УДК 004.6

І. В. Якимчук
С. О. Жуков

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ПЕРЕДБАЧЕННЯ ХВОРИХ НА ДІАБЕТ

Вінницький національний технічний університет

Анотація
Робота присвячена підготовці та розвідувальному аналізу даних для подальшого використання для інформаційної технології передбачення діабету методами машинного навчання. Було проведено аналіз датасету та його ознак.
Ключові слова: машинне навчання, передбачення, інформаційні технології, діабет, аналіз, модель, Python.

Abstract
The work is devoted to the preparation and intelligence analysis of data for further use for the information technology of diabetes prediction by machine learning methods. An analysis of the dataset and its features was carried out.
Keywords: machine learning, prediction, information technology, diabetes, analysis, model, Python.

Вступ

Сучасний світ стрімко змінюється завдяки технологічному прогресу, що відкриває нові можливості для застосування інформаційних технологій у медицині. Високий рівень автоматизації, доступність великих обсягів даних та розвиток алгоритмів машинного навчання дозволяють створювати нові підходи до діагностики, прогнозування та профілактики захворювань.

Одним із ключових напрямів застосування штучного інтелекту є передбачення захворювань на основі наявних медичних даних. Однією з найбільш поширених та соціально значущих хвороб є діабет, що вимагає ефективних методів раннього виявлення для зниження ризиків ускладнень та покращення якості життя пацієнтів.

У цьому дослідженні здійснено розвідувальний аналіз даних (EDA) на основі відкритого набору даних “Pima Indians Diabetes Database”, що доступний на платформі Kaggle [1]. Основною метою є визначення ключових факторів ризику, які впливають на розвиток діабету, та створення ефективної моделі для прогнозування ймовірності виникнення захворювання.

