УДК 004.6

А. Д. Пянкевич
С. О. Жуков

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ПЕРЕДБАЧЕННЯ СТУПЕНЯ ОЖИРІННЯ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Робота присвячена підготовці та розвідувальному аналізу даних для подальшого використання для інформаційної технології передбачення ступеня ожиріння методами машинного навчання. Було проведено аналіз датасету та його ознак.

Ключові слова: ожиріння, інформаційні технології, машинне навчання, аналіз даних, передбачення, ознаки, передбачення ступеня ожиріння.

Abstract

The paper is devoted to the preparation and exploratory analysis of data for further use in the information technology of predicting the degree of obesity using machine learning methods. The dataset and its features were analyzed.

Keywords: obesity, information technology, machine learning, data analysis, predictions, signs, obesity predictions.

Вступ

Кожного дня технології у світі стрімко розвиваються, те що умовно кажучи учора було фантастикою, завтра уже реальність. Особливо це відчувається у сфері інформаційних технологій, які стали основою сьогодення. З їхньою допомогою відкриваються нові можливості для вирішення завдань, які раніше вважались неможливими. Одним із таких завдань, де інформаційні технології допомагають вирішувати проблеми та завдання, є медицина. Наприклад для встановлення діагнозу при наявності певних ознак та аналізів чи прогнозування вірогідності захворіти.

Одним із таких захворювань, які можливо спрогнозувати, є ожиріння. Передбачення даного захворювання дасть змогу завчасно підготуватись до такого розвитку подій, зменшити ризики такого захворювання та в цілому вжити найбільш ефективних заходів для запобігання захворювання.

Виходячи з цього, використання інформаційних технологій для обробки та аналізу даних дає можливості для удосконалення діагностичних методів, методів передбачення та запобігання.

Розвідувальний аналіз

Для проведення аналізу було обрано набір даних, що має назву «Obesity or CVD risk» та опублікований користувачем «ARAVINDPCODER» та має відкритий доступ для загального використання на платформі Kaggle [1]. Даний датасет містить у собі широкий спектр даних про пацієнтів лікарень з оцінками рівня ожиріння у людей. Всього 17 атрибутів та 2111 записів. Серед атрибутів, пов'язаних з харчовими звичками можна виділити такі: споживання висококалорійної їжі (FAVC), частота споживання овочів (FCVC), кількість основних прийомів їжі (NCP), споживання їжі між прийомами їжі (CAEC), споживання води щодня (CH20) та споживання алкоголю (CALC). Атрибути, пов'язані з фізичним станом: моніторинг споживання калорій (SCC), частота фізичної активності (FAF), час використання технологічних пристроїв (TUE), використаний транспорт (MTRANS). Також серед атрибутів є стать, вік, зріст та вага. (рис. 1).

	id	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC
0	0	Male	24.443011	1.699998	81.669950	yes	yes	2.000000	2.983297	Sometimes
1	1	Female	18.000000	1.560000	57.000000	yes	yes	2.000000	3.000000	Frequently
2	2	Female	18.000000	1.711460	50.165754	yes	yes	1.880534	1.411685	Sometimes
3	3	Female	20.952737	1.710730	131.274851	yes	yes	3.000000	3.000000	Sometimes
4	4	Male	31.641081	1.914186	93.798055	yes	yes	2.679664	1.971472	Sometimes

NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
2.983297	Sometimes	no	2.763573	no	0.000000	0.976473	Sometimes	Public_Transportation	Overweight_Level_II
3.000000	Frequently	no	2.000000	no	1.000000	1.000000	no	Automobile	Normal_Weight
1.411685	Sometimes	no	1.910378	no	0.866045	1.673584	no	Public_Transportation	Insufficient_Weight
3.000000	Sometimes	no	1.674061	no	1.467863	0.780199	Sometimes	Public_Transportation	Obesity_Type_III
1.971472	Sometimes	no	1.979848	no	1.967973	0.931721	Sometimes	Public_Transportation	Overweight_Level_II

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

Проведено попереднє очищення даних та розвідувальний аналіз даних. Побудовано розподіл рівнів ожиріння (рис.1).

