Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
ISSN 1994-7836 (print)
ISSN 2519-2477 (online)
https://doi.org/10.36930/40340516
Article received 21.05.2024
Article accepted 23.05.2024
UDC 004.[8+048+622]
Correspondence author: V. V. Petryna volodymyr.v.petryna@lpnu.ua

В. В. Петрина, А. В. Дорошенко
Національний університет "Львівська політехніка", м. Львів, Україна

ЕФЕКТИВНІСТЬ ЗАСТОСУВАННЯ МЕТОДІВ КЛАСИФІКАЦІЇ ДЛЯ ЗАДАЧ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ВЕЛИКИХ ДАНИХ

Проаналізовано ефективність застосування методів класифікації для задач інтелектуального аналізу великих даних на підставі концепції машинного навчання задля підвищення їхньої ефективності у сфері електронної комерції. Проведено порівняльний аналіз застосування таких моделей, як класифікатор методом випадкового лісу (англ. *Random Forest*), класифікатор методом наївного Байєса (англ. *Naïve Bayes*) та класифікатор методом опорних векторів (англ. *Support Vector Machines*, SVM), який також називають опорно-векторними мережами (англ. *Support Vector Networks*, SVN). Для поширеної у сфері електронної комерції задачі класифікації клієнтів розроблено програмне забезпечення для проведення аналізу відповідних алгоритмів. Проаналізовано вхідні дані і здійснено попередню підготовку даних для навчання та тестування вибраних моделей. Здійснено дослідження обраних моделей із використанням попередньо підготовлених даних за допомогою програмного забезпечення відповідно до визначених сценаріїв. Досліджено параметри обраних моделей класифікації та вдосконалено класифікатор методом випадкового лісу шляхом підбору та зміни параметра випадкового стану. Також впроваджено параметри підтримки ймовірностей у класифікаторі методом опорних векторів. Здійснено із використанням попередньо підготовлених даних дослідження обраних моделей за допомогою програмного забезпечення відповідно до визначених сценаріїв. Впроваджено параметру підтримки ймовірностей у класифікаторі методом опорних векторів. Здійснено порівняння результату точності класифікації обраних моделей класифікації. Згідно з результатами дослідження, визначено позитивний тренд на якість навчання моделей за коректної підготовки даних і впливу підбору коректних параметрів для класифікаторів методами випадкового лісу й опорних векторів. Показники ефективності, точності навчання алгоритму показують позитивну динаміку й порівняно із результатами тестування моделі класифікатора методом наївного Байєса базовими значеннями параметрів моделі. На підставі результатів дослідження підтверджується вплив підбору коректних параметрів залежно від вхідного набору даних на результати точності передбачення алгоритмів і їх вплив на навчання, тренування та тестування моделей машинного навчання. Ці результати свідчать про перспективи до подальшого дослідження щодо розроблення оптимальних стратегій оптимізації та підвищення ефективності щодо роботи з алгоритмами машинного навчання у задачах класифікації.

Ключові слова: електронна комерція; великі дані; інтелектуальний аналіз даних; класифікація; випадковий ліс; метод опорних векторів; класифікатор методом наївного Байєса.

Вступ / Introduction

Пришвидшена цифровізація, яка є основою глибокої трансформації бізнесу на всіх рівнях, призводить до того, що дедалі більше людей продають і купують продукти та послуги в мережі Інтернет, а деякі покупки неможливо зробити без інфраструктури мережі Інтернет [26]. Відповідно, маркетингові стратегії постійно переносяться у сферу інтернету, впроваджуючи новітні технології оброблення та аналізу інформації. Оскільки клієнти, які купують онлайн, фактично не можуть побачити, понюхати, спробувати, почути або доторкнутися до товарів, реалізація бізнес-механізмів відбувається непрямим шляхом. Саме тому використання методів оброблення великих даних для вивчення принципів реалізації електронної комерції є актуальним в умовах сьогодення.

Сьогодні галузь електронної комерції зростає шаленими темпами, щоб вирішувати бізнес-проблеми, і всі бізнес-проблеми можна легко вирішити за допомогою великих даних. Враховуючи обсяги даних і швидке зростання компаній електронної комерції, використання великих даних змінює все, включно з ринками електронної комерції. Усе це завдяки мережі Інтернет, соціальним мережам та іншим новим технологіям, які збирають дані та забезпечують значну економію коштів, значні покращення та послуги для виконання обчислювальних завдань, які обробляють різні типи даних легко та без ускладнень. Оскільки канал електронної комерції розширюється, великі дані стають важливим інструментом.

Об'єкт дослідження – процеси збирання, збереження, оброблення та інтелектуального аналізу великих даних в електронній комерції.

Предмет дослідження – методи та моделі збережен-

Інформація про авторів:
Петрина Володимир Васильович, аспірант, кафедра автоматизованих систем управління. **Email:** volodymyr.v.petryna@lpnu.ua
Дорошенко Анастасія Володимирівна, канд. техн. наук, доцент, кафедра автоматизованих систем управління.
Email: anaastasiia.v.doroshenko@lpnu.ua; https://orcid.org/0000-0002-7214-5108

Цитування за ДСТУ: Петрина В. В., Дорошенко А. В. Ефективність застосування методів класифікації для задач інтелектуального аналізу великих даних. Науковий вісник НЛТУ України. 2024, т. 34, № 5. С. 119–128.
Citation APA: Petryna, V. V., & Doroshenko, A. V. (2024). System for intelligent analysis of big data and improvement of object classification methods. *Scientific Bulletin of UNFU*, 34(5), 119–128. https://doi.org/10.36930/40340516

ня, оброблення та інтелектуального аналізу великих даних в електронній комерції, які дають змогу підвищити ефективність їх оброблення та збільшити точність вибраних моделей класифікації.

Мета роботи – проаналізувати ефективність застосування класифікаторів методами випадкового лісу та опорних векторів для задач інтелектуального аналізу великих даних засобами машинного навчання, що дасть змогу збільшити точність класифікації даних для задач електронної комерції.

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

1. Розробити систему збирання, оброблення та зберігання даних із використанням сучасних технологій, що дасть змогу покращити підготовку даних для задач МН.
2. Проаналізувати та порівняти моделі та базові алгоритми машинного навчання для вирішення задач класифікації об'єктів, що дасть змогу визначити методи класифікації, які дають змогу отримати найвищу точність для вирішення конкретних задач, а також визначити вплив зміни параметрів моделей на точність класифікацій.
3. Застосувати параметри випадкового стану для класифікатора методом випадкового лісу і використання ймовірностей у класифікатора методом опорних векторів, що забезпечить підвищення точності методів класифікації для вирішення задачі із визначенням ступеня задоволеності послугами авіакомпанії клієнта.
4. Порівняти результати навчання моделей із використанням класифікатора методом наївного Байєса для визначення оптимальної моделі та вирішення задач класифікації у сфері електронної комерції, що допоможе на підставі базового класифікатора методом наївного Байєса здійснити підбір найефективнішої моделі для вирішення задач класифікації у сфері електронної комерції у задачах інтелектуального аналізу профілювання клієнтів.

