УДК: 004.89:616.12

Д. С. Савчук
С. О. Жуков

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ ПЕРЕДБАЧЕННЯ СЕРЦЕВО-СУДИННИХ ЗАХВОРЮВАНЬ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Робота присвячена підготовці та розвідувальному аналізу даних для подальшого використання в інформаційно-аналітичній системі прогнозування ризиків серцево-судинних захворювань та мінімізації клінічних ризиків методами машинного навчання. Проведено комплексний аналіз структури медичного датасету «Heart Failure Prediction Dataset», досліджено розподіли категоріальних та числових ознак, а також виявлено закономірності розвитку патологій залежно від вікових характеристик, типу болю в грудях, наявності стенокардії та специфічних змін сегмента ST на електрокардіограмі.

Ключові слова: серцево-судинні захворювання, розвідувальний аналіз даних, машинне навчання, діагностичні дані, сегмент ST, системний аналіз.

Abstract

The paper is devoted to the preparation and exploratory data analysis for further use in the information-analytical system of predicting cardiovascular disease risks and minimizing clinical risks using machine learning methods. A comprehensive analysis of the structure of the medical dataset "Heart Failure Prediction Dataset" was conducted, the distributions of categorical and numerical features were investigated, and the regularities of pathology development depending on age characteristics, chest pain type, exercise-induced angina, and specific ST segment changes on the electrocardiogram were revealed.

Keywords: cardiovascular diseases, exploratory data analysis, machine learning, diagnostic data, ST segment, systems analysis.

Вступ

Сучасні інформаційні технології інтелектуального аналізу даних посідають домінуюче місце у сфері біомедичної інженерії та охорони здоров’я, зокрема при розробці систем підтримки прийняття клінічних рішень (Clinical Decision Support Systems, CDSS). Серцево-судинні захворювання (ССЗ) залишаються головною причиною передчасної смертності та інвалідизації населення у всьому світі, що вимагає радикального переходу від реактивного лікування наслідків до проактивної превентивної профілактики.

Алгоритми машинного навчання та ансамблевого моделювання відкривають широкі можливості для побудови гнучких діагностичних контурів, здатних враховувати складні нелінійні взаємозв'язки між різнорідними фізіологічними параметрами пацієнта. Проте першочерговим, фундаментальним етапом проєктування таких технологій є розвідувальний аналіз даних (Exploratory Data Analysis, EDA). Він дозволяє формалізувати структуру вхідних масивів інформації, локалізувати приховані закономірності, виявити проблему асиметрії розподілів чи технічних аномалій вимірювального обладнання та підготувати надійну, очищену основу для етапу безпосереднього математичного моделювання.

Розвідувальний аналіз діагностичних даних

Для проведення системного аналізу та побудови інтелектуального предиктора в межах кваліфікаційної роботи було обрано реальний клінічний набір даних «Heart Failure Prediction Dataset», доступний у відкритому репозиторії обчислювальної платформи Kaggle [1]. Зазначений масив інформації консолідує відомості про 918 пацієнтів та містить 11 незалежних предикторів (фізіологічних параметрів, результатів динамічних тестів під навантаженням та демографічних характеристик), а також 1 бінарну цільову змінну.

Початкова структура матриці даних включає такі ключові параметри, як вік, стать, тип болю в грудях, артеріальний тиск у стані спокою, рівень сироваткового холестерину, цукор натщесерце, результати електрокардіографії у спокої, максимальний досягнутий пульс, наявність індукованої стенокардії та нахил сегмента ST під час навантаження. Схематичне представлення первинної структури та типів даних наведено на рисунку 1.

Розмір датасету: 918 рядків та 12 колонок

Перші 5 рядків датасету

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0	Up	0

Рис. 1. Фрагмент структури вхідного набору даних діагностики серцево-судинних захворювань

Цільовою змінною дослідження є бінарна ознака HeartDisease, яка фіксує наявність підтвердженого серцево-судинного захворювання (1 – патологія наявна, 0 – нормальний стан). Важливою системною характеристикою досліджуваного об'єкта є відсутність критичного дисбалансу класів, що значно підвищує стабільність навчання класифікаторів. Графічний розподіл цільової величини відображено на рисунку 2.

