ISSN 2707-1898 (print)
Український журнал інформаційних технологій
Ukrainian Journal of Information Technology
http://science.lpnu.ua/uk/ujit
https://doi.org/10.23939/ujit2024.01.048
Article received 24.04.2024 p.
Article accepted 30.04.2024 p.
UDK 004.8
Correspondence author
A. V. Doroshenko
anastasia.doroshenko@gmail.com

УДК 004.8

А. В. Дорошенко, Д. Ю. Савчук
Національний університет "Львівська політехніка", м. Львів, Україна

ДОСЛІДЖЕННЯ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ КЛАСИФІКАЦІЇ НЕЗБАЛАНСОВАНИХ НАБОРІВ ДАНИХ

Завдяки стрімкому розвитку інформаційних технологій, які широко використовуються у всіх сферах людського життя та діяльності, сьогодні накопичено надзвичайно великі обсяги даних. Відповідно застосування методів машинного навчання до цих даних дає змогу отримати нові практично корисні знання, які можуть бути використані для маркетингових, управлінських та дослідницьких цілей. Серед завдань інтелектуального аналізу даних – задачі регресії, прогнозування, кластеризації, класифікації та асоціативних правил. У цьому дослідженні розв'язано задачу бінарної класифікації. Основна мета роботи – дослідження різних методів машинного навчання для вирішення завдання класифікації та порівняння їхньої ефективності та точності. Окремим завданням є попереднє оброблення даних, спрямоване на вирішення проблеми незбалансованості вибірки, а також виявлення головних компонент, що використовуватимуться для вирішення завдання класифікації. Для цього досліджено та розроблено інформаційну систему класифікації банкрутства компанії із заданими економічними та фінансовими характеристиками. В дослідженні використано набір даних, на основі якого оцінено ефективність та якість застосування декількох відомих алгоритмів класифікації. Такими класифікаторами є: звичайний та лінійний Support Vector Machine, Extra Trees, Random Forest, Decision Tree, Logistic Regression, Multilayer perceptron Classifier, Gradient Boosting, Naive Bayes Classifier. Для передобробки даних здійснено масштабування, використано SMOTE-метод, щоб позбавитись незбалансованості навчальної вибірки, виконано виділення та аналіз головних компонент і L1 регуляризацію. Аналізування головних компонент дало змогу виявити 15 головних компонент, які найбільше впливають на точність класифікації і, відповідно, використовувати їх для класифікації. Аналізуючи отримані результати, ми встановили, що найкращим класифікатором був Random Forest з 95,9 %, а найгіршим Naive Bayes – 85,1 %. Для оцінювання якості класифікації та вибору найкращого класифікатора використано матрицю помилок (Confusion matrix), в якій враховується кількість істинно позитивних (TP) та істинно негативних значень (TN), а також розраховано кількість хибно негативних (FN) та хибно позитивних (FP) результатів класифікації. Наведено значення таких метрик, як точність, precision, чутливість, F1 та ROC. Точність – відсоток правильних відповідей алгоритму, чутливість (Recall) – це кількість TP, поділена на кількість TP плюс кількість FN. Показник F1 вказує на баланс між точністю та чутливістю. Precision – це кількість істинно позитивних прогнозів, поділена на кількість хибно позитивних та істинно негативних прогнозів. Оцінка ROC AUC – це інструмент вимірювання ефективності для задач класифікації за різних порогових значень, що показує, як модель може розрізняти класи. У висновках наведено найважливіші результати дослідження та вказано основний перспективний напрям розвитку роботи, а саме дослідження результатів класифікації для інших наборів даних та здійснення ефективніших оброблення та аналізу.

Ключові слова: інтелектуальний аналіз даних, класифікація, форматування, масштабування, аналіз, набір даних, вибірка даних, ознака, значення, порівняння, класифікатори.

Вступ / Introduction

Розв'язання задач інтелектуального аналізу даних (ІАД, Data Mining) залишається актуальним внаслідок постійного зростання обсягів даних у всіх сферах діяльності, серед яких бізнес, медицина, наука, інженерія, технології та багато інших [1]. Важливість цього підходу зумовлена здатністю виділяти цінну інформацію із великих обсягів даних, робити прогнози, виявляти закономірності та тренди, а також підтримувати процеси прийняття рішень на основі історичних даних [2].

Наприклад, у бізнесі Data Mining допомагає виявити патерни покупок споживачів, оптимізувати процеси виробництва, управляти запасами, а також прогнозувати попит на товари і послуги. У медицині його можна використати для аналізу клінічних даних пацієнтів, діагностики захворювань та вивчення ефективності методів лікування.

Завдяки постійному розвитку технологій, зокрема машинного навчання та штучного інтелекту, ІАД набуває нових можливостей, таких як автоматизація процесів аналізу даних, підвищення точності прогнозування та здатність створювати інсайти зі складніших наборів даних. Тому постає проблема розвитку та вдосконалення методів вирішення основних завдань Data Mining, зокрема пошуку асоціативних правил або закономірностей, групування об'єктів у певні кластери (кластерний аналіз), побудови регресійних моделей, класифікації об'єктів (до заздалегідь визначених класів) та регресійного аналізу.

У статті розглянуто вирішення завдання класифікації, що характеризується незбалансованою навчаль-

ною вибіркою. Такий тип задач надзвичайно поширений у медицині, якщо необхідно встановити діагноз для рідкісної хвороби (хворих може бути менше ніж 5 % від усіх пацієнтів), виявити шахрайство в електронній комерції тощо. Розглянемо різні підходи для врахування цієї особливості під час формування навчальної вибірки.

Також у роботі здійснено порівняльний аналіз ефективності та точності різних методів машинного навчання для вирішення завдання класифікації.

У загальному випадку існує декілька визначень класифікації, а саме:

- Завданням класифікації є певний розподіл досліджуваних об'єктів, явищ і процесів за родами, видами, типами або будь-якими іншими заданими ознаками з метою їх дослідження, групування цих понять і розташування їх у певній послідовності, що відображає ступінь подібності.
- З іншого боку, класифікація – це сукупність об'єктів зі схожими властивостями. Ці властивості відбирають для того, щоб визначити, наскільки схожі чи різні об'єкти.

Зокрема, зазначимо, що для здійснення класифікації математичними методами необхідно мати деякий загальний опис об'єкта класифікації. Найчастіше таким описом є база даних або набір даних, де кожен запис містить інформацію про певну властивість об'єкта.

Основне завдання роботи – дослідження ефективності методів машинного навчання для розв'язання задач класифікації підприємств, діяльність яких описується певними економічними та фінансовими показниками, на два відомі класи: підприємства, що є банкрутами, і ті, що функцонують успішно [3]. Особливістю задачі є те, що відсоток підприємств-банкрутів достатньо невеликий щодо загальної кількості підприємств, а отже, навчальна вибірка є незбалансованою [4].

Об'єкт дослідження – процес класифікації незбалансованої вибірки із використанням різних методів машинного навчання для здійснення інтелектуального аналізу даних.