Аналіз останніх досліджень та публікацій. Як показує аналіз, сьогодні активно здійснюються дослідження в області машинного навчання та вирішення задач за допомогою даних технологій, зокрема задач класифікації.

Методи оброблення великих даних під час роботи із алгоритмами машинного навчання виділено у праці [13]. Розглянуто основні проблеми передоброблення великих даних і варіанти вирішення. Досліджено інтеграцію бібліотек машинного навчання у систему оброблення великих даних.

Методи, системи та засоби інтелектуального аналізу даних розглянуто у працях [7, 8, 24]. Розглянуто методи вирішення низку інтелектуальних задач, що постають в аналізі даних. Дослідження представляють всеохопний аналіз поточних тенденцій інструментів, методів аналізу великих даних, їхніх сильних і слабких сторін. Аналіз даних досліджень дав змогу у поточному дослідженні проаналізувати вплив інтелектуального аналізу даних на сферу електронної комерції, підкресливши переваги і недоліки збирання великих даних у даній сфері і видалити, що для ефективної оброблення великих даних необхідно дані зменшити до такого формату, щоб доступні на цей момент алгоритми могли ефективно працювати із ними.

Алгоритми пошуку асоціативних правил та об'єкти їх застосування розглянуто у працях [3, 4, 18]. Алгоритми пошуку асоціативних правил стали зараз одним з популярних методів виявлення прихованих закономірностей та побудови знань. Сфера електронної комерції потребує глибокого аналізу інформації для досягнення успіху. Проаналізований матеріал досліджень дає змогу визначити об'єкти електронної комерції, у яких можна використати алгоритми пошуку асоціативних правил, до прикладу, аналіз ринкового кошика.

У роботах [12, 18] розглянуто вплив великих даних на сферу електронної комерції. Виділено способи отримання загальнодоступних даних, використовуючи аналіз веб сайтів підприємств. Інформація із даних джерел дала змогу визначити методи і засоби оброблення даних в електронній комерції, виділити такі засоби, як профілювання даних клієнта, а також персоналізацію послуг за допомогою методів інтелектуального аналізу.

Використання класифікаторів методами наївного Байєса, випадкового лісу, а також опорних векторів розглянуто в роботах [19, 23, 26]. Виділено, що недоліком застосування моделей класифікації є потреба у великих наборах загальнодоступних даних, що наразі є досить важко доступними на ринку. Виділено вплив коректного підбору параметрів під час навчання моделей, що дало змогу в поточному дослідженні здійснити порівняння результатів точності класифікатора методом наївного Байєса без додавання змін у логіку, що доступна за замовчуванням.

Водночас, на противагу попереднім роботам, у роботах [13, 26] виділено вплив розподілу даних за допомогою різних технік генерації штучних даних на підставі оригінальної вибірки на результати точності навчання моделей. У даних дослідження виділено техніки передоброблення даних для заповнення пропущених значень, що дало змогу підібрати необхідний алгоритм для вирішення проблем із даними у наборі даних із дослідження, а саме заповнення значень у колонці запізнення авіалінії у хвилинах.

У роботі [26] досліджено модель на базі класифікатора методом випадкового лісу і вплив на точність навчання моделі за експериментування із різними параметрами класифікатора, що дає змогу у поточному дослідженні зосередитись на проведенні експериментів із класифікаторами методами випадкового лісу й опорних векторів шляхом підбору і зміни параметрів даних моделей.

Матеріали та методи дослідження. У роботі використано методи машинного навчання, що використовуються для задач класифікації – класифікатор методом наївного Байєса, класифікатор методом випадкового лісу й опорних векторів. Для реалізації алгоритмів обрано мову програмування Python 3. Середовище розроблення розгорнуте за допомогою платформи Docker. Для оброблення та застосування алгоритмів обрано бібліотеки і фреймворки "Pyspark", "Sklearn". Для візуалізації використано бібліотеки "Seaborn", "Matplotlib".

Результати дослідження та їх обговорення / Research results and their discussion

Особливості задач інтелектуального аналізу даних. Глобальна електронна комерція створює величезну кількість даних, і аналітика великих даних є хорошою технікою для оброблення цих великих наборів даних і вилучення з них корисної інформації. Компанії електронної комерції мають справу як зі структурованими, так і з неструктурованими даними, коли йдеться про аналіз великих даних [18].

Сьогодні важко уявити діяльність підприємства без використання Інтернету в тій чи іншій формі. Використання великих даних створює можливості для отримання конкурентних переваг, оскільки з технічної точки зору вони пов'язані з обробленням величезних обсягів інформації.

Наприклад, персоналізація в комерції надає повну інформацію про клієнта, що дає змогу надати йому персоналізований сервіс на підставі даних про його переваги. За допомогою великих даних можна отримати інформацію про персональні дані клієнта (адреса, стать, вік, інтереси, хто у нього друзі, які сайти відвідує), а також визначити його комунікативні навички (що він шукає в мережі Інтернет, які відгуки він залишає). Аналіз такої інформації допомагає зрозуміти, що подобається клієнтам, чи готові вони ще робити покупки або їх потрібно стимулювати за допомогою різноманітних бонусів і знижок.

Великі дані є дані великого обсягу, що формуються із різноманітних джерел, мають різний інформаційний зміст та пов'язані з невизначеністю. Такі дані містять поєднання структурованих, напівструктурованих та неструктурованих даних [8].

В інформаційних технологіях для інтерпретації даних, як великі дані, необхідно щоб вони відповідали критеріям чотирьох "V". Перший критерій – це обсяг (Volume) даних, що представляє велике значення, водночас різні обсяги вважаються "великими" у різних середовищах. Наступним критерієм інтерпретації є швидкість (Velocity) того як дані генеруються та надходять у процес. Наступним критерієм є дані, які надходять із широкого спектра джерел даних і тому дуже відрізняються (Variety), наприклад, з точки зору методів запису, аналізу, структуризації тощо. І, на останок, інтерпретованим даним щодо змісту притаманна невизначеність, цілісність та надійність (Veracity). На противагу, звичайними вимогами до якості даних є своєчасність, охоплення та роздільність, що водночас не можна застосовувати для опису великих даних. Відповідно, в конкретному випадку для зручного подальшого оперування інформацією, необхідно спочатку витягти та оцінити дані за допомогою процедур інтелектуального аналізу даних [12].

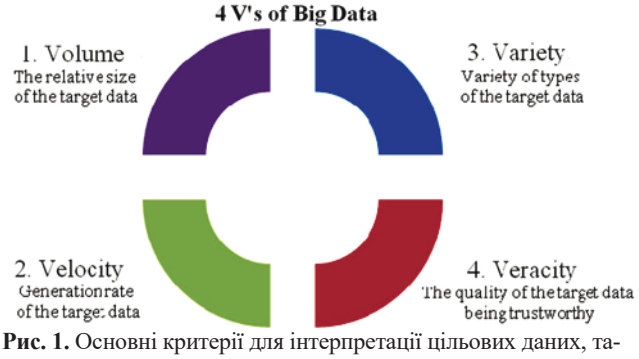

Рис. 1. Основні критерії для інтерпретації цільових даних, таких як "великі дані" / Basic criteria for interpreting target data such as big data

Мета полягає в тому, щоб зменшити розміри даних до такої міри, щоб відомі алгоритми забезпечували інтерпретовані результати за прийнятний час, оскільки потужності сучасних комп'ютерних технологій часто не здатні повністю обчислити великі запаси даних. З цією метою доречно використовувати метод інтелектуального аналізу даних (англ. *Data Mining*) [13].