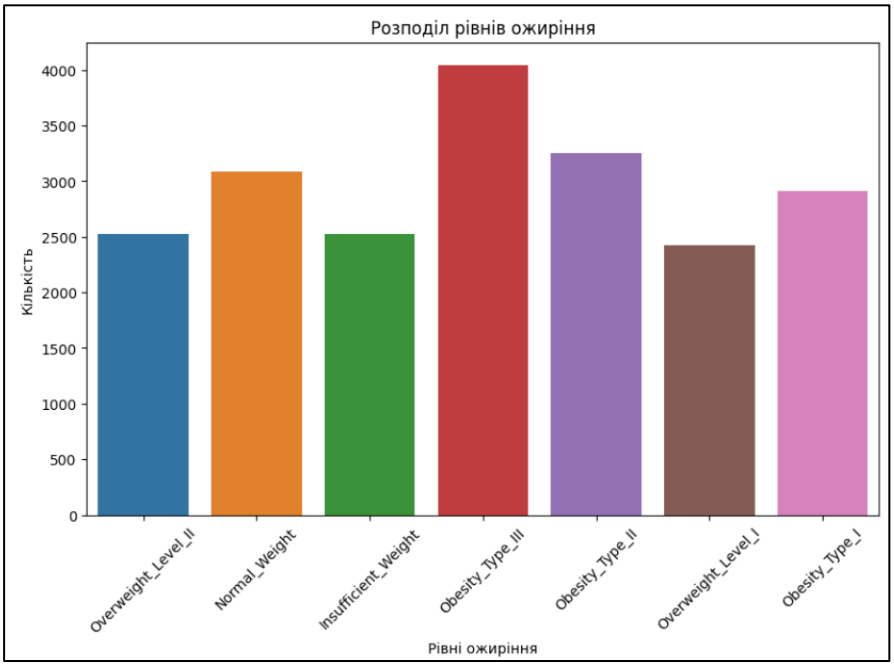

Рис. 1. Гістограма розподілу рівнів ожиріння, що містить набір даних

З гістограми видно, що розподіл рівномірний, що гарно вплине на передбачення. Підготувавши дані було побудовано матрицю кореляції, що дасть змогу виявити явну залежність та зв’язки між різними параметрами та ознаками.

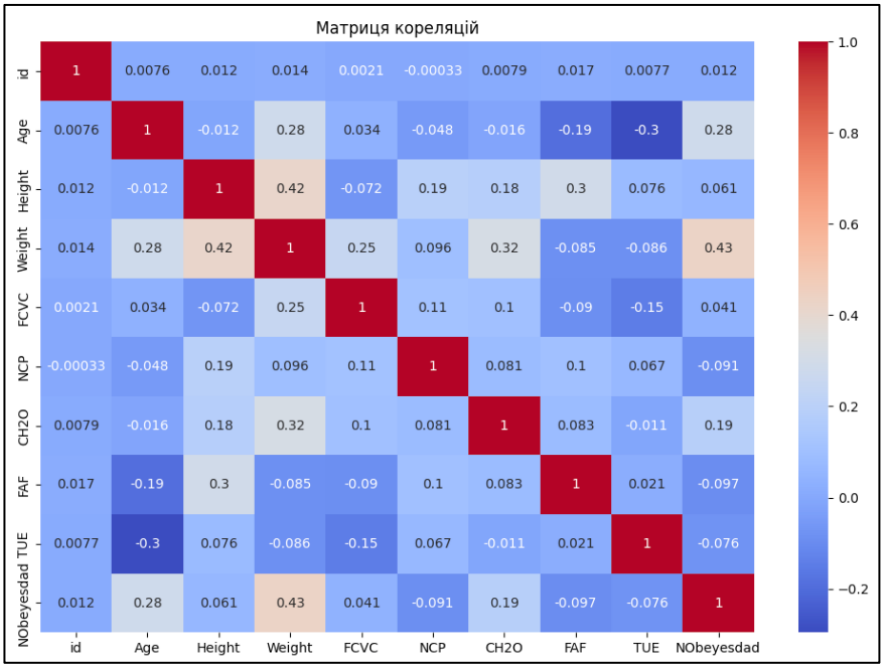

Рис. 2. Матриця кореляції

З рисунку 2 можна зробити висновок що найбільший вплив на рівень ожиріння мають такі параметри, як вага (Weight), вік (Age) та споживання води щодня (CH2O).

Побудовано графік boxplot, який показує як IMT (індекс маси тіла) розподіляється для різних категорій ожиріння (рис. 3).