Предмет дослідження – методи та засоби для класифікації щодо незбалансованої навчальної вибірки, дослідження та порівняння їх ефективності.

Мета роботи – дослідження різних методів машинного навчання для вирішення завдання класифікації та порівняння їх ефективності та точності. Окремим завданням є попереднє оброблення даних, спрямоване на вирішення проблеми незбалансованості вибірки, а також виявлення головних компонент, що використовуватимуться для вирішення завдання класифікації.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження*:

1) для задачі виявлення банкрутства підприємств, яка характеризується значною незбалансованістю екземплярів різних класів у навчальній вибірці, розробити методи та засоби передобробки даних, такі як: масштабування, подолання незбалансованості у тренувальній вибірці, виявлення та усунення кореляції між вхідними ознаками;
2) дослідити останні дослідження та публікації в цій сфері та визначити, які методи забезпечують найкращі результати для таких завдань;
3) розробити методи та засоби, що реалізують такі методи машинного навчання, як SVM, Extra Trees, Random Forest, Decision Tree, Logistic Regression, MLP класифікатор, Gradient Boosting, Naive Bayes Classifier, та застосувати їх для задачі виявлення банкрутства.

Аналіз останніх досліджень та публікацій. Задачу виявлення банкрутства компаній на основі фінансових показників досліджено у декількох статтях. До них належить робота "Фінансові коефіцієнти та індикатори корпоративного управління в прогнозуванні банкрутства: Комплексне дослідження" [4]. У статті згадано, що категорії платоспроможності та прибутковості, а також категорії структури правління та структури власності є найважливішими ознаками для прогнозування банкрутства. Зокрема, найкращих показників моделі прогнозування досягають, поєднуючи точність прогнозування, помилок I/II типу, ROC-кривої та вартість помилкової класифікації.

У статті "Прогнозування банкрутства за допомогою моделей машинного навчання з урахуванням комунікативної цінності річних звітів на основі тексту" також досліджено метод покращення результатів роботи класифікаторів на основі введення текстової інформації з річного звіту, проаналізовано прогнозувальну здатність чотирьох моделей машинного навчання (логістична регресія, випадковий ліс, XGBoost та SVM) [5].

Проблема незбалансованості даних в інтелектуальному аналізі даних – поширене явище, яке виникає через викривлену природу даних. У статті "Проблеми незбалансованості класів в інтелектуальному аналізі даних: Огляд" [6] здійснено комплексне дослідження для виявлення проблем, пов'язаних із виявленням та урахуванням незбалансованості класів під час класифікації за допомогою алгоритмів машинного навчання. Там розглянуто класифікатори, які підтримують упередженість до класу більшості та ігнорують клас меншості.

Розглядаючи проблему незбалансованості даних, автори статті "Підхід класифікатора випадкових лісів для незбалансованої класифікації великих даних для прикладних областей розумного міста" назвали класифікатор випадкових лісів (Random Forest) найкращим методом для оброблення непропорційних розподілів даних, із забезпеченням максимальної точності класифікації [7].

Проаналізовані статті дали змогу з'ясувати основні фінансові показники, що можуть впливати на ризик банкрутства компанії, зокрема, краще дослідити алгоритми, за допомогою яких можна поліпшити результати роботи досліджуваних класифікаторів та вказати основні методи, за допомогою яких можна вирішити проблему незбалансованості даних [8].

Результати дослідження та їх обговорення / Research results and their discussion

Набір даних для моделювання. Для виконання задачі класифікації використано набір даних [4], який містить економічні дані про діяльність різних компаній, а також інформацію про те, збанкрутували вони чи ні. Ці дані взято із Тайванського економічного журналу за 1999–2009 р. Відповідно, банкрутство компанії визначено на

підставі правил ведення бізнесу Тайванської фондової біржі. Зокрема, вибірка даних налічує 6819 прикладів. Кожен приклад має 95 атрибутів і один цільовий атрибут – клас, до якого належить компанія (банкрут чи ні). Опис перших десяти атрибутів наведено в табл. 1.

Табл. 1. Початковий набір даних / Initial Dataset

Набір даних	Характеристика набору даних
	Значення характеристики набору даних
X_1	Рентабельність активів до сплати відсотків та амортизація до сплати відсотків: рентабельність сукупних активів (C)
X_2	ROA (A) до сплати відсотків та відсоток після сплати податків: рентабельність всіх активів (A)
X_3	ROA (B) до сплати відсотків та амортизація після сплати податків: рентабельність сукупних активів (B)
X_4	Операційна валова маржа: валовий прибуток / чистий дохід
X_5	Валова маржа реалізованих продажів: реалізований валовий прибуток / чистий дохід
X_6	Норма операційного прибутку: операційний дохід / чистий продаж
X_7	Чиста відсоткова ставка до оподаткування: прибуток до оподаткування / чистий продаж
X_8	Чиста відсоткова ставка після оподаткування: чистий прибуток / чистий продаж
X_9	Негалузевий дохід і витрати/дохід: коефіцієнт чистого позаопераційного доходу
X_{10}	Безперервна відсоткова ставка (після оподаткування): чистий прибуток без прибутку або збитку від вибуття чистого продажу

У табл. 2 наведено перелік класів, до яких може належати один цільовий об’єкт.

Табл. 2. Класи цільового об’єкта / Objects classes of the target class

Y – значення	Заголовок стовпця таблиці	
	Клас 1	Клас 2
Значення	1.0	0.0
Опис	Компанія збанкрутувала	Компанія тримається на плаву

Аналізуючи навчальну вибірку, зауважимо, що більшість ознак у вибірці даних не підпорядковані нормальному розподілу, що видно з рис. 1.

Передоброблення даних. Щоб виконати якісну та ефективну класифікацію, спочатку нам потрібно попередньо опрацювати наш набір даних: необхідно прочитати дані з файла, перевірити їх на наявність пропусків і дослідити розмірність вибірки даних. Якщо комірка порожня, заповнимо її середнім значенням стовпчика.

Масштабування даних. Для того, щоб на вагу того чи іншого атрибута в прийнятті рішення не впливало абсолютне значення величини, доцільно масштабувати всі значення у навчальній та тестовій вибірках. Існує багато методів масштабування вибірки даних. Для цього завдання використано формулу

$$(\alpha - \beta)/\gamma = \chi\,, \qquad (1)$$

де α – старе значення об’єкта, χ – нове значення, якого набуде наш об’єкт, β – середнє значення навчальної вибірки даних, а γ – стандартне відхилення навчальної вибірки.