Під інтелектуальним аналізом даних розуміють процес ідентифікації необхідних паттернів та їх кореляції один з одним шляхом просіювання великих обсягів даних, що зберігаються в сховищах. Існує декілька інструментів для створення цих даних, які містять абстракції, агрегації, узагальнення та характеристики даних. Інтелектуальний аналіз даних не є специфічним для одного типу даних та може бути пов'язаний з будь-яким типом джерела інформації, однак алгоритми та тактики можуть відрізнятися за застосування до різних типів даних [2].

На першому етапі процесу відбираються та вилучаються дані, які стосуються аналізу. Під час виділення вибирається набір записів даних (вертикальне виділення) та атрибутів (горизонтальне виділення). На етапі підготовки та перетворення (попереднє оброблення) спочатку перевіряється якість даних вибраного пулу даних. Для того, щоб дані, оброблені інструментом розпізнавання образів, використовувалися на наступному етапі, також необхідне перетворення даних у відповідний формат даних. На етапі розпізнавання образів здійснюється вибір і параметризація методу, а також його застосування на підставі підготовленого набору даних. Вибір методів визначається типом шуканих відносин, і отже, має сильну орієнтацію на ситуативну проблему [2].

Існує декілька основних методів інтелектуального аналізу даних, які використовуються, найпоширенішими з яких є: пошук асоціативних правил; кластеризація; прогнозний аналіз.

1. Пошук асоціативних правил, суть якого полягає у вилученні цікавої кореляції та асоціації між наборами елементів у транзакційних базах даних або інших наборах даних. Типове правило асоціації має вигляд A→B, де A – набір елементів, а B – набір елементів, який містить тільки одну атомарну умову. З точки зору електронної комерції, інтелектуальний аналіз асоціативних правил може знайти своє застосування в аналізі ринкового кошика, суть якого полягає у виявленні кореляцій між елементами, отриманими клієнтами в базі даних. Ґрунтовною складовою цього методу є визначення шаблонів, отриманих клієнтом шляхом видобуток-асоціації з транзакційної бази даних роздрібних продавців. Якщо говорити детальніше, то комерційні організації зацікавлені в аналізі асоціативних правил, які визначають шаблони покупок, у такий спосіб, що наявність одного товару в кошику вказуватиме на наявність одного або кількох додаткових товарів. Результат цього аналізу можна використовувати, щоб рекомендувати комбінації продуктів для спеціальних рекламних акцій або розпродажів, розробити більш реальний макет магазину та дати бачення лояльності до бренду та ко-брендінг. Це також приведе менеджерів до ефективного та реального прийняття стратегічних рішень [3].

Метою аналізу асоціативних правил є визначення залежностей між підмножинами набору даних. Ці залежності представлені у вигляді правил асоціації, які описують структуру та значення залежності. Структурними компонентами правил асоціації є передумова та наслідок. Ці два компоненти, які описують структуру шаблону, можуть містити будь-яку кількість елементів, пов'язаних за допомогою логічних операторів. Якість правил асоціації фіксується чинниками підтримки (support) та довіри (confidence). Допоміжний чинник робить

заяви про частоту визначеного правила. Коефіцієнт впевненості є показником надійності або суворості правила [21].

2. Кластеризація є організацією даних у класи шляхом групування подібних об'єктів для формування більш ніж одного класу методів. Іншими словами кластеризація ґрунтується на групуванні набору фізичних або абстрактних об'єктів у класи подібних об'єктів. Цей метод знайшов своє використання у класифікацій пропозицій на веб-платформах електронної комерції. Пропозиції мають різні параметри в багатьох вимірах, і простір окремих параметрів дуже важко порівняти між собою. Пропозиції мають заголовки, атрибути, ціни, інформацію про доставку. Вся дана інформація надається продавцем, однак не вся вона є необхідною. В описуваному методі кластеризації, товар є основною одиницею транзакції, втіленою набором неструктурованих і різнорідних даних. Це містить назву аукціону, опис, пари атрибутів ім'я-вартість, ціну [26].

3. Прогнозний аналіз, який засновують на двох типах передбачень. Перший передбачає прогнозування недоступних значень даних, а другий полягає в передбаченні мітки класу об'єкта на підставі значень атрибутів об'єкта. Прогнозний аналіз в електронній комерції можна застосувати для вирішення таких задач [6]:

- рекомендаційні системи;
- аналіз поведінки клієнтів;
- прогнозування продажу;
- цінова оптимізації.

Застосування інтелектуального аналізу даних в електронній комерції входить до можливих сфер покращення бізнесу. Відвідуючи інтернет-магазин для покупок, користувачі зазвичай залишають певні факти, які компанії можуть зберігати у своїй базі даних. Ці факти є неструктурованими або структурованими даними, які можна отримати, щоб забезпечити конкурентну перевагу компанії. Інтелектуальний аналіз даних може бути застосований у сфері електронної комерції на користь компаній у таких сферах:

1) профілювання клієнта (англ. *Client Profiling*). Це також відомо як орієнтована на клієнта стратегія в електронній комерції. Це дає змогу компаніям використовувати бізнес-аналітику за допомогою аналізу даних клієнтів, щоб планувати свою бізнес-діяльність і операції, а також розробляти нові дослідження продуктів або послуг для успішної електронної комерції. Класифікація потенційних клієнтів, які здійснюють великі покупки, на підставі даних про відвідування може допомогти компаніям зменшити витрати на продаж [19]. Компанії можуть використовувати дані веб-перегляду користувачів, щоб визначити, чи здійснюють вони покупки цілеспрямовано, чи просто переглядають чи купують щось знайоме, чи щось нове. Це допомагає компаніям планувати та вдосконалювати свою інфраструктуру [21];

2) персоналізація послуг (англ. *Personalization of Services*), спрямована на надання вмісту та послуг, орієнтованих на окремих осіб на підставі інформації про їхні потреби та поведінку. Дослідження інтелектуального аналізу даних, пов'язані з персоналізацією, зосереджені в основному на системах рекомендацій і пов'язаних з ними темах, таких як спільна фільтрація. Системи рекомендацій можна поділити на три групи: на підставі вмісту, аналізу соціальних даних і спільної фільтрації. Ці системи культивуються та вивчаються на підставі явних чи неявних відгуків користувачів і зазвичай представлені як профіль користувача. Інтелектуальний аналіз соціальних даних, якщо розглядати джерело даних, створених групою осіб у рамках їх повсякденної діяльності, може бути важливим джерелом інформації для компаній. Навпаки, персоналізація може бути досягнута за допомогою спільної фільтрації, коли користувачі зіставляються з особливим інтересом і в тому ж ключі вподобання цих користувачів для надання рекомендацій [14];