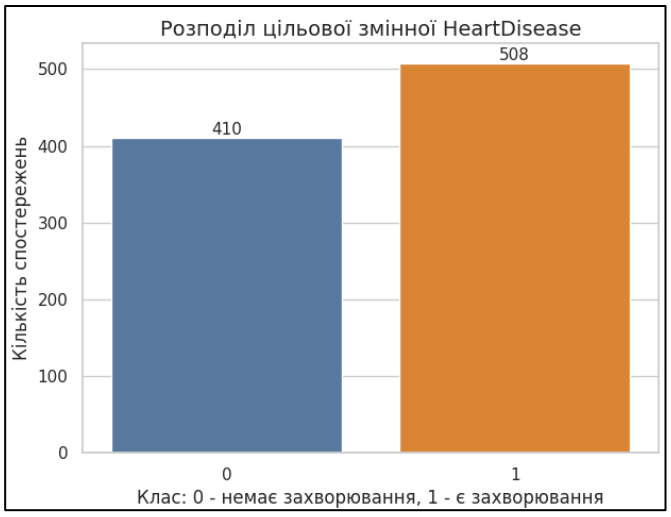

Рис. 2. Графік розподілу класів цільової змінної HeartDisease у вибірці

Статистичний розподіл свідчить, що 508 пацієнтів (55.34%) належать до класу з наявною патологією, тоді як група норми налічує 410 спостережень (44.66%). Таке рівномірне співвідношення дозволяє відмовитися від методів штучного балансування (наприклад, SMOTE), проте вимагає використання обов’язкової стратифікації при поділі даних на навчальну та тестову підвибірки для виключення викривлення оцінок.

У межах багатофакторного аналізу категоріальних ознак було встановлено чіткі клінічні закономірності. Виявлено значну гендерну асиметрію: чоловіки мають суттєво вищу частоту діагностованих патологій порівняно з жінками. Аналіз типу болю в грудях (ChestPainType) показав, що найбільш підступним є безсимптомний характер перебігу (категорія ASY), де частка підтверджених захворювань є максимальною, що доводить небезпеку латентного розвитку ішемії. Наявність стенокардії, викликаної фізичним навантаженням (ExerciseAngina = Y), демонструє критично сильний зв'язок із позитивним класом захворювання. Також встановлено, що нахил ділянки сегмента ST на ЕКГ під час навантаження є ключовим розділяючим маркером: висхідний нахил (Up) у 80% випадків відповідає нормі, тоді як плоский (Flat) та низхідний (Down) строго асоціюються з наявністю ССЗ.

Оскільки вік є фундаментальним чинником деградації судинної системи, було досліджено динаміку захворюваності за віковими групами пацієнтів (рис. 3).

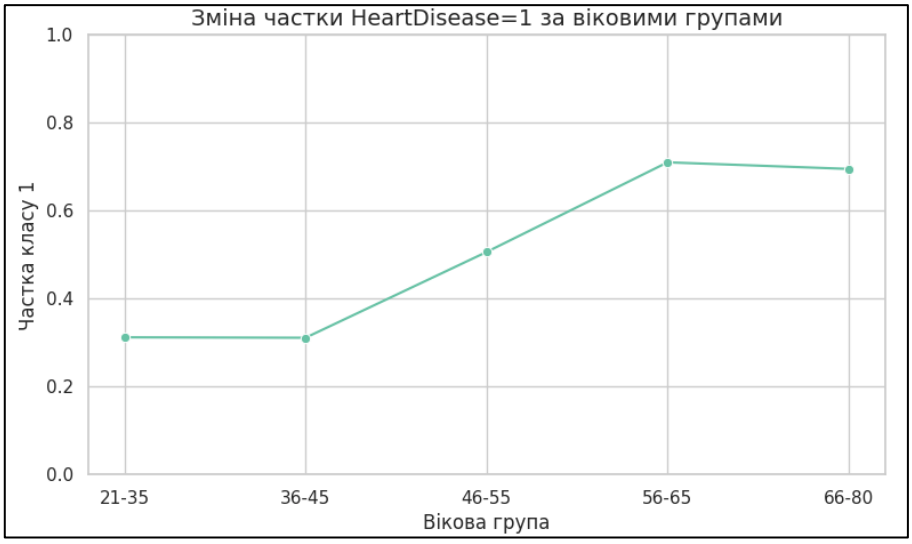

Рис. 3. Розподіл частки серцево-судинних захворювань за віковими групами пацієнтів

Візуалізація (рис. 3) чітко підтверджує нелінійне стрибкоподібне зростання ризиків патології з віком. У молодих групах пацієнтів (до 45 років) частка захворювань є мінімальною і становить близько 31%. Однак, починаючи з вікової когорти 46–55 років, показник стрімко долає межу у 50.6%, а для пацієнтів віком понад 56 років досягає свого максимуму – понад 70%, що обґрунтовує доцільність проведення примусової вікової дискретизації (бакетізації) ознаки на етапі Feature Engineering.