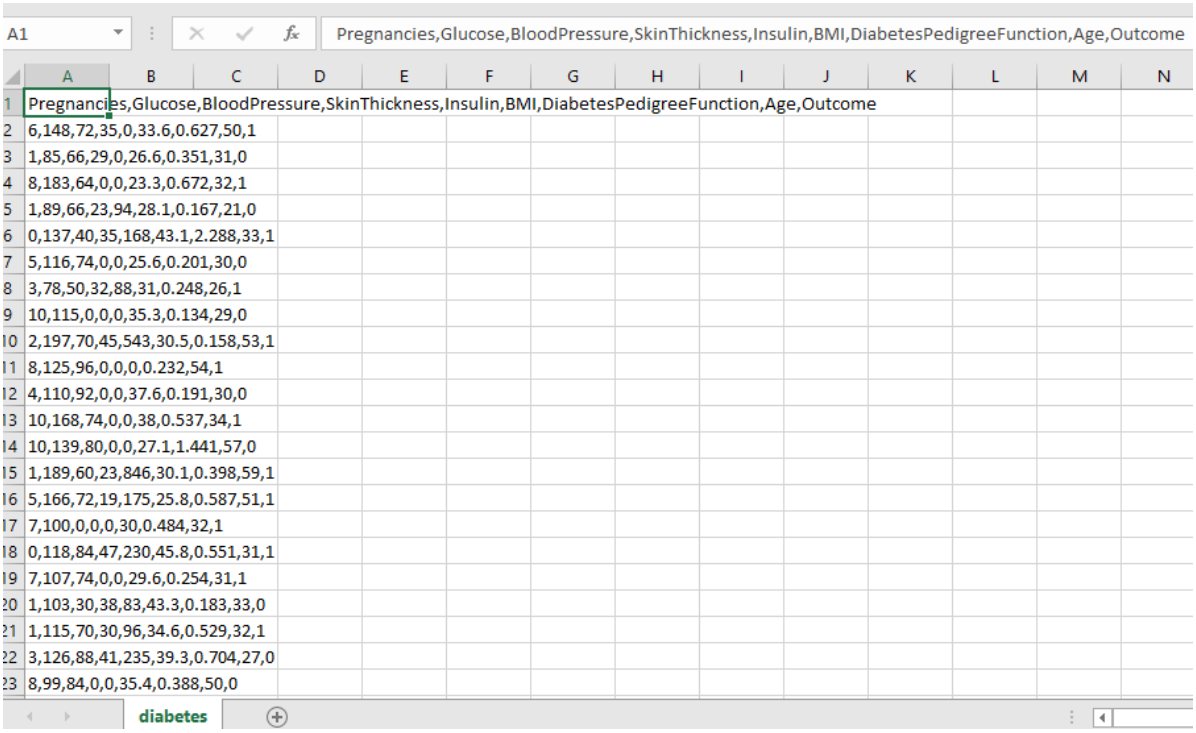
Розвідувальний аналіз

Датасет “Pima Indians Diabetes Database” містить 768 записів та включає 8 основних атрибутів, що є потенційними факторами ризику розвитку діабету:

- Pregnancies — кількість вагітностей;
- Glucose — рівень глюкози натще;
- BloodPressure — діастолічний артеріальний тиск (мм рт. ст.);
- SkinThickness — товщина шкірної складки трицепса (мм);
- Insulin — рівень інсуліну (мкУ/мл);
- BMI — індекс маси тіла (кг/м²);
- DiabetesPedigreeFunction — показник генетичної схильності до діабету;
- Age — вік пацієнта.

Цільова змінна Outcome визначає наявність або відсутність діабету у пацієнта (1 — діабет, 0 — відсутність діабету).

Цей набір даних надає широкий спектр інформації, що дозволяє провести розвідувальний аналіз, визначити ключові фактори, які впливають на розвиток діабету, та створити модель для прогнозування хвороби. Розподіл атрибутів та їхній взаємозв'язок будуть проаналізовані для кращого розуміння природи даних (рис. 1).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Pregnancies,Glucose,BloodPressure,SkinThickness,Insulin,BMI,DiabetesPedigreeFunction,Age,Outcome													
2	6,148,72,35,0,33.6,0.627,50,1													
3	1,85,66,29,0,26.6,0.351,31,0													
4	8,183,64,0,0,23.3,0.672,32,1													
5	1,89,66,23,94,28.1,0.167,21,0													
6	0,137,40,35,168,43.1,2.288,33,1													
7	5,116,74,0,0,25.6,0.201,30,0													
8	3,78,50,32,88,31,0.248,26,1													
9	10,115,0,0,0,35.3,0.134,29,0													
10	2,197,70,45,543,30.5,0.158,53,1													
11	8,125,96,0,0,0,0.232,54,1													
12	4,110,92,0,0,37.6,0.191,30,0													
13	10,168,74,0,0,38,0.537,34,1													
14	10,139,80,0,0,27.1,1.441,57,0													
15	1,189,60,23,846,30.1,0.398,59,1													
16	5,166,72,19,175,25.8,0.587,51,1													
17	7,100,0,0,0,30,0.484,32,1													
18	0,118,84,47,230,45.8,0.551,31,1													
19	7,107,74,0,0,29.6,0.254,31,1													
20	1,103,30,38,83,43.3,0.183,33,0													
21	1,115,70,30,96,34.6,0.529,32,1													
22	3,126,88,41,235,39.3,0.704,27,0													
23	8,99,84,0,0,35.4,0.388,50,0													

Рис. 1. Приклад ознак пацієнтів, що містить набір даних

Проведено попереднє очищення даних для усунення пропущених значень та аномалій. Було здійснено візуалізацію розподілу ознак залежно від цільової змінної (рис. 2).

