УДК: 004.8:519.2:663.2

Є. М. Лавренюк
С. О. Жуков

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ МЕТОДУ ГОЛОВНИХ КОМПОНЕНТ ДЛЯ ЗНИЖЕННЯ РОЗМІРНОСТІ ОЗНАК У ЗАДАЧАХ МЕДИЧНОЇ ДІАГНОСТИКИ

Вінницький національний технічний університет

Анотація

У роботі досліджено застосування методу головних компонент (PCA) для виявлення інформативних ознак у медичних даних та зниження їх розмірності в задачах автоматизованої класифікації. Використано набір даних Breast Cancer Wisconsin (Diagnostic), який містить числові характеристики клітинних ядер пухлин. Проведено попередню обробку даних, стандартизацію ознак, аналіз мультиколінеарності та побудову класифікаційних моделей. Реалізовано порівняння логістичної регресії на повному наборі ознак і моделі після застосування PCA. Встановлено, що використання 9 головних компонент дозволяє зберегти понад 94% дисперсії та зменшити розмірність даних у 3,3 раза без втрати якості класифікації.

Ключові слова: метод головних компонент, PCA, машинне навчання, медичні дані, класифікація, логістична регресія, зниження розмірності.

Abstract

The paper investigates the application of Principal Component Analysis (PCA) for feature extraction and dimensionality reduction in medical data classification tasks. The Breast Cancer Wisconsin (Diagnostic) dataset was used, containing numerical characteristics of tumor cell nuclei. Data preprocessing, feature standardization, multicollinearity analysis, and classification model development were performed. A comparison was made between logistic regression based on the full feature set and a model using the principal component space. It was found that using 9 principal components preserves more than 94% of the variance and reduces data dimensionality by 3.3 times without loss of classification quality.

Keywords: Principal Component Analysis, PCA, machine learning, medical data, classification, logistic regression, dimensionality reduction.

Вступ

Сучасний розвиток інформаційних технологій та цифрової медицини спричиняє постійне зростання обсягів медичних даних, які потребують ефективної обробки, аналізу та інтерпретації. Особливо актуальними є задачі автоматизованої діагностики, де необхідно швидко й точно класифікувати медичні об’єкти на основі великої кількості параметрів.

Медичні набори даних часто характеризуються високою розмірністю, наявністю корельованих ознак, шумів та надлишкової інформації. Це може ускладнювати побудову моделей машинного навчання, знижувати їх стабільність і підвищувати обчислювальні витрати. Одним із ефективних підходів до вирішення цієї проблеми є метод головних компонент (Principal Component Analysis, PCA), який дозволяє перетворити початковий простір ознак у простір меншої розмірності зі збереженням основної частини інформації[2, 3].

Метою роботи є дослідження ефективності застосування методу головних компонент для зниження розмірності медичних даних у задачах класифікації та оцінка його впливу на якість моделей машинного навчання.

Основна частина

Для дослідження було використано набір даних Breast Cancer Wisconsin (Diagnostic), який містить інформацію про морфологічні характеристики клітинних ядер пухлин[1]. У наборі представлено 569 спостережень та 30 числових ознак, що описують такі параметри, як радіус, текстура, периметр, площа, гладкість, компактність, ввігнутість, симетрія та інші характеристики клітин. Цільовою змінною є діагноз, який визначає тип пухлини: доброякісна або злоякісна.

На етапі розвідувального аналізу даних було досліджено структуру набору, типи змінних, наявність пропущених значень та розподіл класів. Встановлено, що набір даних є придатним для подальшого моделювання, однак має певний дисбаланс класів: більшість записів належить до доброякісних пухлин, а менша частина — до злоякісних. Такий розподіл необхідно враховувати під час оцінювання моделей класифікації, оскільки в медичних задачах важливо не лише досягти високої загальної точності, але й мінімізувати кількість помилкових пропусків злоякісних випадків.