3) аналіз кошика (англ. *Shopping Cart Analysis*). Кожен кошик покупця має свою історію, а аналіз ринкового кошика є поширений інструмент роздрібної торгівлі та бізнес-аналітики, який допомагає роздрібним торговцям краще знати своїх клієнтів. Існують різні способи отримати більшу користь від аналізу ринкового кошика, і вони містять [4]:

- виявлення спорідненості продукту. Клієнти, які купують певні товари, виявляють прихильність до одного з трьох товарів в асортименті, що є недослідженим зв'язком, який можна виявити за допомогою вдосконаленої аналітики ринкового кошика для планування ефективніших маркетингових заходів;
- кампанії перехресних продажів; тут показано продукти, придбані разом, тому клієнтів, які купують певний товар, можна переконати вибрати інший товар, який безпосередньо пов'язаний з товаром, який обрав клієнт (принтери та папір з картриджами);
- планограми та комбінації продуктів – використовуються для кращого контролю запасів на підставі спорідненості продуктів, розроблення комбінованих пропозицій і розроблення ефективних зручних планограм для зосередження на продуктах, які продаються разом;
- профіль покупців; в аналізі ринкового кошика за допомогою аналізу даних протягом певного часу, щоб отримати уявлення про те, ким є покупці, отримуючи уявлення про їхній вік, діапазон доходів, купівельні звички, симпатії та антипатії, купівельні переваги;

4) прогнозування продажів (англ. *Sales Forecasting*), яке містить особливість часу, який окремий клієнт витрачає на покупку товару, і в цьому процесі відбувається спроба передбачити, чи клієнт зробить покупку знову. Цей тип аналізу можна використовувати для визначення стратегії запланованого морального старіння або визначення додаткових продуктів для продажу. У прогнозуванні продажу грошовий потік можна спроектувати за трьома групами, які містять песимістичний, оптимістичний і реалістичний. Це допомагає мати план щодо достатньої кількості доступного капіталу [22];

5) планування товарів (англ. *Planning of goods*). У разі онлайн-бізнесу планування товарів допоможе визначити варіанти зберігання запасів [16]. Використання правильного підходу до планування товарів обов'язково призведе до відповідей щодо того, що робити з:

- ціноутворенням: особливість аналізу бази даних допоможе визначити найкращу ціну продуктів або послуг у процесах виявлення чутливості клієнтів;
- вибір продуктів; інтелектуальний аналіз даних надає підприємствам електронної комерції інформацію про те, які продукти насправді бажають клієнти, що містить особливість інформації про товари конкурентів;
- балансування запасів; під час аналізу бази даних роздрібної торгівлі це допомагає визначити правильну та конкретну кількість необхідних запасів;

6) сегментація ринку (англ. *Market Segmentation*) дає змогу велику кількість отриманих даних розбити на різні та значущі сегменти, як-от дохід, вік, стать, професія клієнтів, і це можна використовувати, коли компанії проводять маркетингові кампанії. Особливість сегментації ринку також може допомогти компанії визначити своїх власних конкурентів. Ця надана інформація сама по собі може допомогти роздрібній компанії визначити, що періодичні респонденти зазвичай не єдині, хто вказує клієнтам ті самі гроші, що й поточна компанія. Сегментація бази даних роздрібної компанії покращить коефіцієнти конверсії, оскільки компанія може зосередитися на просуванні на тісному ринку. Це також допомагає компанії роздрібної торгівлі зрозуміти конкурентів, які беруть участь у кожному сегменті процесу, даючи змогу персоналізувати продукти, які фактично задовольняють цільову аудиторію в загальному вигляді [28]. У роботі досліджено методи та інструменти, що використовуються для аналізу великих даних у контексті електронної комерції, в т.ч. й моделі машинного навчання, методи статистичного аналізу та методи візуалізації даних для розв'язання задачі класифікації.

Підготовка даних. Для проведення дослідження обрано набір даних, що містить опитування пасажирів авіакомпанії [9]. Підготовка даних поділилась на такі етапи:

1) завантаження вхідного набору даних у систему;
2) виокремлення основних компонентів із набору даних, профайлінг даних для визначення цілісності і коректного розподілу даних;
3) виконано перевірку набору даних на наявність пропущених значень;
4) здійснено заповнення пропущених значень у стовпці "Arrival Delay In Minutes" медіаною цього ж стовпця для навчального та тестового наборів даних.

Класифікатор методом випадкового лісу (англ. *Random Forest*). Формула для прийняття рішення методом випадкового лісу полягає в узагальненні результатів багатьох дерев прийняття рішень, які вирішуються шляхом використання випадкових підвибірок функцій та даних. Класифікатор методом випадкового лісу обчислює прогнози шляхом узагальнення прогнозів декількох рішень, отриманих від кожного дерева рішень у лісі. На рис. 2, наведено спрощену блок-схему класифікатора методом випадкового лісу.

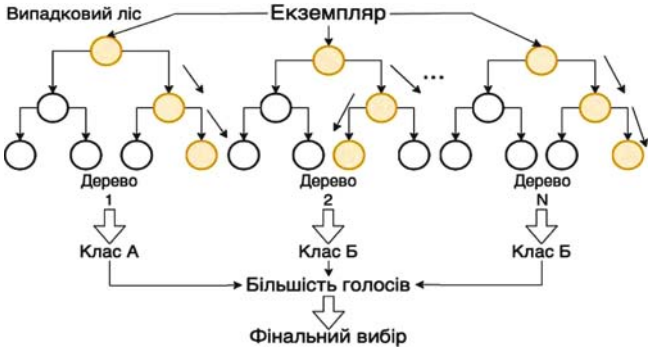

Рис. 2. Спрощена блок-схема класифікатора методом випадкового лісу / Simplified block diagram of the Random Forest classifier

Класифікатор методом наївного Байєса (англ. *Naive Bayes*) використовує теорему Байєса для обчислення ймовірності кожного класу на підставі вхідних ознак, згідно з якою всі ознаки незалежні між собою, що є ідеалізованим припущенням (звідси і назва "наївний"). Формула для обчислення ймовірності класу *Ck* для вхідних ознак x*x* може бути записана як:

$$P(Ck|x) = P(x)\ P(x|Ck) \cdot P(Ck),$$

де: $P(Ck|x)$ – ймовірність класу *Ck* при заданих вхідних ознаках *x*; $P\ (x|Ck)$ – ймовірність вхідних ознак *x* при умові класу *Ck*; $P(Ck)$ – апріорна ймовірність класу *Ck*; $P(x)$ – ймовірність вхідних ознак *x*.

Блок-схему класифікатора методом наївного Байєса зображено на рис. **3.**

Рис. 3. Спрощена блок-схема класифікатора методом наївного Байєса / Simplified block diagram of the Naive Bayes classifier

Класифікатор методом опорних векторів (англ. *Support Vector Networks*, SVN). Метод опорних векторів вирішує задачу класифікації, знаходячи оптимальну розділову гіперплощину між класами. Він шукає таку гіперплощину, що максимізує відстань між найближчими до неї точками кожного класу (опорні вектори). Формула для розділової гіперплощини може бути записана як:

$$w^{Tx} + b = 0,$$

де: *w* – вектор ваг; *x* – вхідні ознаки; *b* – зсув (параметр зміщення); w^{Tx} – транспонований вектор ваг.