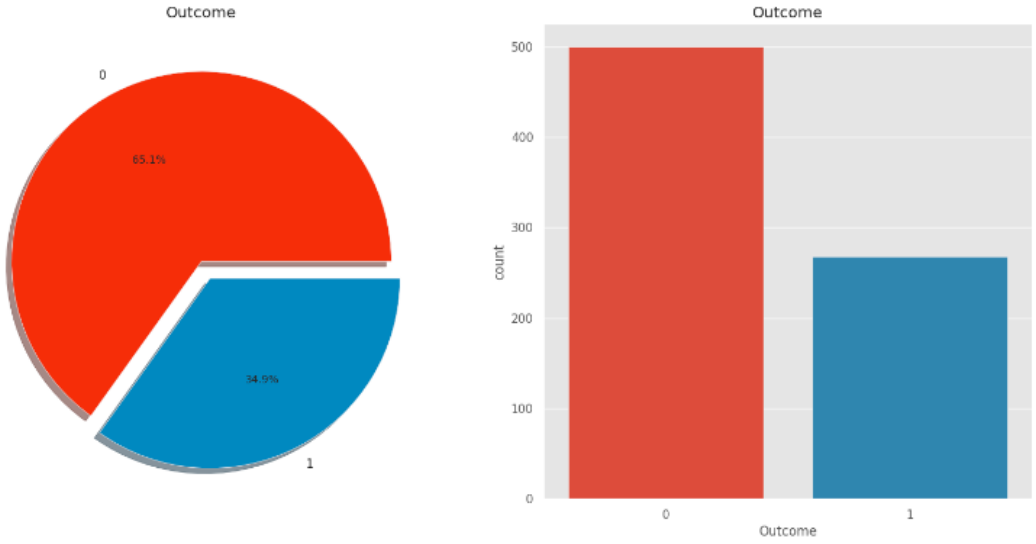


Рис. 2. Гістограма рівнів розподілу категорії здоровий та діабет

Було побудовано кореляційну матрицю (рис. 3), яка виявила сильну залежність між рівнем глюкози та наявністю діабету.

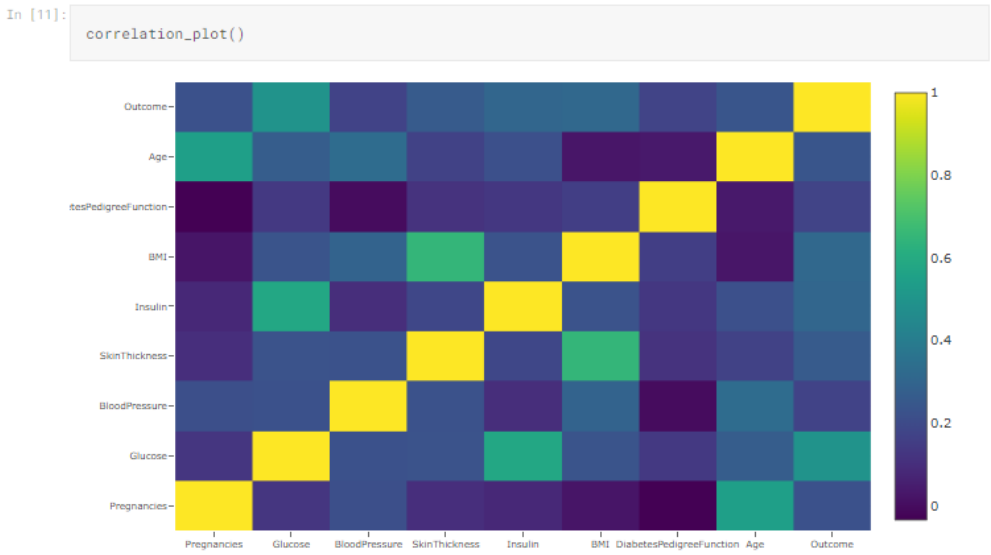


Рис. 3. Матриця кореляції

Порівняння розподілу ознак між здоровими та діабетичними пацієнтами дозволило виділити ключові предиктори для подальшого моделювання.

Побудовано графік, який дозволяє нам отримати важливі інсайти про розподіл даних у наборі даних, що може бути корисним для розвідувального аналізу та подальшого моделювання. (рис. 4).