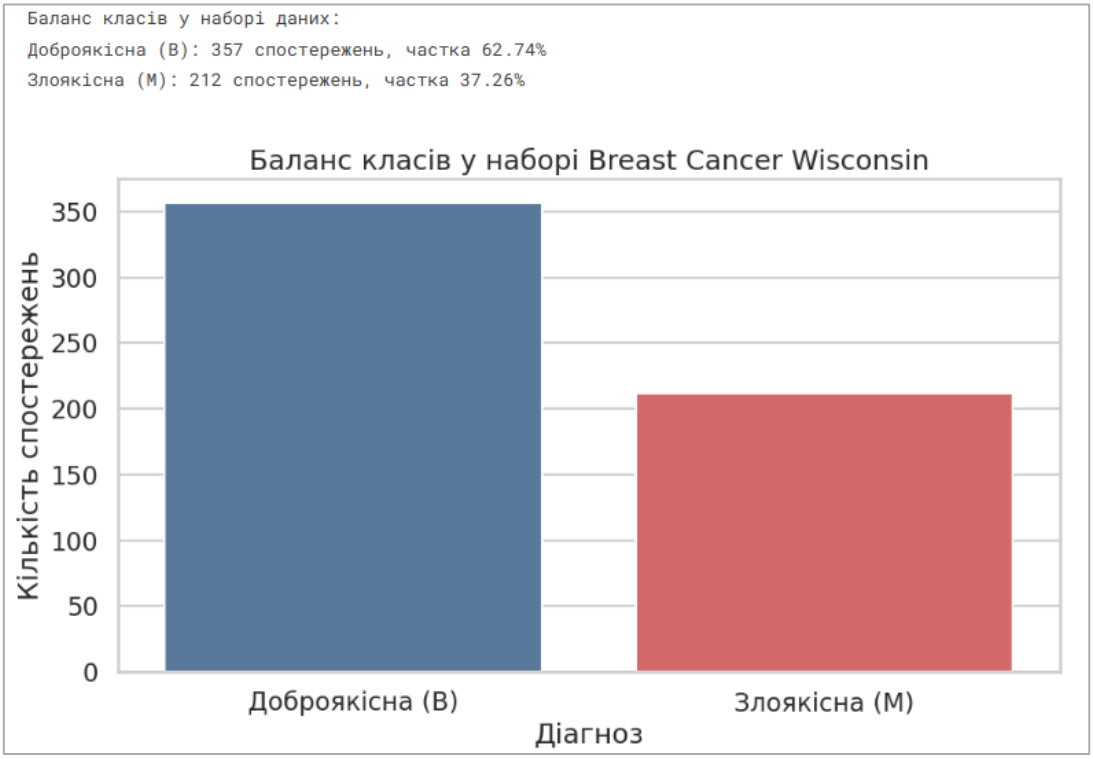


Рис. 1. Розподіл класів у наборі даних Breast Cancer Wisconsin

Рисунок 1 демонструє співвідношення між доброякісними та злоякісними випадками у досліджуваному наборі даних. Наявність двох класів підтверджує, що задача належить до задач бінарної класифікації. При цьому дисбаланс класів не є критичним, однак він обґрунтовує необхідність використання не лише метрики Accuracy, а й Precision, Recall та F1-score[5–7].

Подальший аналіз було спрямовано на дослідження взаємозв'язків між ознаками. Кореляційний аналіз показав, що значна частина числових параметрів має сильні взаємні залежності. Найбільш виражена кореляція спостерігається між характеристиками, пов'язаними з геометричними параметрами клітин, зокрема радіусом, периметром і площею. Це свідчить про наявність мультиколінеарності та надлишкової інформації у початковому просторі ознак.