Для формалізації лінійних зв'язків між неперервними величинами та виключення явища мультиколінеарності було розраховано матрицю коефіцієнтів кореляції Пірсона (рис. 4).

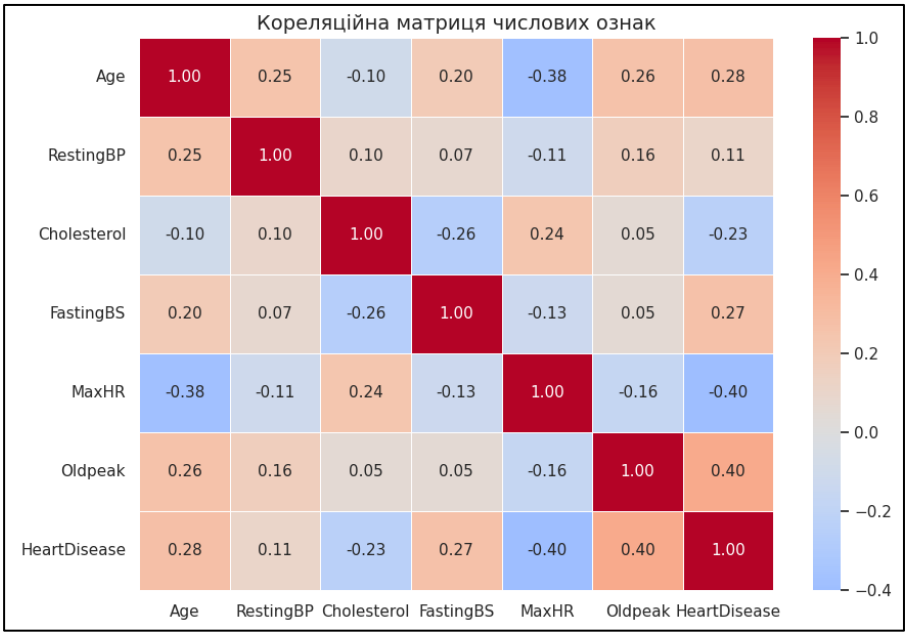

Рис. 4. Матриця лінійної кореляції числових діагностичних ознак

З аналізу кореляційної матриці (рис. 4) встановлено відсутність мультиколінеарності: жодна пара незалежних чинників не наближається до небезпечного порогу 0.7, що свідчить про високу стабільність простору ознак. Найбільший прямий зв'язок із цільовим класом демонструє депресія сегмента ST (Oldpeak, r = 0.40) та максимальна частота серцевих скорочень (MaxHR, r = -0.40), яка має природну обернену залежність.

Проте системний аналіз матриці виявив важливу аномалію: рівень холестерину (Cholesterol) показав стійкий від’ємний зв'язок із хворобами серця (r = -0.23), що суперечить медичній теорії. Поглиблена перевірка виявила наявність 172 нульових значень у цій колонці, які є прихованими пропусками даних (технічний шум вимірювань). Даний факт строго доводить необхідність обов’язкового застосування процедури групової імпутації пропусків медіаною та логарифмічного згладжування перед початком моделювання.

Висновки

У ході розвідувального аналізу даних було повністю досліджено структуру, внутрішні закономірності та приховані аномалії клінічного набору даних «Heart Failure Prediction Dataset». Встановлено, що ключовими функціональними предикторами серцево-судинного ризику є геометричні зміни сегмента ST на ЕКГ, максимальний досягнутий пульс пацієнта та характер больових відчуттів у грудях. Виявлений латентний шум у вигляді нульових значень холестерину та нелінійний стрибок ризику після 46 років обґрунтовують необхідність проектування конвеєрів автоматизованого препроцесингу (Pipeline, ColumnTransformer), логарифмування та дискретизації віку. Отримані результати сформували надійний фундамент для побудови та успішного налаштування високоточних ансамблевих моделей класифікації.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Heart Failure Prediction Dataset. [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data
2. Pandas Tutorial. [Електронний ресурс] – Режим доступу: https://www.w3schools.com/python/pandas/default.asp
3. Matplotlib Pyplot Documentation. [Електронний ресурс]. – Режим доступу: https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html
4. Seaborn: Statistical Data Visualization. [Електронний ресурс]. Режим доступу: https://seaborn.pydata.org/

Савчук Дмитро Сергійович – студент групи СА-22б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця. e-mail: savchukodmytro@gmail.com

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

Savchuk Dmytro – student of Faculty of Intellectual Information Technologies and Automation, SA-22b, Vinnytsia National Technical University, Vinnytsia, e-mail savchukodmytro@gmail.com

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