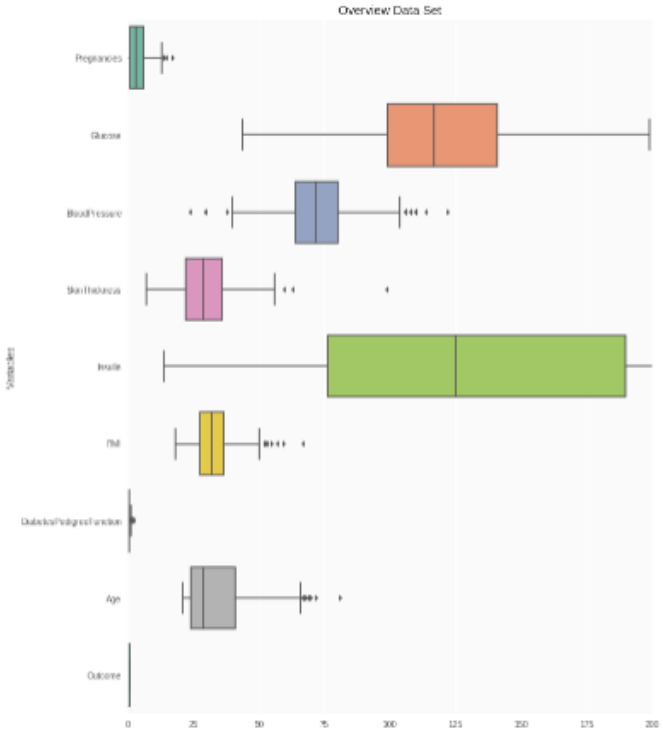


Рис. 4. Графік ящика з вусами дозволяє виявити потенційні викиди або аномалії в розподілі даних.

Аналізуючи boxplot, можна швидко оцінити основні характеристики вибірки та визначити, чи є потреба в подальшому очищенні даних або застосуванні трансформацій для покращення моделі.

Висновки

Під час розвідувального аналізу набору даних «Pima Indians Diabetes Database», що містить інформацію про параметри та фактори ризику розвитку діабету, було досліджено вплив різних ознак на цільову змінну – наявність діабету.

На основі матриці кореляції визначено залежність між ключовими параметрами, що впливають на ймовірність виникнення діабету. Зокрема, позитивна кореляція виявлена між рівнем глюкози в крові натще, індексом маси тіла (BMI) та рівнем інсуліну з ймовірністю діагностування діабету (Outcome). Це свідчить про те, що ці фактори мають найбільший вплив на розвиток захворювання.

Аналіз розподілу категорій здорових та хворих на діабет пацієнтів показав, що розподіл є достатньо збалансованим, що позитивно впливає на побудову моделей передбачення, оскільки зменшує ризик зміщення у бік більшої або меншої групи.

Крім того, побудовано графік ящика з вусами (boxplot), що дозволило виявити потенційні викиди, які можуть свідчити про аномалії у вибірці, що потребують додаткового дослідження для підвищення точності моделі.

Отримані результати підтверджують, що обрані ознаки (рівень глюкози, індекс маси тіла, рівень інсуліну) є важливими факторами для побудови інформаційної технології прогнозування ризику захворювання на діабет.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Ivan Yakymchuk notebook [Електронний ресурс] – Режим доступу: https://www.kaggle.com/code/marlos11/notebook308ced0f9b
2. Pandas Getting started. 2024 [Електронний ресурс] – Режим доступу: https://pandas.pydata.org/docs/getting_started/index.html
3. Matplotlib Pyplot Documentation. 2024 [Електронний ресурс]. – Режим доступу: https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html
4. Seaborn Tutorial. 2023 [Електронний ресурс]. – Режим доступу: https://seaborn.pydata.org/tutorial.html

Якимчук Іван Володимирович – студент групи ІСТ-23м, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: iyakim2211@gmail.com

Жуков Сергій Олександрович– к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, e-mail: sazhukov@gmail.com

Yakymchuk Ivan V. - student of Faculty of Intelligent Information Technology and Automation, IST-23m, Vinnytsia National Technical University, Vinnytsia, e-mail iyakim2211@gmail.com

Zhukov Serhii O. - Ph.D., Assistant Professor of the Department of Systems Analysis and Information Technology, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@gmail.com