На рис. 4 відображено спрощену блок-схему цього методу.

Навчання та тренування моделей. Для навчання та тестування моделей здійснено такі етапи:

- розподіл вхідних даних на тренувальну і тестувальну вибірку. Поділ відбувався відповідно 75 % до 25 %;
- проведено визначення вчителя, що є колонкою *satisfaction*;
- обрано метрики для оцінювання результатів навчання. Набір даних містить якісно розподілену інформацію, що дало змогу робити експеримент, використовуючи усі колонки як метрики окрім учителя [16];
- у класифікатора методом випадкового лісу змінено значення параметра випадкового стану для покращення точності аналізу даних.

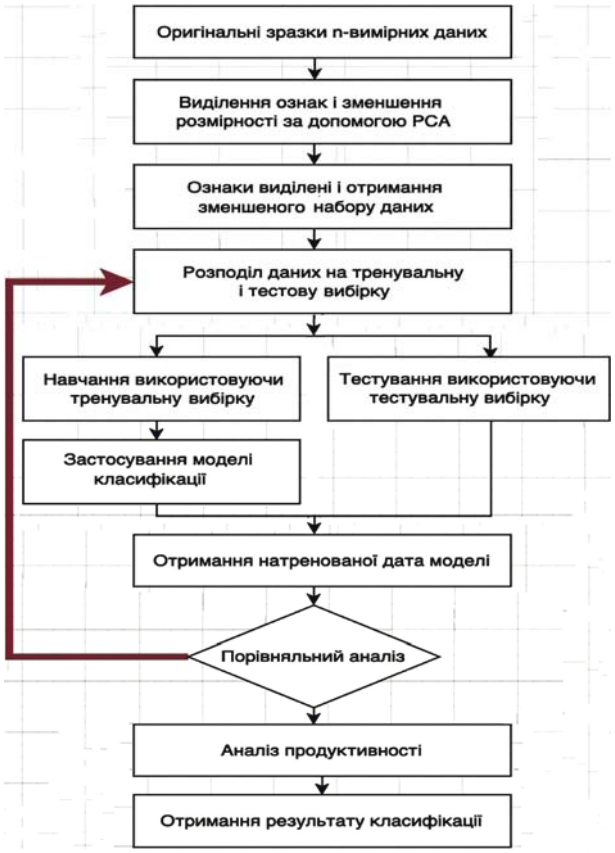

Рис. 4. Блок-схема класифікатора методом опорних векторів / Simplified block diagram of the classifier by the support vector method

Важливою особливістю під час роботи із великими даними і аналізом показників клієнтів у електронній комерції є можливість визначення певних показників і на підставі вхідних критеріїв. Одним із основних і найважливіших чинників у роботі із даними електронної комерції є визначення ступеня задоволеності певним сервісом клієнта. До цього зводиться увесь процес попереднього збирання, оброблення даних про комунікацію клієнта із сервісами. Отримання інформації про ступінь задоволеності клієнта дасть змогу бізнесу визначити критерії, щоб відповідати вимогам клієнтів.

Для проведення дослідження обрано набір даних, що містить опитування пасажирів авіакомпанії. Поточний набір даних містить більше двадцяти п'яти тисяч рядків даних. Водночас ця задача розрахована на вибірки із значно більшим потоком даних. Цей набір містить таку інформацію:

- стать: стать пасажирів (жінка, чоловік);
- тип клієнта: тип клієнта (лояльний клієнт, нелояльний клієнт);
- вік: фактичний вік пасажирів;
- тип подорожі: Мета рейсу пасажирів (Особиста подорож, Ділова подорож);
- клас: Клас подорожі в літаку пасажирів (Бізнес, Еко, Еко Плюс);
- відстань польоту: відстань польоту цієї подорожі;
- послуга Wi-Fi на борту: рівень задоволеності службою Wi-Fi на борту (0: не застосовується; 1-5);
- зручний час відправлення/прибуття: рівень задоволеності зручним часом відправлення/прибуття;
- простота онлайн-бронювання: рівень задоволеності онлайн-бронюванням.
- розташування воріт: рівень задоволення розташуванням воріт;
- їжа та напої: рівень задоволеності їжею та напоями;
- онлайн-посадка: рівень задоволення онлайн-посадкою;
- комфорт сидіння: рівень задоволеності комфортом сидіння;
- розваги на борту: рівень задоволення розвагами на борту;
- обслуговування на борту: рівень задоволеності обслуговуванням на борту;
- обслуговування номерів ніг: рівень задоволення обслуговуванням номерів ніг;
- оброблення багажу: рівень задоволеності обробкою багажу;
- послуга реєстрації: рівень задоволеності послугою реєстрації;
- обслуговування на борту: рівень задоволеності обслуговуванням на борту;
- чистота: рівень задоволеності чистотою;
- затримка відправлення у хвилинах: хвилини затримки відправлення;
- затримка прибуття у хвилинах: хвилини затримки прибуття;
- задоволеність: рівень задоволеності авіакомпанією (задоволення, нейтральне або незадоволене) [9].

На рис. 5 наведено статистику розподілу даних за рівнем задоволеності сервісами клієнтів.

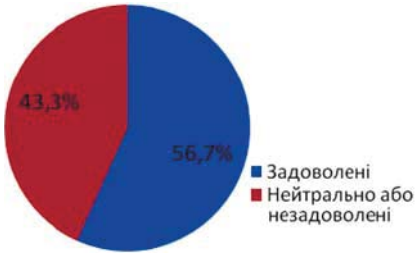

Рис. 5. Розподіл показника задоволеності клієнта у наборі даних / Distribution of the customer satisfaction score in the data set

Суть задачі полягає у навчанні моделей класифікації для визначення ступеня задоволеності клієнта на підставі набору показників. За основу потрібно взяти класифікатор методом випадковго лісу, а також порівняння результатів цього класифікатора із іншими моделями задач класифікації. Для вирішення задачі обрано такий набір класифікаторів методами:

- випадковий ліс (Random Forest);
- наївного Байєса (Naive Bayes);
- опорних векторів (SVM).

Під час експерименту дані було поділено на 75 % тренувальна вибірка і 25 % тестова. Далі набори даних було передано у необхідні об'єкти. Для оцінювання результатів задач класифікації використовують різні метрики, які вимірюють різні особливості точності та ефективності класифікатора. Суть завдання: було порівняння точності моделей на підставі класифікаторів методами випадкового лісу і опорних векторів під час зміни налаштувань цих моделей, порівняно із класифікатором методом наївного Байєса. Під час експериментів основну увагу приділено параметрам випадкового стану для класифікатора методом випадкового лісу і параметру ймовірність у класифікатора методом опорних векторів.

Випадковий стан – це параметр, що визначає початкове значення для генератора псевдовипадкових чисел, який використовується класифікатором методом випадкового лісу. У разі використання алгоритмів, які мають компонент випадковості, випадковий стан дає змогу забезпечити відтворюваність результатів. Це означає, що під час кожного запуску моделі з однаковим значенням

параметру випадкового стану, результати будуть однаковими. Якщо випадковий стан не задано, то модель використовуватиме внутрішній генератор псевдовипадкових чисел, що може призводити до різних результатів під час кожного запуску [23].