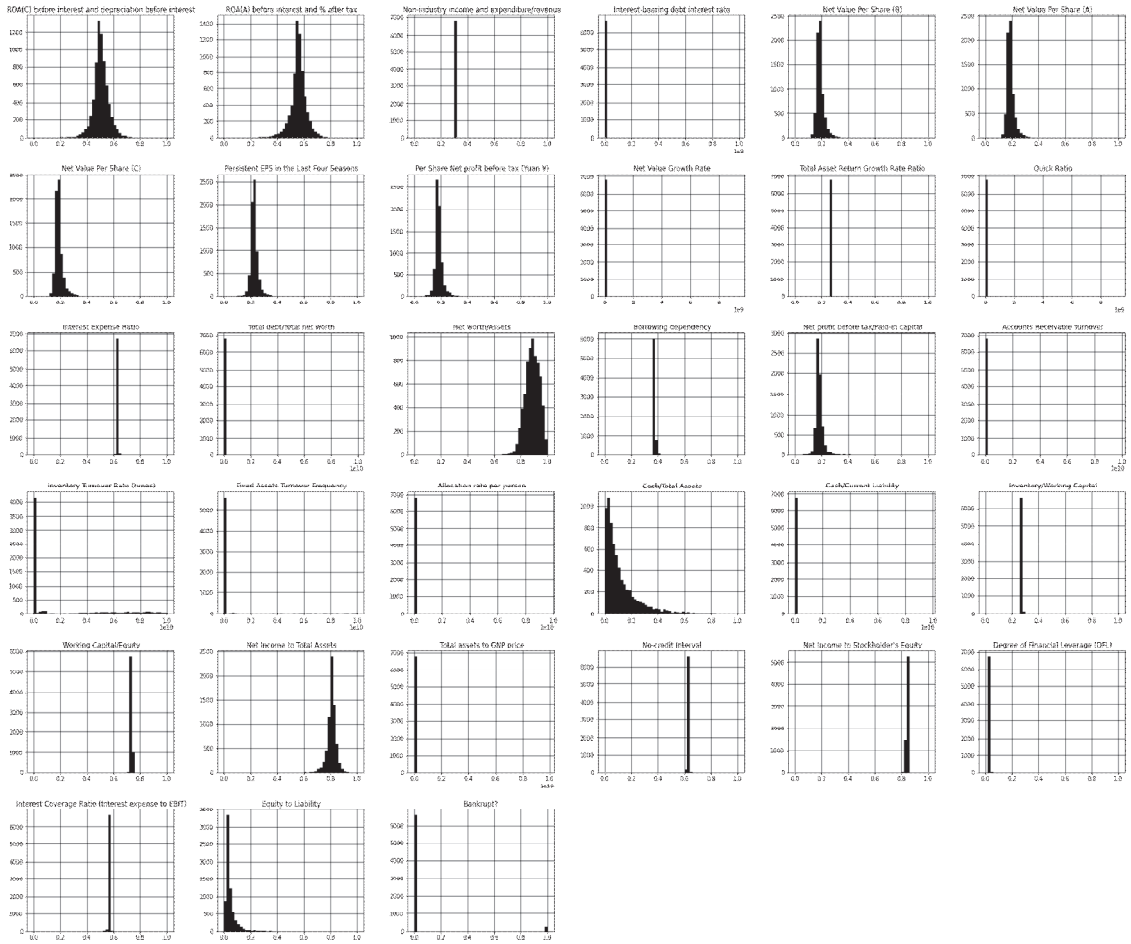

Рис. 1. Розподіл ознак у наборі даних / Distribution of features in the dataset

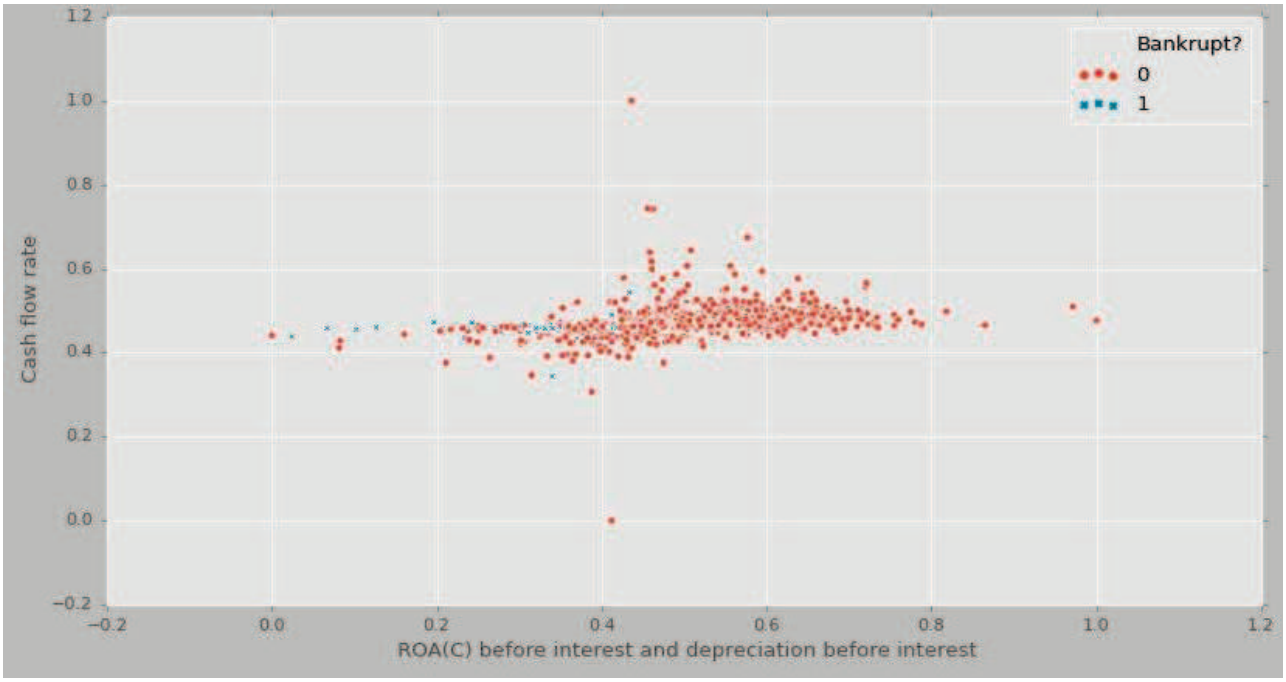

Рис. 2. Кількість об’єктів цільового класу / Number of objects of the target class

Збалансованість тренувальної вибірки. Наступний крок – перевірка даних на збалансованість. Це можна зробити, побудувавши гістограми, що відображають кількість об’єктів потрібного нам класу (рис. 2).

З рис. 2 видно, що дані незбалансовані. Для цього набору даних кількість об’єктів із нульовим значенням становить близько 7000, а об’єктів з одиничним значенням – не більше ніж 100.

Проблема незбалансованості навчальної вибірки критична для багатьох методів класифікації, оскільки менший клас може або не розпізнаватись взагалі, або класифікуватись із дуже низькою точністю. Тому сьогодні відомі різні підходи до вирішення цієї проблеми. Найпоширеніші з них можна поділити на три групи відповідно до того, як вони поводяться із дисбалансом класів: зовнішній підхід (підхід на рівні даних), внутрішній підхід (за якого проблема незбалансованості вирішується на алгоритмічному рівні) та підхід, чутливий до витрат. Також хороші результати показують класифікатори на основі ансамблевого навчання, які також відіграють істотну роль у класифікації незбалансованих даних [9].

