та діабет, складає 1,5%;

– фізично активні пацієнти мають діабет у 3,5 рази менше, ніж інші.

Для прогнозування методом Decision Tree [6] обрано такі гіперпараметри: criterion= 'entropy', max_depth=40, а інші параметри за замовченням. Результати наведено у таблиці 2.

Таблиця 2

Результати моделі Decision Tree Classifier

class	precision	recall	f1-score	accuracy
0 (не має діабет)	0.92	0.89	0.91	0.92
1 (перед діабет)	0.93	0.95	0.94	
2 (має діабет)	0.89	0.89	0.89	

З таблиці 2 можна зробити висновки, що точність моделі для всіх класів дорівнює 0.92, що свідчить про її загальну ефективність у класифікації. Модель має високі показники precision та recall для класів "Не має діабету" і "Пред діабет", що робить її корисною для виявлення цих станів у пацієнтів. Однак для класу "Має діабет" її ефективність менша.

Результати тестування моделі Random Forest Classifier з наступними гіперпараметрами: n_estimators=100, max_features=16, max_depth=16 (інші параметри за замовченням) наведено в таблиці 3.

Таблиця 3

Результати моделі Random Forest Classifier

class	precision	recall	f1-score	accuracy
0 (не має діабет)	0.93	0.89	0.91	0.95
1 (перед діабет)	0.86	0.86	0.86	
2 (має діабет)	0.79	0.82	0.81	

Загальна точність цієї моделі для всіх класів становить 0.95, що свідчить про її загальну ефективність у класифікації. Модель має найкращі показники для класу "Не має діабету", що робить її корисною для виявлення цього стану, проте для класів "Пред діабет" і "Має діабет" точність менша.

Результати тестування моделі Ada Boost Classifier з гіперпараметрами n_estimators=100, max_features=16, max_depth=16 (інші параметри за замовченням).наведені в таблиці 4.

Таблиця 4

Результати моделі Ada Boost Classifier

class	precision	recall	f1-score	accuracy
0 (не має діабет)	0.84	0.77	0.80	0.63
1 (перед діабет)	0.59	0.55	0.57	
2 (має діабет)	0.54	0.62	0.58	

Загальна точність даної моделі для всіх класів становить 0.63, що свідчить про її загальну ефективність у класифікації. Проте точність та повнота для класів "Пред діабет" і "Має діабет" є недостатньо високими, вказуючи на помилкові класифікації для цих груп.

Результати тестування моделі K-NN Classifier з гіперпараметром - кількість сусідів = 5 (інші за замовченням) наведені в таблиці 5.

Таблиця 5

Результати моделі K-NN Classifier

class	precision	recall	f1-score	accuracy
0 (не має діабет)	0.91	0.98	0.94	0.9
1 (перед діабет)	0.00	0.00	0.00	
2 (має діабет)	0.31	0.10	0.15	

З таблиці 5 можна зробити висновок, що модель здійснила не збалансовану класифікацію даних, оскільки показники precision, recall, f1-score доволі різні для трьох класів, а це вказує на те, що модель не навчилася розрізняти ці три класи. Тому потрібно було зробити балансування даних за допомогою методу SMOTEENN, після цього

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

дані стали збалансовані, де кожен клас мав таку кількість значень: «0» – 27948 значення, «1» – 39597, «2» – 35628. Після цього ще раз проведено тренування та тестування моделі, результати наведено у таблиці 6.

Таблиця 6

Результати моделі K-NN Classifier

class	precision	recall	f1-score	accuracy
0 (не має діабет)	0.99	0.87	0.93	0.96
1 (перед діабет)	0.96	1.00	0.98	
2 (має діабет)	0.94	0.98	0.96	

Узагальнюючи отримані результати, можна сказати, що модель може бути корисною для точної класифікації пацієнтів з діабетом і перед діабетом, але може не виявити всіх пацієнтів без діабету. Точність моделі в цілому є високою.

Висновки. Після проведення прогнозування виявлення цукрового діабету в організмі людини, використовуючи показники та звички людей чотирма методами машинного навчання, можна зробити висновок, що найбільшу точність 95% має метод Random Forest, а найменшу 63% – Ada Boost, проте не треба одразу поспішати та обирати найкращою моделлю Random Forest Classifier. Оскільки, якщо порівняти показники precision, recall та f1-score, можна побачити, що модель Decision Tree Classifier має найвищі показники f1-score для всіх трьох класів (0.94, 0.93, 0.89). Це вказує на кращу здатність моделі розрізняти всі три класи (не має діабету, перед діабет, має діабет) порівняно з іншими результати моделей. Тому найкращим методом для розв’язання поставленої задачі є Decision Tree.

Отримані результати дослідження мають велике значення, оскільки вони можуть бути використані для покращення роботи медичних фахівців у виявленні цукрового діабету до його появи аби запобігти його розвитку в організмі людини та підвищити шанси порятунку життя та здоров'я пацієнтів, які стикаються з цією серйозною та невиліковною хворобою. Результати роботи також можуть бути корисними для розробки та моделювання інформаційних систем з прогнозу виявлення цукрового діабету.

ЛІТЕРАТУРА

1. American Diabetes Association. Diagnosis and classification of diabetes mellitus // Diabetes Care. *Clinical Practice Recommendations*. 2012. Vol. 35 (Supplement_1). P. 64-71. URL: https://doi.org/10.2337/dc12-s064.