Параметр ймовірність у класифікатора методом опорних векторів пов'язаний з можливістю оцінки ймовірностей належності класів, а не тільки жорсткої класифікації. Коли параметр ймовірність встановлено на True, модель обчислює ймовірності класів для кожного передбачення. Це корисно, коли потрібно мати не тільки рішення про клас, але й впевненість моделі у своєму рішенні. Модель сама по собі не генерує ймовірностей. Щоб отримати ймовірності, використовують методи калібрування, такі як Platt Scaling [23].

Для оцінення впливу на точність класифікації для класифікатора методом випадкового лісу проведено експерименти із зміною параметра випадковий стан. Для поточною задачі найкращі результати отримано у діапазоні параметру випадковий стан від 40 до 50 одиниць, що отримано шляхом зміни, добору і аналізу результатів точності класифікації. У класифікатора методом опорних векторів для цього експерименту здійснено зміну і аналіз результатів точності класифікації даної задачі у разі використанні параметра ймовірності у моделі. Згідно з результатами, у поточній задачі у разі використання параметра ймовірностей отримано кращі результати точності класифікатора методом опорних векторів. Оцінювання результатів навчання моделей проведемо за допомогою таких метрик [11]:

1) точність (Accuracy);
2) влучність (Precision);
3) повнота (Recall);
4) F-міра (F-measure).

Також для відображення результатів класифікації обрано матрицю невідповідностей. Матриця невідповідностей водночас представляє співвідношення помилок до коректно спрогнозованих результатів. Далі наведено приклад експерименту [19].

Таблиця. Метрики оцінювання класифікацій / Classification evaluation metrics

	Класифікатор методом випадкового лісу	Класифікатор методом наївного Байєса	Класифікатор методом опорних векторів
Метрика	Результат	Результат	Результат
Точність	0,96	0,86	0,95
Влучність	0,97	0,86	0,95
Повнота	0,94	0,82	0,93
F-міра	0,96	0,84	0,94

Результати показників, зазначених у таблиці, вимірюють у шкалі від 0 до 1, тому їх одиницею виміру є частка, а саме певне співвідношення. Матриці невідповідностей для зазначених у таблиці результатів класифікацій наведено на рис. 6.

Далі порівнюються результати експерименту для кожної моделі класифікації. Діаграма порівняння класифікацій наведена на рис. 7.

Згідно з рис. 7, можна дійти висновку, що на підставі вхідного набору даних найкраще для вирішення саме певної задачі надаються класифікатори методами випадкового лісу і опорних векторів, оскільки точності класифікації є близько 90 %, що приблизно на 10 % випереджає модель класифікатора методом наївного Байєса.

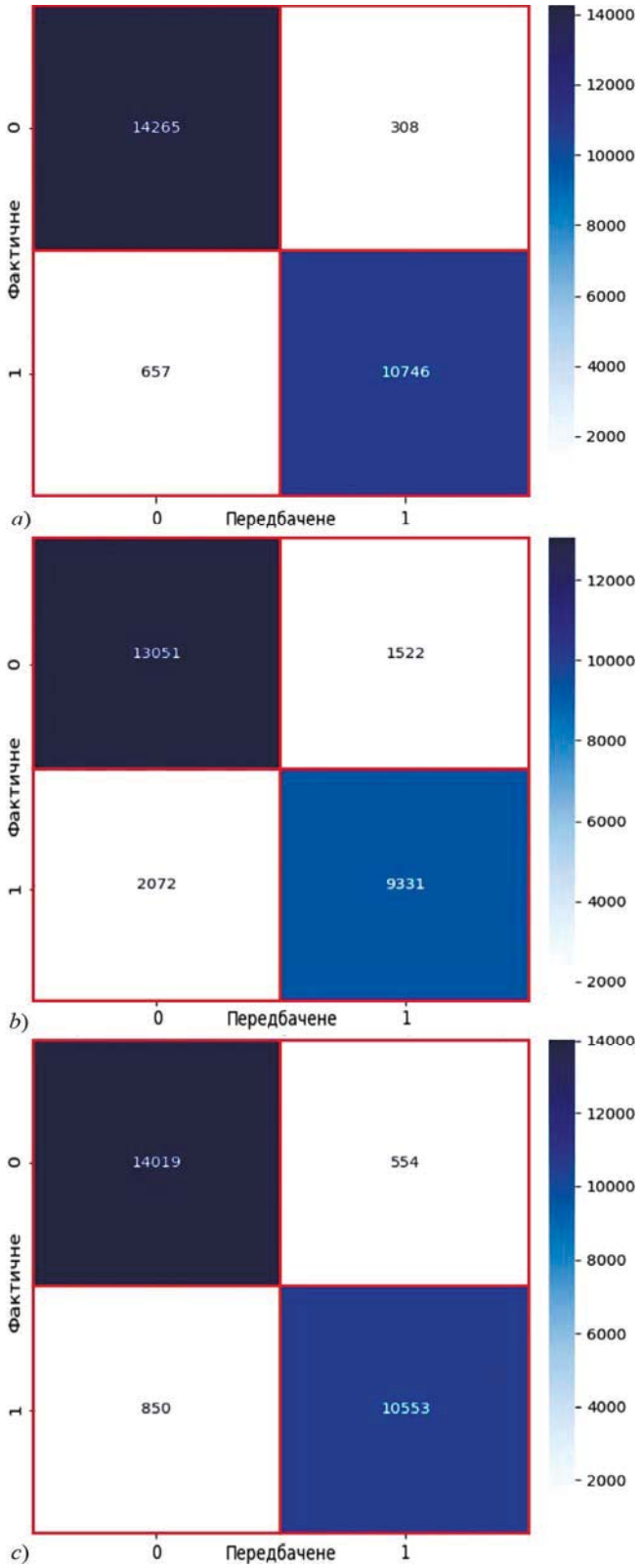

Рис. 6. Матриці невідповідностей класифікатора / Matrices of inconsistencies of the classifier using the: a) методом випадкового лісу / random forest method; b) методом наївного Байєса / Naive Bayes method; c) методом опорних векторів / SVN method

Обговорення результатів дослідження. Порівнюючи отримані в ході експериментів результати із результатами, наведеними в дослідженнях [5, 15] варто зазначити, що аналіз вхідних даних і вибір коректних атрибутів під час навчання моделі мають значний вплив на результати точності калькуляцій. У цих дослідженнях використовуються вибірки, що містить такі дані, як

стать людини. Потрібно виділити різницю між класифікатором методом опорних векторів, що дав змогу досягнути точності 87 %, тоді як класифікатор методом випадкового лісу працює найменш оптимально з показником 80 % під час тренування на одному і тому ж наборі даних, виділивши основним атрибутом стать. У поточному дослідженні ці два методи на обраному наборі даних показали майже однакові результати із незначною перевагою у класифікатора методом випадкового лісу.