У цій роботі використовуємо зовнішній підхід на рівні даних (попереднє оброблення).

Методи формування вибірки, використовувані для попереднього оброблення незбалансованих даних, які належать до цього підходу, можна розподілити на три типи:

1. Метод випадкової недостатньої вибірки, який довільно зменшує кількість зразків більшого класу і генерує підмножину первинного набору даних так, щоби збалансувати співвідношення. Це може призвести до втрати потенційних даних через їх видалення [10].
2. Метод випадкової надлишкової вибірки, за яким збільшується кількість зразків у меншому класі через випадкове повторення наявних зразків і генерується надмножина первинних даних. Але це може збільшити ймовірність перепідганяння через реплікацію.
3. Гібридна техніка, яка об’єднує обидва методи вибірки для збалансування співвідношення [11].

Для збалансування вибірки даних у дослідженні використано SMOTE (Synthetic Minority Over-Sampling Technique) [12]. Цю техніку надлишкової вибірки використовують, коли кількість даних недостатня. Суть цього методу полягає у синтезі нових прикладів для меншого класу на основі наявної вибірки даних за допомогою інтерполяції зразків меншого класу, розміщених поруч у просторі ознак. SMOTE випадково вибирає одного з *k*-найближчих сусідів (kNN) вибраних зразків і створює новий вектор, генеруючи значення за допомогою інтерполяції даних двох випадково вибраних зразків. Для меншого класу ми вирішили обмежити розподіл у просторі більшого класу. Цей метод дасть змогу уникнути проблеми надмірного підганяння, але водночас створює шумні та граничні зразки, які можуть спричиняти проблеми.

Для таких проблем, з якими стикається SMOTE, використовують деякі методи на основі фільтрації, щоб уникнути шуму в незбалансованих наборах даних (SMOTE-TL та SMOTE-EL). З іншого боку, для оброблення незбалансованих даних оригінальні методи вибірки також модифікують за допомогою методу найближчих сусідів – збалансованої упаковки (NBBag).

Також існує модифікована техніка передискретизації синтетичної меншості (MSMOTE), яка є вдосконаленою формою SMOTE. За допомогою розрахунку відстаней між усіма зразками в цьому алгоритмі менший клас поділено на три групи: із прихованим шумом, безпечні та граничні зразки. Він відхиляє приховані плями шуму на основі методу класифікації kNN, коли MSMOTE створює нові приклади. Однак він нічого не робить для випадків прихованого шуму, а також не визначає пріоритету важливих характеристик. Розширення SMOTE за допомогою ітераційно-розподільного фільтра використовують для оброблення шумів і регулювання меж класів. Модифікації SMOTE також можна використовувати для потужніших методів рівня даних, таких як розширення B1-SMOTE та B2-SMOTE для нормалізації незбалансованих даних.

В експерименті, розглянутому в статті, метод SMOTE застосовано в його базовому варіанті для того, щоб збалансувати вибірку: для цього, використовуючи вектори меншого класу, генеруємо SMOTE-методом додаткову кількість векторів, щоб кількість екземплярів меншого класу стала такою самою, як і більшого.

Відповідно, якщо початковий розподіл становив 6700 екземплярів більшого класу і 100 меншого, то після застосування SMOTE-методу вибірка стала збалансованою й обидва класи представлені однаковою кількістю векторів.

Кореляційний аналіз. Щоб перевірити, наскільки елементи вектора ознак, використані у вибірці, незалежні між собою, доцільно також побудувати кореляційну матрицю.

На рис. 3 наведено кореляційну матрицю для вказаної вибірки даних.

Аналізуючи кореляційну матрицю, можна зробити висновки, що деякі ознаки тісно взаємопов'язані. Тому наступним кроком є зменшення цієї залежності за допомогою техніки усунення ознак (Feature elimination technique); у нашому дослідженні ми використовували "L1 Regularization" [14, 15] та PCA (Principal Component Analysis) [16].

L1 Regularization. У пошуках механізму визначення надлишкових характеристик замість аналізу головних компонент ми дослідили метод, який називають "L1 Regularization".

Набір даних містить велику кількість незалежних змінних. Тому для полегшення класифікації та підвищення точності моделей машинного навчання необхідно зменшити розмірність цих даних. Для цього використано метод елімінації ознак під назвою "L1 Regularization". Результати наведено на рис. 4.