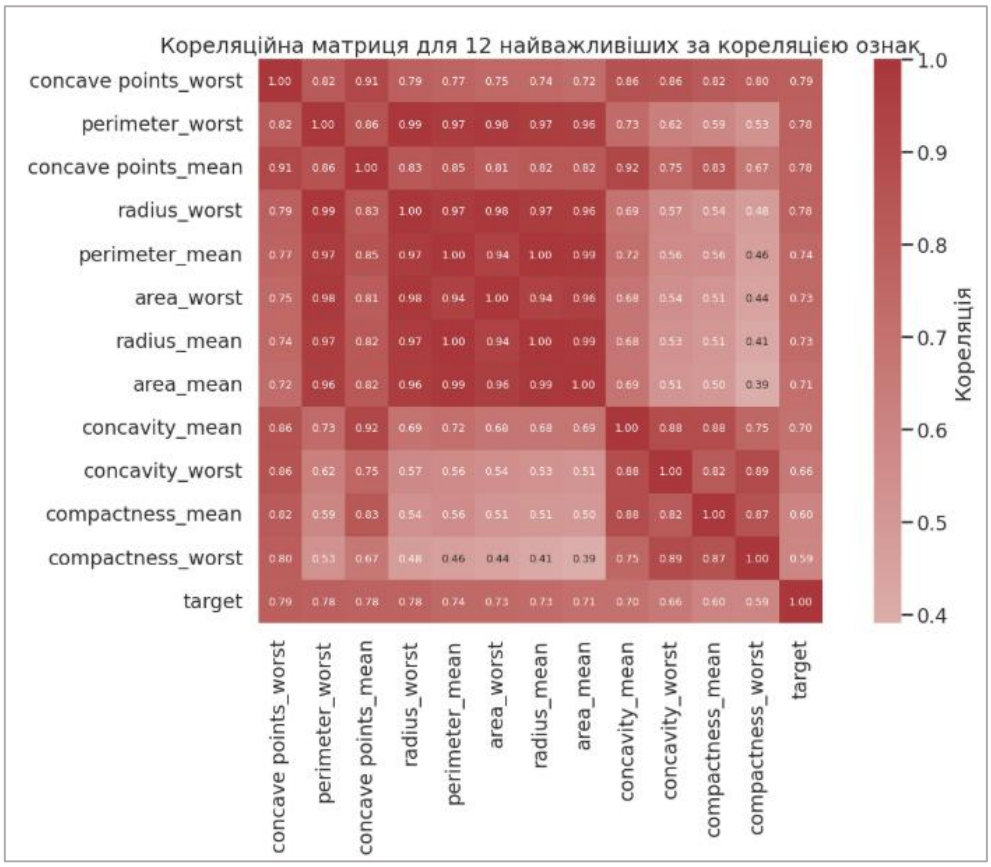


Рис. 2. Кореляційна матриця ознак набору Breast Cancer Wisconsin

Рисунок 2 ілюструє наявність сильних кореляцій між окремими числовими ознаками. Така структура даних є типовою для медичних наборів, у яких різні параметри можуть описувати близькі за змістом характеристики одного об’єкта. Саме тому застосування методу головних компонент є доцільним, оскільки PCA дозволяє перетворити корельовані ознаки у нові незалежні компоненти та зменшити надлишковість даних.

Перед побудовою моделей було виконано попередню обробку даних. Із набору видалено службові стовпці, які не несуть діагностичної інформації. Категоріальну цільову змінну було перетворено у числовий формат, де злоякісна пухлина позначалася як 1, а доброякісна — як 0. Оскільки ознаки мають різні масштаби вимірювання, було застосовано стандартизацію за допомогою StandardScaler[3]. Це дозволило привести всі параметри до єдиного масштабу та забезпечити коректну роботу алгоритмів машинного навчання.

У роботі реалізовано два підходи до класифікації. Перший підхід передбачав навчання логістичної регресії на повному наборі ознак[4]. Другий підхід використовував метод головних компонент для зниження розмірності даних із подальшим навчанням логістичної регресії у просторі головних компонент. Такий підхід дав змогу оцінити, чи можна скоротити кількість вхідних параметрів без погіршення якості класифікації.

Для визначення оптимальної кількості головних компонент було проведено експеримент із різними значеннями параметра k у діапазоні від 1 до 30. Для кожної моделі оцінювалися метрики Accuracy, Precision, Recall та F1-score. Особливу увагу було приділено Recall, оскільки в задачах медичної діагностики критично важливо мінімізувати кількість хибнонегативних результатів, коли злоякісний випадок може бути помилково класифікований як доброякісний.

