УДК 004.89

Сидоренко К.В., магістр спеціальності 124 Системний аналіз
Науковий керівник: Хом'як Т.В., к. ф.-м. н., доцент кафедри системного аналізу та управління
(Національний технічний університет «Дніпровська політехніка», м. Дніпро, Україна)

ПРОГНОЗУВАННЯ ВИЯВЛЕННЯ ЦУКРОВОГО ДІАБЕТУ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Сьогодні кількість людей, які живуть з невиліковними хворобами зростає. Цукровий діабет - це серйозне захворювання, яке може призвести до численних ускладнень та проблем зі здоров'ям. Іноді люди через ускладнення діабету, які не контролюються, помирають, а саме може статися інфаркт, гіпоглікемія та інші. Зараз цукровий діабет є однією з найпоширеніших хронічних захворювань у світі, яким страждає близько 530 мільйонів людей, з яких 1 300 000 – громадяни України на червень 2023 року [3]. Це захворювання впливає на рівень цукру (глюкози) в крові. Поява цукрового діабету зазвичай обумовлена генетичними, середовищними та стильовими факторами. Основні причини включають генетичну схильність, ожиріння, неправильну харчову поведінку, інсулінорезистентність та шкідливі звички [4].

Значущим є той факт, що 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. Важливо підкреслити, що ці симптоми можуть бути ранніми показниками високого рівня глюкози у крові [3, 4].

Отже, для значно більшої ймовірності виявлення цукрового діабету до його появи потрібно мати не тільки більш досвідчених лікарів, а й навчитися прогнозувати дану хворобу для того, щоб у майбутньому не збільшувалась кількість хворих у рік.

З цією метою пропонується зробити прогнозування методоми машинного навчання: Decision Tree, Random Forest, K-NN, Ada Boost [1, 5-8].

Отже, перейдемо до результатів тренування та тестування розроблених моделей з наступними гіперпараметрами: кількість сусідів дорівнювала п'яти, а інші параметри за замовченням [10]. Результати тестування моделі K-NN Classifier наведені у таблиці 1.

Таблиця 1

Результати моделі K-NN Classifier

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

З таблиці 1 можна зробити наступні висновки, що модель здійснила не збалансовану класифікацію даних, бо 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 зробити балансування даних, наприклад, за допомогою методу SMOTEENN, після цього дані стали збалансовані, де кожен клас мав таку кількість значень: : «0» – 27948 значень, «1» – 39597, «2» – 35628. Після цього ще раз проведено тренування та тестування моделі, результати наведено у таблиці 2, з якої можна зробити наступні висновки: модель має високу точність для класів "перед діабет" і "має діабет", але меншу повноту для класу "не має діабет." Це означає, 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 [2, 10].

Таблиця 2

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	

Проведено тренування та тестування розроблених моделей з наступними гіперпараметрами: criterion= 'entropy', max_depth=40, а інші параметри за замовченням [6, 10]. Результати тестування моделі Decision Tree Classifier наведені у таблиці 3.

Таблиця 3

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	

Загальна 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, але прийнятна.

Проведено тренування та тестування розроблених моделей з наступними гіперпараметрами: n_estimators=100, max_features=16, max_depth=16, а інші параметри за замовченням [10, 11]. Результати тестування моделі Random Forest Classifier наведені у таблиці 4.

Таблиця 4

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 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, однак для інших класів точність і повнота менші, що може вказувати на більше помилкових класифікацій для цих груп.

В таблиці 5 наведено результати тестування моделі Ada Boost Classifier з наступними гіперпараметрами: n_estimators=100, max_features=16, max_depth=16, а інші параметри за замовченням [10].

Таблиця 5

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 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. Загальний рівень точності (accuracy) може бути збільшений для покращення ефективності моделі [2, 12].

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.

Список використаних джерел:

1. Машинне навчання простими словами. Частина 1. URL: http://www.mmf.lnu.edu.ua/ar/1739 (дата звернення: 01.10.2023 року)
2. Оцінка якості моделі класифікації. URL: https://studfile.net/preview/9974842/page:22/ (дата звернення: 05.10.2023 року)
3. Кількість діабетиків у світі до 2050 року може зрости майже втричі. URL: https://thepage.ua/ua/news/kilkist-diabetikiv-v-sviti-mozhe-zrosti-do-13-milyarda-lyudej-do-2050-roku (дата звернення: 02.10.2023 року)
4. 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. 265-271. (DOI – 10.46299/ISG.2023.2.6)
5. Ada Boost Classifier. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html (дата звернення: 13.10.2023 року)
6. Decision Tree Classifier. URL: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html (дата звернення: 05.10.2023 року)
7. Isolated Forest. URL: https://medium.com/@corymaklin/isolation-forest-799fceacdda4 (дата звернення: 03.10.2023 року)
8. K-Neighbors Classifier. URL: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (дата звернення: 09.10.2023 року)
9. Outliers in Machine Learning. URL: https://medium.com/analytics-vidhya/outliers-in-machine-learning-e830b2bd8660 (дата звернення: 01.10.2023 року)
10. Parameters and Hyperperemeters in Machine Learning and Deep Learning. URL: https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac (дата звернення: 05.10.2023 року)
11. Random Forest Classifier. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (дата звернення: 11.10.2023 року)
12. Standard Scaler. URL: learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (дата звернення: 03.10.2023 року)