2. Larabi-Marie-Sainte S., Aburahmah L., Almohaini R., Saba T. Current techniques for diabetes prediction: review and case study // Appl. Sci. 2019. Vol. 9, no. 4604. URL: https://doi.org/10.3390/app9214604.

3. Zia U.A., Khan N. Predicting diabetes in medical datasets using machine learning techniques // Int. J. Sci. Eng. Res. 2017. Vol. 8(5). P. 257–267.

4. Neha D., Chande P., Neha M. A Study of Machine Learning Techniques for Diabetes Prediction // *International Journal of Emerging Trends.in Engineering Research.* 2022. Vol. 10, no. 2. P. 133-140. URL: https://doi.org/10.30534/ijeter/2022/181022022.

5. Kamadi, V., Allam A., Thummala, S. A computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach // *Applied Soft Computing*, 2016. Vol.49, P.137-145. URL: https://doi.org/10.1016/j.asoc.2016.05.010.
6. Сидоренко Є.В., Хом'як Т.В. Аналіз причин та прогнозування виявлення цукрового діабету методом машинного навчання Decision Tree // The 6th International scientific and practical conference"Methodical and practical methods of creating inventions" (October 24 – 27, 2023), Sofia, Bulgaria. International Science Group. 2023. с. 266-271. (URL: https://doi.org/10.46299/ISG.2023.2.6).

REFERENCES

1. American Diabetes Association. Diagnosis and classification of diabetes mellitus // Diabetes Care. *Clinical Practice Recommendations*. 2012. Vol. 35 (Supplement_1). P. 64-71. URL: https://doi.org/10.2337/dc12-s064.
2. Larabi-Marie-Sainte S., Aburahmah L., Almohaini R., Saba T. Current techniques for diabetes prediction: review and case study // Appl. Sci. 2019. Vol. 9, no. 4604. URL: https://doi.org/10.3390/app9214604.
3. Zia U.A., Khan N. Predicting diabetes in medical datasets using machine learning techniques // Int. J. Sci. Eng. Res. 2017. Vol. 8(5). P. 257–267.
4. Neha D., Chande P., Neha M. A Study of Machine Learning Techniques for Diabetes Prediction // *International Journal of Emerging Trends.in Engineering Research.* 2022. Vol. 10, no. 2. P. 133-140. URL: https://doi.org/10.30534/ijeter/2022/181022022.
5. Kamadi, V., Allam A., Thummala, S. A computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach // *Applied Soft Computing*, 2016. Vol.49, P.137-145. URL: https://doi.org/10.1016/j.asoc.2016.05.010.
Sydorenko K.V., Khomiak T.V. Analiz prichyn ta prognozyvannya vyiavlennya tsykrovogo diabety metodom mashinnogo navchannya Decision Tree // The 6th International scientific and practical conference"Methodical and practical methods of creating inventions" (October 24 – 27, 2023), Sofia, Bulgaria. International Science Group. 2023. p. 266-271. (URL: https://doi.org/10.46299/ISG.2023.2.6).

Received 14.01.2025.
Accepted 17.01.2025.

Prediction causes of diabetes detection using machine learning methods

Diabetes is one of the most common chronic diseases in the world, affecting about 530 million people. The main causes of its occurrence include genetic predisposition, obesity, improper eating behavior, insulin resistance and bad habits. Early detection of the disease can prevent its development. Many symptoms of diabetes, such as dry mouth, frequent urination, blurred vision, weight loss, constant hunger, are not always immediately considered as signs of the disease. But these symptoms can be early indicators of high blood glucose levels. The paper analyzes the factors and causes that affect the risk of developing diabetes and makes predictions using the Decision Tree, Random Forest, k-NN and Ada Boost machine learning methods. The results are analyzed and the accuracy of the methods used is assessed. The re-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

sults obtained will allow for the detection of significantly more cases of diabetes before it occurs, early and effective treatment, and reduction of healthcare costs.

Keywords: diabetes, prediction, machine learning, Decision Tree, Random Forest, Ada Boost, k-NN.

Хом'як Тетяна Валеріївна – доцент кафедри системного аналізу та управління, к.ф.-м.н., Національний технічний університет «Дніпровська політехніка».
Сидоренко Катерина Віталіївна – студентка кафедри системного аналізу та управління, Національний технічний університет «Дніпровська політехніка».
Малієнко Андрій Вікторович – доцент кафедри системного аналізу та управління, к.т.н., Національний технічний університет «Дніпровська політехніка».
Мінєєв Олександр Сергійович – доцент кафедри системного аналізу та управління, к.т.н., Національний технічний університет «Дніпровська політехніка».

Khomiak Tetiana – Associate Professor of the Department of System Analysis and Control, Candidate of Physical and Mathematical Sciences, National Technical University "Dnipro University of Technology".
Sydorenko Kateryna – student of the Department of System Analysis and Control, National Technical University "Dnipro University of Technology".
Maliienko Andrii – Associate Professor of the Department of System Analysis and Control, Candidate of Technical Sciences, National Technical University "Dnipro University of Technology".
Mineyev Oleksandr – Associate Professor of the Department of System Analysis and Control, Candidate of Technical Sciences, National Technical University "Dnipro University of Technology".