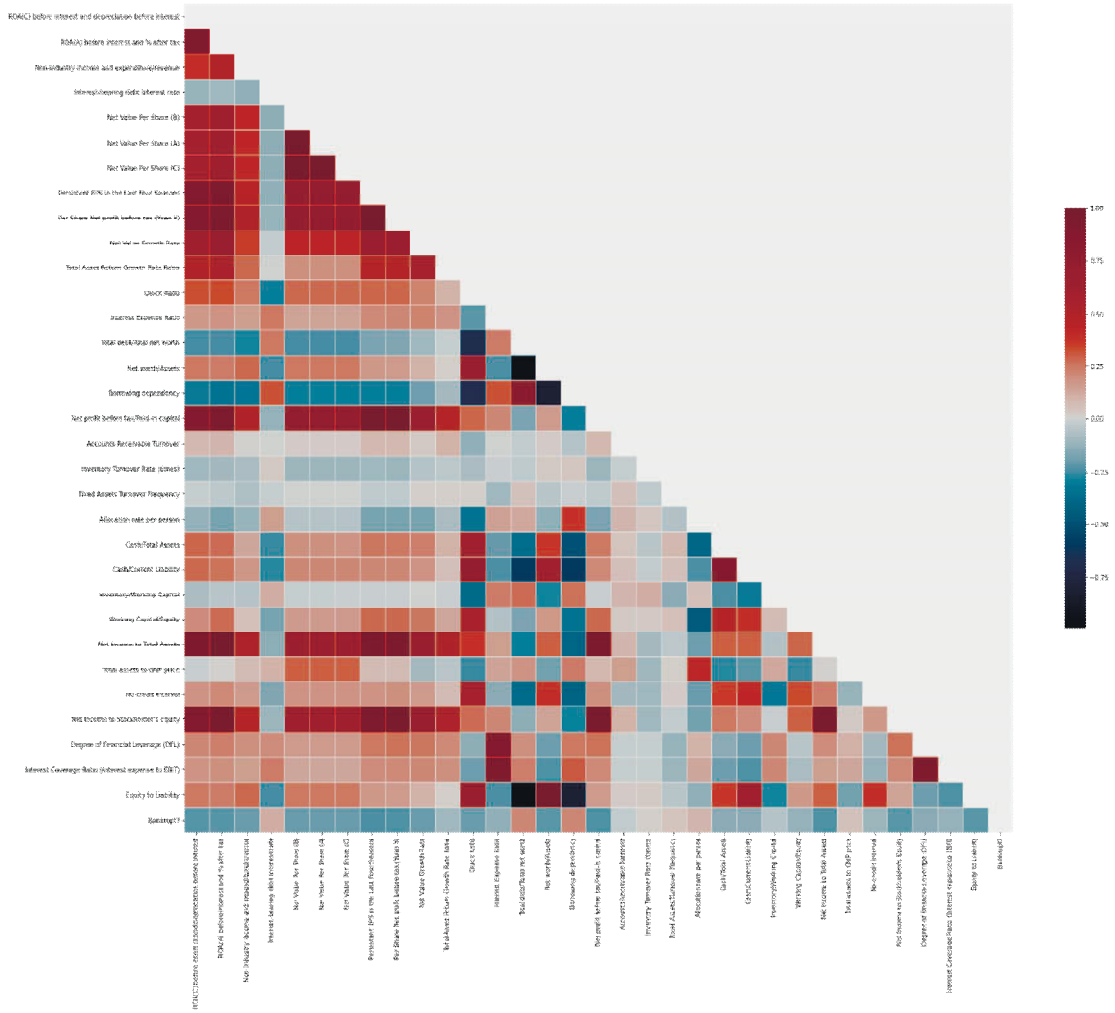

Рис. 3. Кореляційна матриця / Correlation matrix

```
total features: 95
selected features: 80
features with coefficients shrank to zero: 15
15
(10558, 80) (10558,)
(2640, 80) (2640,)
```

Рис. 4. Результати регуляризації L1 / Results of L1 Regularization

Видно, що після “L1 Regularization” кількість ознак, які потрібно вилучити, становить 15. Це порівняно невелике значення, тому ми вирішили використати іншу техніку вилучення ознак.

Аналіз головних компонент (PCA). Як можна зрозуміти, вибірка даних містить багато незалежних ознак. Тому необхідно позбутися тих, які не впливають на результати класифікації.

Для того, щоб з’ясувати, скільки ознак потрібно залишити, необхідно побудувати Scree діаграму (рис. 5.). У багатовимірній статистиці осипний графік – це лінійний графік власних значень факторів або головних компонентів у аналізі. Графік осипу використовують для визначення кількості факторів, які необхідно зберегти в дослідному факторному аналізі (FA), або основних компонентів, які треба зберегти в аналізі основних компонентів (PCA) [17]. Процедура пошуку статистично значущих факторів або компонентів із використанням осипу також відома як тест на осип. Раймонд Б. Кеттелл представив графік на осипах у 1966 р.

Осиповий графік завжди відображає власні значення у вигляді низхідної кривої, впорядковуючи власні значення від найбільшого до найменшого. Відповідно до тесту на осип, виявляється “вигин” графіка, де власні значення згладжені, а фактори або компоненти зліва від цієї точки повинні залишатися значущими.

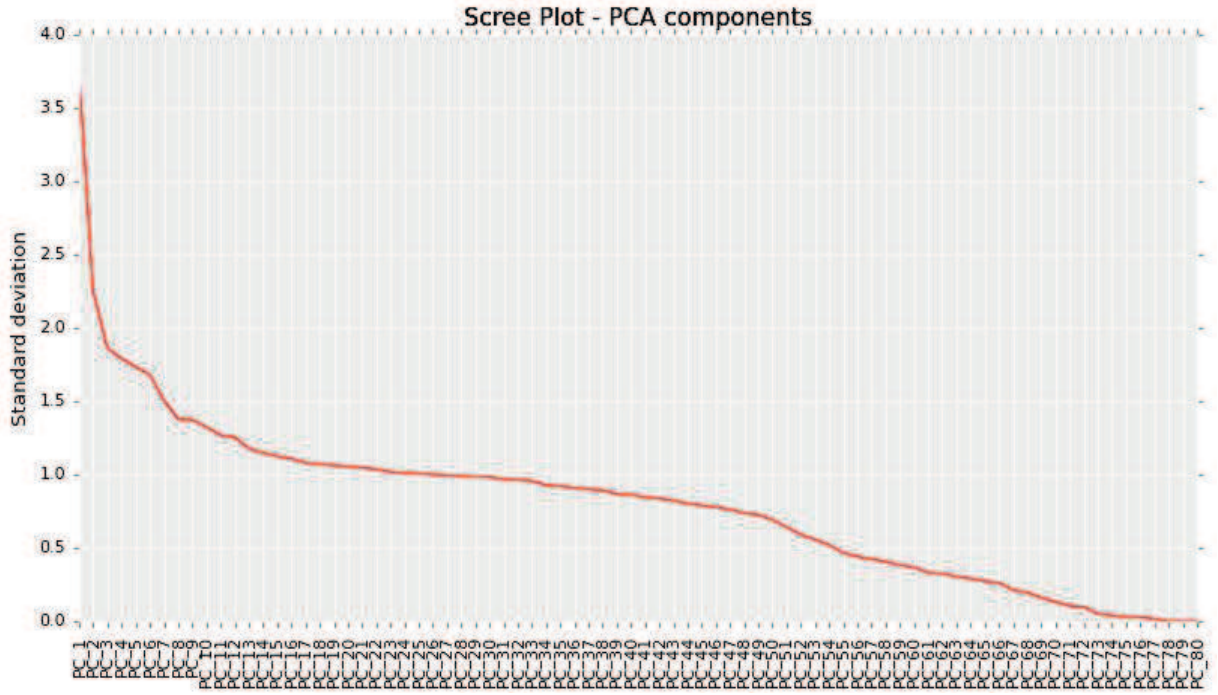

Рис. 5. Осипова діаграма основних компонентів / Scree Plot for PCA Components

Зважаючи на викладене вище, ми побудували діаграму (рис. 5.), де Y відповідає за середньоквадратичне відхилення, а X – за кількість компонент (ознак), використовуваних для розрахунку середньоквадратичного відхилення [18].

Для вибору кількості головних компонент застосовують метод ліктя. Зрозуміло, що крива є порівняно прямою, коли розміщена на PC_15, тому для подальшого аналізу потрібно взяти лише 15 головних компонент.

Аналіз викидів. Наступний крок – перевірка набору даних на наявність пропусків або аномальних спостережень. Це важливий крок в аналізі даних, оскільки він дає змогу вилучити помилкові або неточні спостереження, які спотворюють висновки [19].

Для аналізу викидів застосовують широкий спектр методів та інструментів [20], [21], [22]. Для розв’язання цієї задачі ми використали коробкову діаграму (рис. 6).

Побудова моделей та класифікація. Наявні класифікатори. В роботі розглянуто та використано різні типи класифікаторів [23], [24], [25], які наведено в табл. 3.