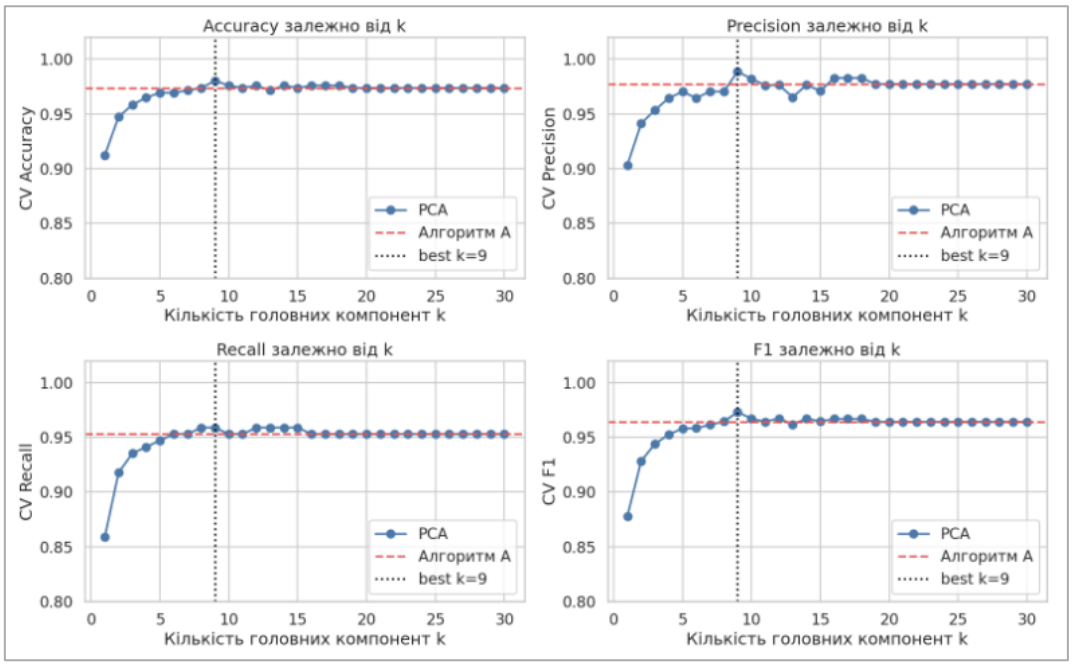


Рис. 3. Залежність метрик якості класифікації від кількості головних компонент k

На рисунку 3 показано, як змінюються основні метрики якості залежно від кількості використаних головних компонент. При малих значеннях k модель демонструє нижчу якість, оскільки значна частина інформації втрачається. Зі збільшенням кількості компонент показники поступово покращуються та стабілізуються. Це дозволяє визначити оптимальне значення k, при якому забезпечується баланс між скороченням розмірності та збереженням високої якості класифікації.

Експериментальні результати показали, що оптимальним є використання 9 головних компонент. Така модель зберігає понад 94% початкової дисперсії даних і забезпечує високу якість класифікації. При цьому розмірність даних зменшується з 30 ознак до 9 компонент, тобто приблизно у 3,3 раза. Це свідчить про те, що значна частина початкової інформації зосереджена у перших головних компонентах, а решта ознак містить переважно надлишкову або менш інформативну інформацію.