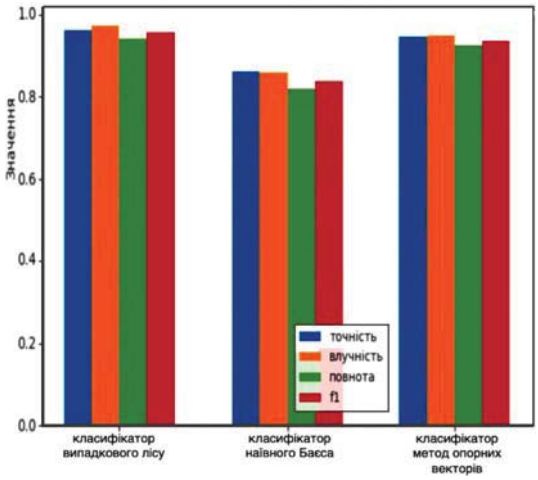

Рис. 7. Діаграма порівняння результатів навчання моделей / Comparison diagram of model training results

У роботі [1] досліджують класифікатори методами випадкового лісу і опорних векторів. Визначено вплив кількості вхідних даних на результати дослідження та здійснено порівняння класифікатора методом опорних векторів до класифікатора методом випадкового лісу, де на підставі найбільшої вибірки отримано результати точності моделей, більше 92 % для класифікатора методом випадкового лісу. У поточному дослідженні після проведення експериментів зі зміни параметра випадкового стану досягнуто точності 96 % за схожим підбором атрибутів для навчання.

Під час розв'язання задачі адаптації студентів [25] до онлайн навчання за допомогою класифікатора методом випадкового лісу досягнуто точності алгоритму 91 % використовуючи коректний набір вхідних параметрів.

Водночас у роботі [20] виокремлено вплив ітеративного алгоритму зважування RIM на результати точності класифікації моделей. Із використанням цього алгоритму вдалось досягнути точності 95 % порівняно із аналогічними результатами без використання алгоритму. Для порівняння, класифікатор методом наївного Байєса у поточному дослідженні показав результати точності 86 %, що в середньому на десять одиниць менша, ніж у класифікаторів методами випадкового лісу і опорних векторів.

Описано вплив характеристик даних на ефективність класифікації у роботі [10], що дало змогу досягнути для класифікатора методом випадкового лісу 85,37 % правильно класифікованих результатів. Розглянуто методологію, що включає збір, аналіз, моделювання і оцінку ефективності класифікатора.

Отже, за результатами виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – отримали подальший розвиток моделі класифікації даних шляхом проведення відповідних експериментів класифікаторів методами випадкового лісу і опорних векторів, що ґрунтуються, відповідно, на підборі показника випадкового стану і параметра ймовірності.

Практична значущість результатів дослідження – розроблені моделі класифікації даних дають змогу проводити аналіз даних у сфері електронної комерції для визначення потреб клієнтів, будуть корисними для використання у відділі аналітики споживчих підприємств, авіакомпаній та інших об'єктів господарської діяльності.

Висновки / Conclusions

Проаналізовано ефективність застосування класифікаторів методами випадкового лісу та опорних векторів для задач інтелектуального аналізу великих даних засобами машинного навчання, що дасть змогу збільшити точність класифікації даних для задач електронної комерції. За результатами проведеного дослідження можна зробити такі основні висновки.

1. Проведено аналіз особливостей задач інтелектуального аналізу даних, методологій оброблення великих даних у сфері електронної комерції з використанням інтелектуального аналізу даних. Проведено аналіз вхідних даних і здійснено попередню підготовку даних для навчання та тестування вибраних моделей.
2. Висвітлено актуальність роботи з великими даними у сфері електронної комерції, оскільки профілювання даних клієнтів, інформації про їх дії, покупки, а також аналіз інформації конкурентів дає змогу отримати конкурентні переваги.
3. Проаналізовано можливість застосування інтелектуального аналізу даних для підтримки прийняття рішень у процесі комерційного управління. Окрім цього, цей процес забезпечує інформаційну основу для розроблення адаптивних систем магазинів, які дають змогу еволюційно адаптувати онлайн-магазини до вподобань покупців. Це відкриває політику адаптації параметрів дій для конкретного каналу для роздрібної торгівлі в мережі Інтернет, що сприяє збуту. Результати інтелектуального аналізу даних у електронній комерції дають змогу створити базу даних щодо комплексного аналізу інформації та купівельної поведінки.
4. Розроблено систему підготовки даних для їх оброблення. Проведено експеримент з аналізу даних клієнтів на підставі набору даних про опитування клієнтів авіакомпанії. Обраним завданням дослідження було вирішення задачі класифікації для визначення ступеня задоволення клієнта сервісами авіакомпанії. Досліджено класифікатор методом опорних векторів, класифікатор методом випадкового лісу та класифікатор методом наївного Байєса.
5. Проведено порівняння результатів точності даних алгоритмів. Здійснено аналіз результатів класифікатора методом випадкового лісу у разі підбору і зміни параметра випадкового стану, а також впровадження параметра ймовірності у класифікатор методом опорних векторів. Здійснено порівняння результатів дослідження із результатами останніх досліджень, в яких використано ці ж методи. На підставі проведеного дослідження виявлено позитивний тренд у результатах точності класифікаторів методами випадкового лісу та опорних векторів, особливо порівняно із класифікатором методом наївного Байєса із використанням параметрів за замовчуванням.