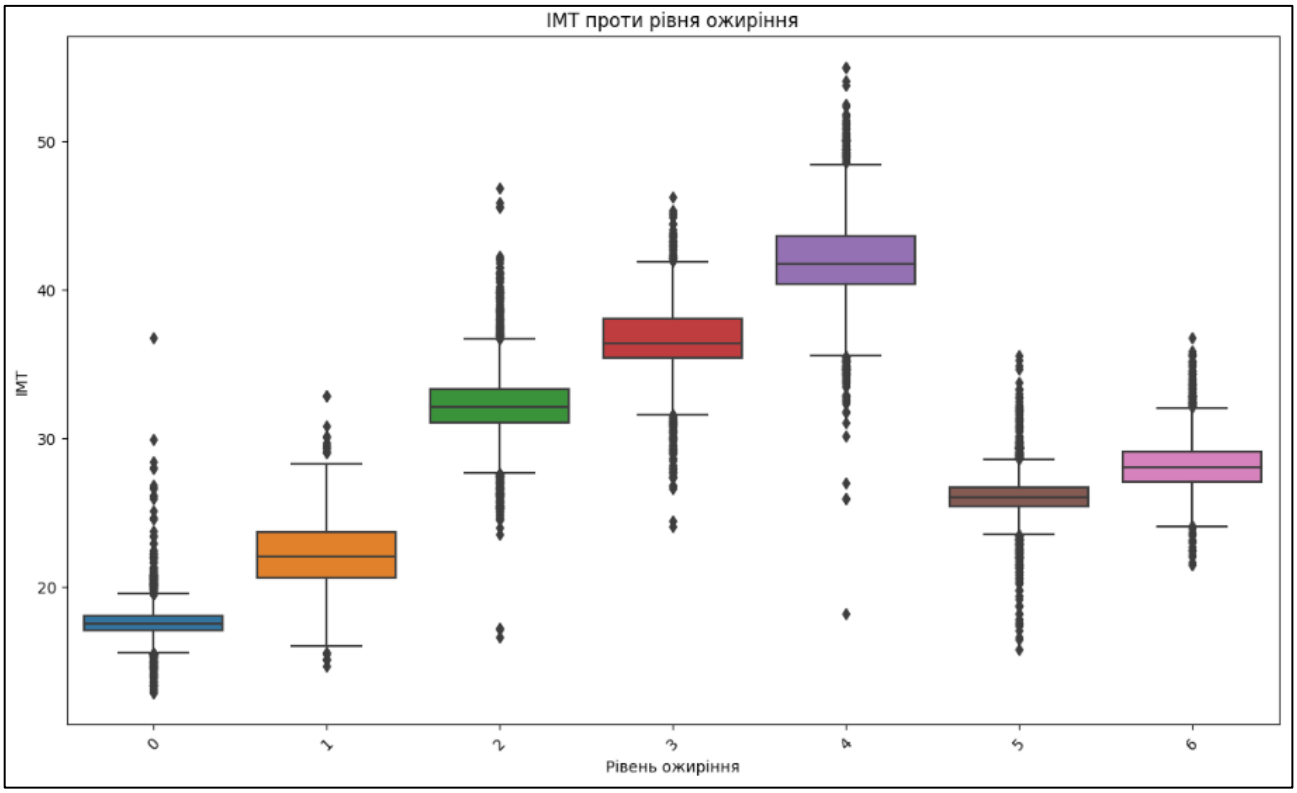

Рис. 3. Графік boxplot залежності індексу маси тіла (IMT) від рівня ожиріння

З рисунку 3 видно, що є чітка тенденція – вищі значення IMT відповідають вищим рівням ожиріння. Кожен boxplot відображає: мінімальне значення IMT в групі, нижній квартиль (25-й перцентиль),

медіану (50-й перцентиль), верхній квартиль (75-й перцентиль), максимальне значення ІМТ в групі та аномальні значення або «викиди» (outliers), якщо вони є. Такий графік допомагає зрозуміти, як змінюється ІМТ в залежності від рівня ожиріння, та виявити потенційні аномалії.

Висновки

При розвідувальному аналізі набору даних «Obesity or CVD risk», що містить у собі інформацію про параметри та ризики захворювання на ожиріння було досліджено вплив різник ознак на показник ступеня ожиріння. Побудовано матрицю кореляції, яка показує залежність між усіма параметрами та зроблено висновки щодо залежності параметру ступеня ожиріння (NObeyesdad) до інших ознак.

Побудовано гістограму розподілу рівнів ожиріння, яка показала, що розподіл достатньо рівномірний, що позитивне вплине на передбачення.

Також було побудовано графік boxplot, який показує як ІМТ (індекс маси тіла) розподіляється для різних категорій ожиріння. Виявлена чітка тенденція – вищі значення ІМТ відповідають вищим рівням ожиріння

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Obesity Prediction Dataset. Kaggle. 2023 [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster/data
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

Пянкевич Андрій Дмитрович – студент групи СА-20б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: piankevych2003@gmail.com

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

Piankevych Andriy D. - student of Faculty of Intelligent Information Technology and Automation, SA-20b, Vinnytsia National Technical University, Vinnytsia, e-mail piankevych2003@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