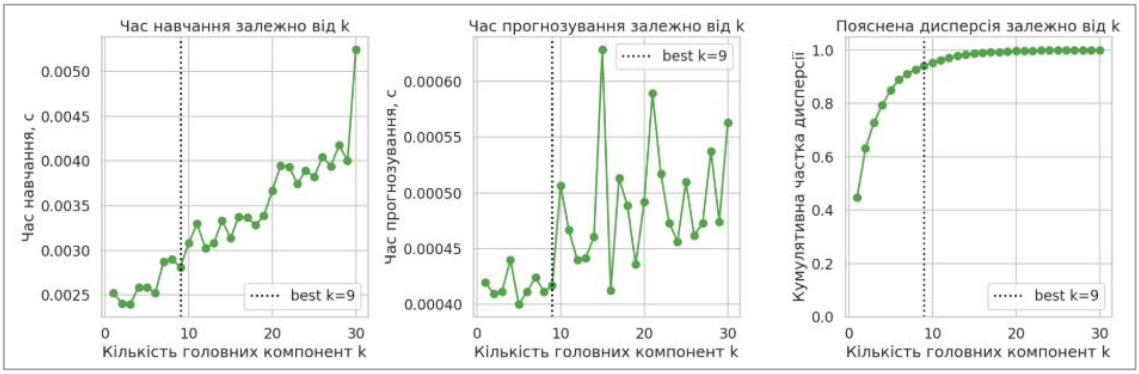
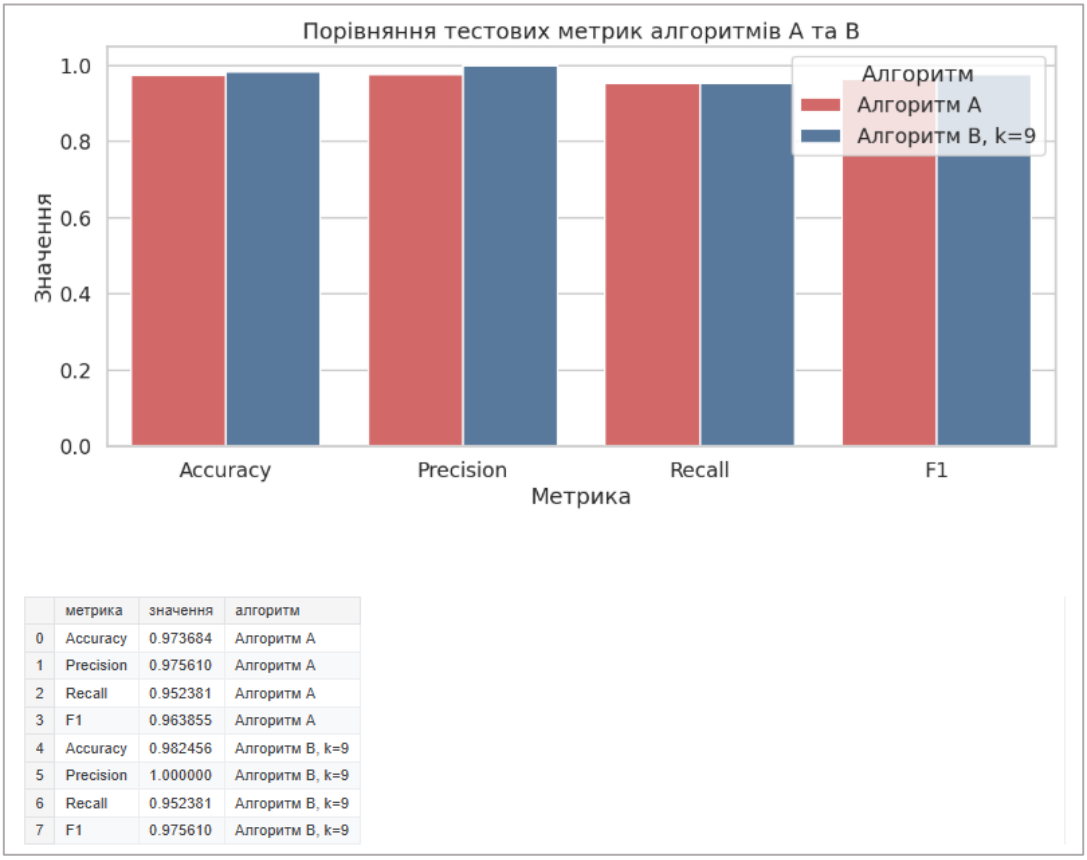


Рис. 4. Час навчання, час прогнозування та пояснена дисперсія залежно від кількості компонент k

Рисунок 4 демонструє залежність часу навчання, часу прогнозування та частки поясненої дисперсії від кількості головних компонент. Зі збільшенням кількості компонент зростає обсяг інформації, що зберігається моделлю, однак також може збільшуватися обчислювальна складність. Оптимальне значення k = 9 дозволяє зберегти понад 94% дисперсії та водночас суттєво зменшити кількість вхідних параметрів.

Для остаточного оцінювання ефективності запропонованого підходу було виконано порівняння базової моделі логістичної регресії, побудованої на всіх ознаках, із моделлю після застосування PCA. Порівняння показало, що модель у просторі головних компонент не поступається базовій моделі за

основними показниками якості. У деяких випадках вона демонструє більш стабільні результати, що пояснюється усуненням надлишкових і сильно корельованих ознак.