Табл. 3. Доступні класифікатори / Available Classifiers

№	Назва класифікатора
1	Logistic Regression
2	Decision Tree
3	Exrta Trees
4	Random Forest
5	SVC
6	Linear SVC
7	MLP Classifier
8	Gradient Boosting
9	Naïve Bayes

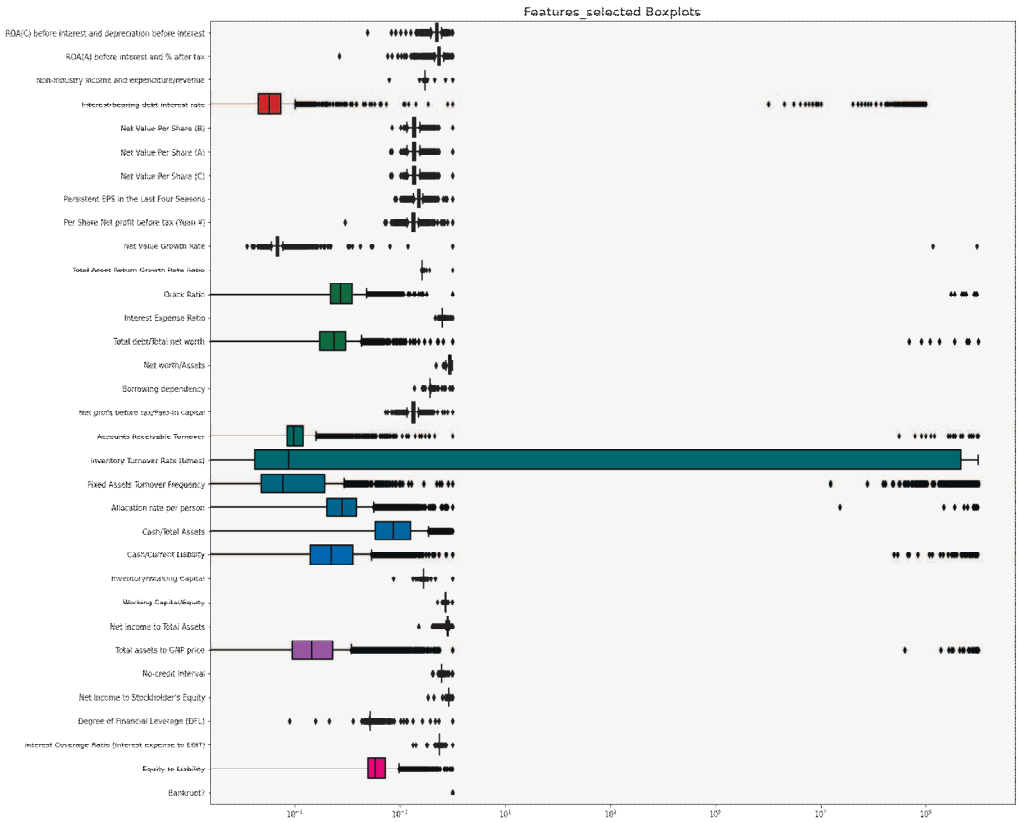

Рис. 6. Коробкова діаграма з викидами / Boxplot with outliers

Побудова моделі. Вибірку даних розділено на тестову та навчальну у співвідношенні 1/5. Параметри для класифікаторів вибрано за замовчуванням.

Як видно з рис. 7, стандартний вигляд результатів класифікації для будь-якого з класифікаторів такий:

- назва класифікатора;
- матриця помилок;
- accuracy;
- precision;
- recall;
- F1 – оцінка;
- support;
- macro average;
- weighted average.

Після здійснення деякої кількості ітерацій системи ми знайшли середні результати точності класифікаторів, а потім ці результати згрупували в таблицю (рис. 8).

```
Classification: MLPClassifier

Confusion Matrix
 [[1247   22]
 [  72 1299]]

Accuracy Score
 0.9643939393939394

Classification Report
              precision    recall  f1-score   support

           0       0.95      0.98      0.96      1269
           1       0.98      0.95      0.97      1371

    accuracy                           0.96      2640
   macro avg       0.96      0.97      0.96      2640
weighted avg       0.97      0.96      0.96      2640
```

Рис. 7. Результати класифікації / Classification Results

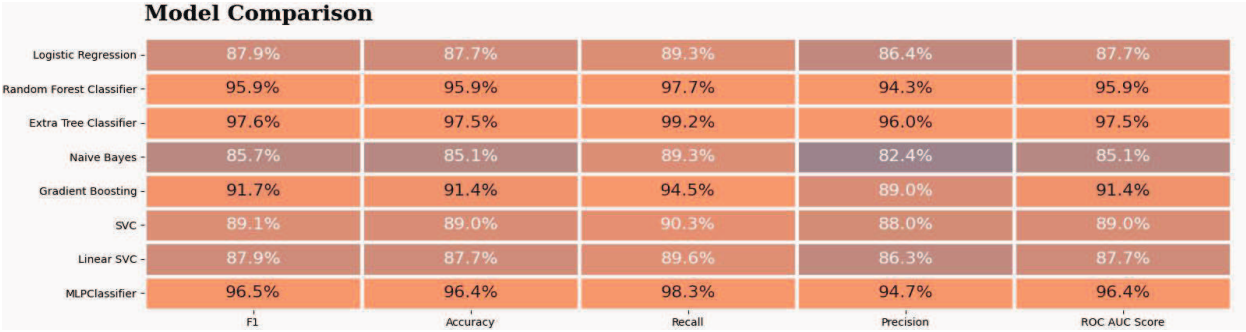

Рис. 8. Результати класифікації / The results of models comparison

Табл. 4. Приклад матриці помилок / Confusion Matrix

№	Цільове значення	
	$y = 1$	$y = 0$
$y^* = 1$	Істинно позитивні (TP)	Хибно позитивні (FP)
$y^* = 0$	Хибно негативні (FN)	Істинно негативні (TN)

де y^* – результат роботи алгоритму на об'єкті; y – істинна мітка класу на цьому об'єкті.

Обговорення результатів дослідження. Перш ніж аналізувати результати класифікації, необхідно встановити визначення матриці помилок (Confusion matrix example) за рис. 7. Її вигляд для досліджуваного набору подано у табл. 4.

Отже, помилки класифікації бувають двох типів: хибно негативні (FN) та хибно позитивні (FP).

Основні метрики, які оцінюють та порівнюють [10]:

- **F1** також називають F-оцінкою або F-мірою. Вимірюється за (2).

$$F1 = 2 * \frac{accuracy * recall}{accuracy + recall}, \quad (2)$$

інакше кажучи, показник F1 вказує на баланс між accuracy та recall.

- **Precision** – це кількість істинно позитивних прогнозів, поділена на кількість хибно позитивних та істинно негативних прогнозів. Тобто це кількість позитивних передбачень, поділена на загальну кількість передбачених позитивних значень класу (3). Його також називають позитивною прогностичною цінністю (PPV):

$$PPV = \frac{TP}{FP + TN}, \quad (3)$$

- **Accuracy** – відсоток правильних відповідей алгоритму, розраховується за (4):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (4)$$

- **Recall** – це кількість TP, поділена на кількість TP плюс кількість FN (5). Його також називають чутливістю, або реальним коефіцієнтом TP:

$$\mathrm{Re}call = \frac{TP}{TP + FN}. \quad (5)$$

- Оцінка **ROC AUC** – це інструмент вимірювання ефективності для задач класифікації за різних порогових значень. Він показує, як модель може розрізняти класи. Що вище AUC, то краще модель приймає 0 за 0 і 1 за 1.