References

1. Avcı, C., Budak, M., Yağmur, N., & Balçık, F. (2023). Comparison between random forest and support vector machine algorithms for LULC classification. *International Journal of Engineering and Geosciences*, 8(1). https://doi.org/10.26833/ijeg.987605
2. Dai, H.-N., Wang, H., Xu, G., Wan, J., Imran, M., Dai, H.-N., & Xu, G. (2019). Big Data Analytics for Manufacturing Internet of Things: Opportunities, Challenges and Enabling Technologies. *Enterprise Information Systems*, 14. https://doi.org/10.1080/17517575.2019.1633689
3. Deng, R. (2022). Research on value mining of management accounting non-financial data based on association rules algorithm. In: 2022 2nd *International Conference on Networking, Communications and Information Technology* (NetCIT), Manchester, United Kingdom, 175–178. https://doi.org/10.1109/NetCIT57419.2022.00051
4. H. J. V. L., & Rajan, D. (2023). Enhancing customer experience and sales performance in a retail store using association rule mining and market basket analysis. In: 2023 14th *International Conference on Computing Communication and Networking Technologies* (ICCCNT), Delhi, India, 1–5. https://doi.org/10.1109/ICCCNT56998.2023.10307411
5. Hammoumi, L., Maanan, M., & Rhinane, H. (2024). Characterizing Smart Cities Based on Artificial Intelligence. *Smart Cities*, *7*(3). https://doi.org/10.3390/smartcities7030056
6. Hastie, T., Tibshirani, R., & Friedman, J. (n.d.). The elements of statistical learning: Data mining, inference, and prediction. URL: https://web.stanford.edu/~hastie/ElemStatLearn/
7. Horohovatskyi, V. O., & Tvoroshenko, I. S. (2021). Methods of intellectual analysis and data processing. URL: https://openarchive.nure.ua/handle/document/15868
8. Ikegwu, A., Nweke, H., Anikwe, C., Alo, U., & Okonkwo, O. (2022). Big data analytics for data-driven industry: A review of data sources, tools, challenges, solutions, and research directions. *Cluster Computing*, 25, 3343–3387. https://doi.org/10.1007/s10586-022-03568-5
9. Kaggle. (2023). Airline passenger satisfaction. URL: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction
10. Kalinina, Iryna, & Gozhyj, Alexander. (2021). Study of the efficiency of classification methods in forecasting in machine learning tasks. *Management of Development of Complex Systems*, 46, 173–180. https://doi.org/10.32347/2412-9933.2021.46.173-180
11. Keskar, V., Yadav, J., & Kumar, A. (2022). Perspective of anomaly detection in big data for data quality improvement. *Materials Today: Proceedings*, 51, 532–537. https://doi.org/10.1016/j.matpr.2021.06.145
12. Kumar, A., Kalia, P., & Mann, B. S. (2020). Utilization of big data in e-commerce business. *Journal of Information and Optimization Sciences*, *41*(6), 1391–1401. https://doi.org/10.1080/02522667.2020.1794434
13. L'Heureux, A., Grolinger, K., El Yamany, H., & Capretz, M. (2017). Machine Learning With Big Data: Challenges and Approaches, 5, 7776–7797. https://doi.org/10.1109/ACCESS.2017.2696365
14. Li, H. J., & Yang, D. X. (2006). Study on data mining and its application in e-business. *Journal of Gansu Lianhe University* (*Natural Science*), *20*(6), 30–33. URL: https://core.ac.uk/download/pdf/11784915.pdf
15. Musleh, D., Alkhwaja, A., Alkhwaja, I., Alghamdi, M., Abahussain, H., Albugami, M., Alfawaz, F., El-Ashker, S., & Al-Hariri, M. (2024). Machine Learning Approaches for Predicting Risk of Cardiometabolic Disease among University Students. *Big Data and Cognitive Computing*, *8*(3). https://doi.org/10.3390/bdcc8030031
16. Natchiar, S. U., & Baulkani, S. (2014). Customer relationship management classification using data mining techniques. In: 2014 *International Conference on Science Engineering and Management Research* (ICSEMR), Chennai, India, 1–5. https://doi.org/10.1109/ICSEMR.2014.7043662
17. Nti, I., Quarcoo, J., Fosu, G., & Aning, J. (2022). A Mini-Review of Machine Learning in Big Data Analytics: Applications, Challenges, and Prospects. *Big Data Mining and Analytics*, 5, 81–97. https://doi.org/10.26599/BDMA.2021.9020028
18. Painuly, S., Sharma, S., & Matta, P. (2021). Big data driven e-commerce application management system. In: 2021 6th *International Conference on Communication and Electronics Systems* (ICCES), Coimbatore, India, 1–5. https://doi.org/10.1109/ICCES51350.2021.9489108
19. Pan, W., Washizaki, H., Yoshioka, N., Fukazawa, Y., Khomh, F., & Guéhéneuc, Y. (2023). A machine learning based approach to detect machine learning design patterns. In: 2023 30th *Asia-Pacific Software Engineering Conference* (APSEC), Seoul, Korea, Republic of, 574–578. https://doi.org/10.1109/APSEC60848.2023.00073
20. Petryna, V. V., Doroshenko, A. V., Sydorenko, R. V., & Teslyuk, V. M. (2023). Model and means of data collection and processing using machine learning. *Scientific Bulletin of UNFU*, *33*(3), 102–109. https://doi.org/10.36930/40330315
21. Raghavan, S. N. R. (2005). Data mining in e-commerce: A survey. *Sadhana*, 30, 275–289. https://doi.org/10.1007/BF02706248
22. Rao, T. R., Mitra, P., Bhatt, R., & Goswami, A. (2019). The big data system, components, tools, and technologies: A survey. *Knowledge and Information Systems*, 60(3), 1165–1245. https://doi.org/10.1007/s10115-018-1248-0
23. Raschka, S., & Mirjalili, V. (2019). Python machine learning. URL: https://www.w3schools.com/python/python_ml_getting_started.asp
24. Subach, I., & Mykytiuk, A. (2023). The method pf forming associative rules from the SIEM database – systems based on the theory of fuzzy sets and linguistic terms. *Electronic professional scientific publication "Cybersecurity: education, science, technology"*, *3*(19). https://doi.org/10.28925/2663-4023.2023.19.2033
25. Teslyuk, V., Doroshenko, A., & Savchuk, D. (2023). Intelligent Methods and Models for Assessing Level of Student Adaptation to Online Learning https://www.scopus.com/authid/detail.uri?authorId=57202210835
26. Turban, E., Whiteside, J., King, D., & Outland, J. (2017). Introduction to electronic commerce and social commerce. Springer Link. https://doi.org/10.1007/978-3-319-50091-1
27. Uncertainty Based Under-Sampling for Learning Naive Bayes Classifiers Under Imbalanced Data Sets, IEEE Journals & Magazine, IEEE Xplore. (n.d.). (2024). URL: https://ieeexplore.ieee.org/document/8939418
28. Xu, X., Jia, L., Wang, Z., Zhang, H., Liang, S., & Zhou, C. (2005). Fast algorithm for mining item profit in retails based on microeconomic view. In: 2005 *International Conference on Cyberworlds* (CW'05), Singapore, 353. https://doi.org/10.1109/CW.2005.44

V. V. Petryna, A. V. Doroshenko

Lviv Polytechnic National University, Lviv, Ukraine

SYSTEM FOR INTELLIGENT ANALYSIS OF BIG DATA AND IMPROVEMENT OF OBJECT CLASSIFICATION METHODS

The study reviewed the methods of intelligent data analysis based on the concept of machine learning in order to increase their effectiveness in the field of e-commerce. A comparative analysis of the use of such models as a random forest classifier, a naive Bayes classifier, and a classifier based on the support vector method was conducted. Algorithm analysis software has been developed

for the problem of customer classification, which is common in the field of e-commerce. An analysis of the input data was carried out and preliminary data preparation was performed for training and testing the selected models. The study of the selected models was conducted using pre-prepared data with the help of software according to the defined scenarios. The parameters of the selected classification models were studied and the random forest classifier was improved by selecting and changing the random state parameter. The support vector method is also implemented in the classifier for the probability support parameter. Pre-prepared research data of selected models using software according to defined scenarios was involved. The support vector method has been introduced to the probability support parameter in the classifier. A comparison of the classification accuracy results of the selected classification models was made. According to the results of the study, a positive trend was determined for the quality of model training with correct data preparation and the influence of selecting the correct parameters for random forest algorithms and the support vector method. Indicators of efficiency and accuracy of algorithm training show positive dynamics and are comparable with the results of testing the naive Bayes classifier model with the basic values of the model parameters. Therefore, the results of the study reveal that the influence of the selection of correct parameters based on the input data set on the results of prediction accuracy of algorithms and their influence on training and testing of machine learning models is confirmed. These results emphasize the prospects for further research in the development of optimal strategies for optimization and efficiency improvement in working with machine learning algorithms in classification tasks.

Keywords: e-commerce; big data; intelligent data analysis; classification; random forest; support vector method; naive Bayes classifier.