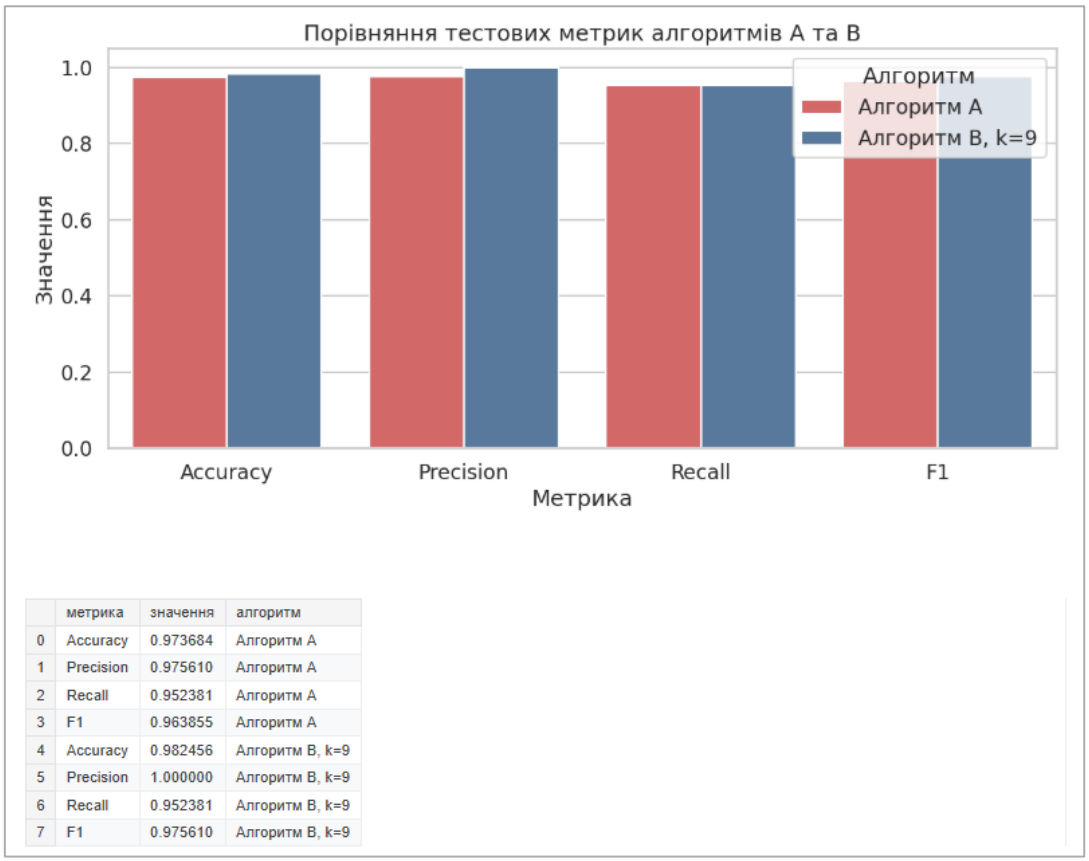


	метрика	значення	алгоритм
0	Accuracy	0.973684	Алгоритм A
1	Precision	0.975610	Алгоритм A
2	Recall	0.952381	Алгоритм A
3	F1	0.963855	Алгоритм A
4	Accuracy	0.982456	Алгоритм B, k=9
5	Precision	1.000000	Алгоритм B, k=9
6	Recall	0.952381	Алгоритм B, k=9
7	F1	0.975610	Алгоритм B, k=9

Рис. 5. Порівняння базової моделі та моделі з PCA при k = 9

На рисунку 5 наведено порівняння метрик якості для двох підходів: логістичної регресії на повному наборі ознак та логістичної регресії після застосування PCA. Отримані результати підтверджують, що зниження розмірності не призводить до погіршення класифікації, а навпаки може покращити узагальнювальну здатність моделі за рахунок зменшення мультиколінеарності.

Додатково було проаналізовано матриці помилок для базової моделі та PCA-моделі. Такий аналіз дозволив визначити кількість правильно та неправильно класифікованих об’єктів кожного класу. У медичних задачах це має особливе значення, оскільки помилки класифікації можуть мати різну вагу. Найбільш небезпечними є хибнонегативні результати, коли злоякісна пухлина визначається як доброякісна.