Під час роботи досліджено методи інтелектуального аналізу даних на підставі незбалансованих даних про банкрутство компаній для вирішення проблеми класифікації. З усіх восьми класифікаторів найточнішими виявились моделі із використанням Random Forest Classifier – 95,9 %, Extra Tree Classifier – 97,5 %, та MLP Classifier – 96,4 %, а найнижча точність у Naive Bayes – 85,1 % та Linear SVC – 87,7 %. Отже, найкращими моделями для вирішення поставленого завдання є класифікатори на основі ієрархічних дерев та нейронних мереж.

Зокрема, зазначимо, що інші метрики, такі як F1, recall, precision, ROC AUC, також показали хороші результати, тобто можна вважати доцільним використання заданих класифікаторів для дослідження проблеми банкрутства компаній.

Відповідно, сформулюємо наукову новизну та практичну значущість одержаних результатів.

Наукова новизна отриманих результатів дослідження – полягає у комбінації різних методів передоброблення даних для підвищення точності класифікації для задачі із істотною незбалансованістю даних. Досліджено також різні типи моделей інтелектуального аналізу даних для вирішення проблеми класифікації із використанням незбалансованих даних.

Практична значущість результатів дослідження – можливість використати результати дослідження для того, щоб вибрати необхідну модель для вирішення проблеми класифікації із використанням незбалансованих даних.

Висновки / Conclusions

У роботі досліджено особливості вирішення завдання класифікації для незбалансованих наборів даних, які дуже поширені у медицині, електронній комерції, наукових дослідженнях. Описано рекомендовані підходи до роботи із даними на всіх етапах життєвого циклу: від передоброблення й масштабування до вирівнювання вибірок та формування навчальних і тестових наборів даних.

Досліджено ефективність застосування таких методів машинного навчання, як SVM, Extra Trees, Ran-

dom Forest, Decision Tree, Logistic Regression, MLP класифікатор, Gradient Boosting, Naive Bayes Classifier та використання їх для завдання виявлення банкрутства. З усіх восьми класифікаторів найвищою виявилась точність моделей із використанням Random Forest Classifier – 95,9 %, Extra Tree Classifier – 97,5 %, та MLP Classifier – 96,4 %, а найнижчою – Naive Bayes – 85,1 % та Linear SVC – 87,7 %. Отже, найкращими моделями для вирішення поставленого завдання є класифікатори на основі ієрархічних дерев та нейронних мереж.

Також розглянуто процес вибору метрик для алгоритмів, які допоможуть приблизно прогнозувати час, необхідний для класифікації даних на практиці, та точність, якої можна досягти.

Однією із головних перспектив є дослідження результатів класифікації для інших наборів даних та здійснення їх ефективного оброблення та аналізу залежно від ситуації. Завдяки цьому можна буде створити систему, яка оброблятиме, аналізуватиме та класифікуватиме будь-які набори даних.

Подяки / Acknowledgments

Цю роботу виконано в межах програми Erasmus+ Jean Monnet Module "Trustworthy artificial intelligence: the European approach" (101085626 – TrustAI – ERASMUS- JMO-2022-HEI-TCH-RSCH).

References

1. Teslyuk, V., Doroshenko, A., & Savchuk, D. (2023). Intelligent Methods and Models for Assessing Level of Student Adaptation to Online Learning, 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Kharkiv, Ukraine. CEUR Workshop Proceedings, 3387, 331-343.
2. Akhavan, F., & Hassannayebi, E. (2024). A hybrid machine learning with process analytics for predicting customer experience in online insurance services industry. *Decision Analytics Journal, 11*, art. no. 100452. https://doi.org/10.1016/j.dajour.2024.100452
3. Guha, A., & Veeranjaneyulu, N. (2019). Prediction of bankruptcy using big data analytic based on fuzzy C-means algorithm. *IAES International Journal of Artificial Intelligence, 8*(2), 168–174. https://doi.org/10.11591/ijai.v8.i2.pp168-174
4. Liang, D., Lu, C.-C., Tsai, C.-F., & Shih, G.-A. (2016). Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A Comprehensive Study. *European Journal of Operational Research, 252*(2), 561–572. https://doi.org/10.1016/j.ejor.2016.01.012
5. Chen, T.-K., Liao, H.-H., Chen, G.-D., Kang, W.-H., & Lin, Y.-C. (2023). Bankruptcy Prediction Using Machine Learning Models with the Text-based Communicative Value of Annual Reports. *Expert Systems with Applications, 120714*. https://doi.org/10.1016/j.eswa.2023.120714
6. Ali, H., Mohd Salleh, M. N., Saedudin, R., Hussain, K., & Mushtaq, M. F. (2019). Imbalance class problems in data mining: a review. *Indonesian Journal of Electrical Engineering and Computer Science, 14*(3), 1552. https://doi.org/10.11591/ijeecs.v14.i3.pp1552-1563
7. More, S., & Rana, Anjali and P. (2018). Dipti and Agarwal, Isha, Random Forest Classifier Approach for Imbalanced Big Data Classification for Smart City Application Domains. I*nternational Journal of Computational Intelligence & IoT, 1*(2). Retrieved from: https://ssrn.com/abstract=3354727
8. Santos, M. S., Abreu, P. H., Japkowicz, N. et al. (2022). On the joint-effect of class imbalance and overlap: a critical review. *Artif Intell Rev, 55*, 6207-6275. https://doi.org/10.1007/s10462-022-10150-3
9. Doroshenko, A. & Tkachenko, R. (2018). Classification of Imbalanced Classes Using the Committee of Neural Networks. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), 400–403. https://doi.org/10.1109/STC-CSIT.2018.8526611
10. Basha, S. J., Madala, S. R., Vivek, K., Kumar, E. S., & Ammannamma, T. (2022). A Review on Imbalanced Data Classification Techniques. 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA), Coimbatore, India, 1–6. https://doi.org/10.1109/ICACTA54488.2022.9753392
11. Zhongqiang, Sun, Wenhao, Ying, Wenjin, Zhang, & Shengrong, Gong (2024). Undersampling method based on minority class density for imbalanced data. *Expert Systems with Applications, 249*(Part A), 123328. https://doi.org/10.1016/j.eswa.2024.123328
12. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357. https://doi.org/10.1613/jair.953
13. Srividya, Mohanavalli, S., Sripriya, N., & Poornima, S. (2018). Outlier Detection using Clustering Techniques. *International Journal of Engineering & Technology, 7*(3.12), 813. https://doi.org/10.14419/ijet.v7i3.12.16508
14. Regularization path of L1- Logistic Regression. (б. д.). scikit-learn. https://scikit-learn.org/stable/auto_examples/linear_model/plot_logistic_path.html
15. Pan, H., Badawi, D., Bassi, I., Ozev, S. & Cetin, A. E. (2022). Detecting Anomaly in Chemical Sensors via L1-Kernel-Based Principal Component Analysis. *IEEE Sensors Letters, 6*(10), art no. 7004304, 1–4. https://doi.org/10.1109/LSENS.2022.3209102
16. Soomro, G. M., Krayem, S., Amur, Z. H., Chramcov, B., Jasek, R., & Noordin, I. (2023). Tumor Detection of Breast Tissue Using Random Forest with Principal Component Analysis. IEEE 8th International Conference on Engineering Technologies and Applied Sciences (ICETAS), Bahrain, Bahrain, 1–7. https://doi.org/10.1109/ICETAS59148.2023.10346582
17. Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). *Computers & Geosciences, 19*(3), 303–342. https://doi.org/10.1016/0098-3004(93)90090-r
18. Doroshenko, Anastasiya (2019). Application of global optimization methods to increase the accuracy of classification in the data mining tasks. In: Luengo D., Subbotin S. (Eds.): Computer Modeling and Intelligent Systems. Proc. 2-nd Int. Conf. CMIS-2019, Vol-2353: Main Conference Zaporizhzhia, Ukraine, April 15–19, 98–109. https://doi.org/10.32782/cmis/2353-8
19. Jadhav, T. et al. (2023). Predicting Urban Land Cover Using Classification: A Machine Learning Approach. IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC), Rajkot, India, 450–454. https://doi.org/10.1109/R10-HTC57504.2023.10461930
20. Savchuk, D. & Doroshenko, A. (2021). Investigation of machine learning classification methods effectiveness. IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 33–37. https://doi.org/10.1109/CSIT52700.2021.9648582
21. Ahmed, T., Paul, R. R., Alam, M. A., Hasan, M. T., & Rab, M. R. (2022). Performance Comparison of Different Machine Learning Classifiers in Categorizing Bangla News Articles. 4th International Conference on Natural Language Processing (ICNLP), Xi'an, China, 376–379. https://doi.org/10.1109/ICNLP55136.2022.00069