	метрика	значення	алгоритм
0	Accuracy	0.973684	Алгоритм A
1	Precision	0.975610	Алгоритм A
2	Recall	0.952381	Алгоритм A
3	F1	0.963855	Алгоритм A
4	Accuracy	0.982456	Алгоритм B, k=9
5	Precision	1.000000	Алгоритм B, k=9
6	Recall	0.952381	Алгоритм B, k=9
7	F1	0.975610	Алгоритм B, k=9

Рис. 6. Матриці помилок для базової моделі та PCA-моделі

Рисунок 6 відображає кількість правильних та помилкових класифікацій для кожного класу. Аналіз матриць помилок показує, що модель після застосування PCA зберігає високу здатність до правильного розпізнавання злоякісних випадків. Це є важливим показником адекватності моделі для задач медичної діагностики.

Отримані результати свідчать, що метод головних компонент є ефективним інструментом для обробки високорозмірних медичних даних. Він дозволяє усунути надлишковість, зменшити мультиколінеарність, скоротити обчислювальні витрати та зберегти високу діагностичну точність класифікаційної моделі.

Висновки

У роботі досліджено застосування методу головних компонент для зниження розмірності медичних даних у задачі бінарної класифікації пухлин молочної залози. Для експерименту використано набір Breast Cancer Wisconsin (Diagnostic), який містить 569 спостережень та 30 числових ознак, що описують морфологічні характеристики клітинних ядер. У процесі аналізу було встановлено наявність мультиколінеарності між окремими параметрами, зокрема між ознаками, пов’язаними з радіусом, периметром і площею клітин, що обґрунтувало доцільність застосування PCA.

Експериментально було порівняно базову логістичну регресію на всіх 30 ознаках та модель логістичної регресії після застосування PCA. Оптимальною виявилася модель з 9 головними компонентами, яка зберігає 94,04% дисперсії початкових даних і зменшує розмірність із 30 ознак до 9 компонент, тобто приблизно у 3,3 раза. На тестовій вибірці PCA-модель отримала Accuracy = 0,9825, Precision = 1,0000, Recall = 0,9524 та F1-score = 0,9756, тоді як базова модель мала Accuracy = 0,9737, Precision = 0,9756, Recall = 0,9524 та F1-score = 0,9639.

Порівняння результатів показало, що застосування PCA не погіршило здатність моделі виявляти злоякісні випадки, оскільки показник Recall залишився на рівні 0,9524. Водночас загальна точність зросла на 0,0088, Precision — на 0,0244, а F1-score — на 0,0118. Це свідчить про те, що перехід до

простору головних компонент дозволив зменшити кількість надлишкових ознак, зберегти критично важливу діагностичну інформацію та підвищити стабільність класифікації.

Отже, метод PCA є доцільним інструментом попередньої обробки високорозмірних медичних даних. Він дозволяє скоротити кількість вхідних параметрів, зменшити вплив мультиколінеарності та зберегти високу якість автоматизованої діагностики. Отримані результати підтверджують можливість використання такого підходу в медичних системах підтримки прийняття рішень, де важливими є не лише точність, а й надійність виявлення потенційно небезпечних випадків.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Breast Cancer Wisconsin (Diagnostic) Data Set. Kaggle [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.
2. Jolliffe I. T. Principal Component Analysis. Springer Series in Statistics. Springer, 2002.
3. Scikit-learn Documentation. Principal Component Analysis (PCA) [Електронний ресурс]. – Режим доступу: https://scikit-learn.org
4. Scikit-learn Documentation. Logistic Regression [Електронний ресурс]. – Режим доступу: https://scikit-learn.org
5. Pandas Documentation [Електронний ресурс]. – Режим доступу: https://pandas.pydata.org
6. Matplotlib Documentation [Електронний ресурс]. – Режим доступу: https://matplotlib.org
7. Seaborn Documentation [Електронний ресурс]. – Режим доступу: https://seaborn.pydata.org

Лавренюк Євгеній Михайлович – студент групи СА-23б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця. e-mail: lavrenyuk0629@gmail.com

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

Lavreniuk Yevhemii – student of Faculty of Intellectual Information Technologies and Automation, SA-23b, Vinnytsia National Technical University, Vinnytsia, e-mail lavrenyuk0629@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