22. Tanouz, D., Subramanian, R. Raja, Eswar, D., Parameswara Reddy, G. V., Ranjith Kumar, A., Praneeth, CH. V. N. M. (2021). Credit Card Fraud Detection Using Machine Learning. 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 967–972. https://doi.org/10.1109/ICICCS51141.2021.9432308
23. Izonin, I., Tkachenko, R., Pidkostelnyi, R., Pavliuk, O., Khavalko, V., Batyuk, A. (2021). Experimental evaluation of the effectiveness of ann-based numerical data augmentation methods for diagnostics tasks CEUR Workshop Proceedings, 3038, 223-232.
24. Md. Shojeb Hossain Shojol, Md. Abu Ismail Siddique, Fariha Haque (2023). Enhanced Convolutional Neural Networks for Early Detection and Classification of Ophthalmic Diseases. International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Dhaka, Bangladesh, 2023, 209–213. https://doi.org/10.1109/ICICT4SD59951.2023.10303558
25. Singh, A. K. (2022). Detection of Credit Card Fraud using Machine Learning Algorithms. 11th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2022, 673–677. https://doi.org/10.1109/SMART55829.2022.10047099
26. Subbotin, S., Tabunshchyk, G., Arras, P., Tabunshchyk, D., & Trotsenko, E. (2021). Intelligent Data Analysis for Individual Hypertensia Patient's State Monitoring and Prediction. IEEE International Conference on Smart Information Systems and Technologies (SIST), Nur-Sultan, Kazakhstan, 2021, 1–4. https://doi.org/10.1109/SIST50301.2021.9465989

A. V. Doroshenko, D. Yu. Savchuk

Lviv Polytechnic National University, Lviv, Ukraine

RESEARCH OF DATA MINING METHODS FOR CLASSIFICATION OF IMBALANCED DATA SETS

With the rapid development of information technology, which is widely used in all spheres of human life and activity, extremely large amounts of data have been accumulated today. By applying machine learning methods to this data, new practically useful knowledge can be obtained. The main goal of this paper is to study different machine learning methods for solving the classification problem and compare their efficiency and accuracy. A separate task is data pre-processing aimed at solving the problem of sample imbalance, as well as identifying the principal components that will be used to solve the classification problem. For this purpose, an information system for classifying the bankruptcy of a company with specified economic and financial characteristics was researched and developed. The study uses a dataset on the basis of which the efficiency and quality of application of several existing classification algorithms are evaluated. These classifiers are: conventional and linear Support Vector Machine, Extra Trees, Random Forest, Decision Tree, Logistic Regression, Multilayer perceptron Classifier, Gradient Boosting, Naive Bayes Classifier. For data pre-processing, we scaled the data, used the SMOTE method to get rid of the imbalance of the training sample, and performed principal component analysis and L1 regularisation. Principal component analysis allowed us to identify 15 principal components that have the greatest impact on classification accuracy and, accordingly, use them in the classification process. Analysing the results, we found that the best classifier was Random Forest with 95.9 % accuracy, and the worst was Naive Bayes with 85.1 %. To evaluate the quality of classification and select the best classifier, the Confusion matrix is used, which takes into account the number of true positive (TP) and true negative (TN) values, as well as the number of false negative (FN) and false positive (FP) classification results, and the values of such metrics as accuracy, precision, sensitivity, F1, and ROC. Accuracy is the percentage of correct answers given by the algorithm, while Recall is the number of TPs divided by the number of TPs plus the number of FNs. F1 indicates the balance between accuracy and sensitivity. Precision is the number of true positive predictions divided by the number of false positive and true negative predictions. ROC AUC is a tool for measuring performance for classification tasks at different thresholds. It shows how well a model can distinguish between classes. The conclusions present the main results of the study and indicate the main future direction of the work, namely, the study of classification results for other datasets and more efficient processing and analysis.

Keywords: data mining, classification, formatting, scaling, analysis, dataset, data sample, feature, value, comparing, classifiers.

Інформація про авторів:

Дорошенко Анастасія Володимирівна, канд. техн. наук, доцент, кафедра автоматизованих систем управління. **Email**: anaastasiia.v.doroshenko@lpnu.ua; https://orcid.org/0000-0002-7214-5108

Савчук Дмитро Юрійович, аспірант, кафедра автоматизованих систем управління. **Email**: dmytro.y.savchuk@lpnu.ua; https://orcid.org/0009-0006-0937-6161

Цитування за ДСТУ: Дорошенко А. В., Савчук Д. Ю. Дослідження методів інтелектуального аналізу даних для класифікації незбалансованих наборів даних. *Український журнал інформаційних технологій*. 2024, т. 6, № 1. С. 48–57.

Citation APA: Doroshenko, A. V., & Savchuk, D. Yu. (2024). Research of data mining methods for classification of imbalanced data sets. *Ukrainian Journal of Information Technology, 6*(1), 48–57. https://doi.org/10.23939/ujit2024.01.048