DOI 10.31891/2307-5732-2023-323-4-25-35
УДК 004.042

БОЙКО НАТАЛІЯ.
Національний університет «Львівська політехніка»
ORCID ID: 0000-0002-6962-9363
e-mail: Nataliya.i.boyko@lpnu.ua
МИХАЙЛИШИН ВЛАДИСЛАВ
Національний університет «Львівська політехніка»
ORCID ID: 0000-0003-1889-9053
e-mail: vladyslavmykhailyshyn@gmail.com

ОЦІНКА ЕФЕКТИВНОСТІ РЕКУРСИВНОГО ПРОЦЕСУ РОЗПОДІЛУ НАБОРУ ДАНИХ З ВИКОРИСТАННЯМ АЛГОРИТМУ CART

В роботі наведено результати досліджень та порівняння результатів зарубіжних і вітчизняних праць, які показали високу ефективність моделі CART у прогнозуванні ефективності рекламних кампаній, що збігається з висновками інших дослідників. Наведене порівняння, що дозволяє підтвердити переваги і стабільність алгоритму у контексті оцінки рекламних кампаній. Наведено алгоритм збору, обробки та аналізу даних для застосування методу CART. Розглянуто процес ділення вузлів, який здійснюється до досягнення заданої кількості вузлів або до досягнення певного рівня глибини дерева. Наведена оціночна функція, що використовується для ділення вузлів та базується на Gini-індексі, який оцінює нечистоту у вузлі. Чим менше нечистота вузла, тим більше вважається його вагомим для подальшого ділення. Розроблено модель оцінки ефективності рекламних кампаній використовуючи алгоритм CART. Наводиться методика перевірки точності розробленої моделі. Порівнюються результати роботи моделі з реальними даними. Новизною дослідження є використання алгоритму CART для оцінки ефективності рекламних кампаній. Аналізується метод, який дозволяє швидко та точно аналізувати великі обсяги даних та визначати найважливіші чинники, які впливають на ефективність рекламних кампаній. Обґрунтовується практичне значення дослідження, яке полягає в тому, що розроблений алгоритм дозволяє раціонально використовувати бюджет на маркетингові заходи та оптимізувати рекламні кампанії з метою досягнення найкращих результатів.

Ключові слова: алгоритм, Classification and Regression Tree, Gini-індекс, Receiver Operating Characteristic, Area Under the Curve.

BOYKO NATALIYA
Lviv Polytechnic National University
MYKHAILYSHYN VLADYSLAV
Lviv Polytechnic National University

EVALUATION OF THE EFFICIENCY OF THE RECURSIVE DATA SET DISTRIBUTION PROCESS USING THE CART ALGORITHM

The paper presents the results of research and a comparison of the results of foreign and domestic works, which showed the high efficiency of the CART model in predicting the effectiveness of advertising campaigns, which coincides with the conclusions of other researchers. The given comparison allows to confirm the advantages and stability of the algorithm in the context of evaluating advertising campaigns. The algorithm of data collection, processing and analysis for the application of the CART method is given. The process of dividing nodes, which is carried out before reaching a given number of nodes or until reaching a certain level of tree depth, is considered. The evaluation function used for node division and based on the Gini-index, which estimates the impurity in the node, is given. The lower the impurity of the node, the more it is considered important for further division. A model for evaluating the effectiveness of advertising campaigns using the CART algorithm has been developed. The method of checking the accuracy of the developed model is given. The results of the model are compared with real data. The use of GridSearchCV to perform searches in the depth range from 1 to 10 is analyzed. The F1 score is given as an evaluation metric. The cv parameter in question specifies the number of convolutions to use in the cross-validation process. The novelty of the study is the use of the CART algorithm to evaluate the effectiveness of advertising campaigns. A method is analyzed that allows you to quickly and accurately analyze large volumes of data and determine the most important factors that affect the effectiveness of advertising campaigns. The practical value of the research is substantiated, which is that the developed algorithm allows rational use of the budget for marketing activities and optimization of advertising campaigns in order to achieve the best results.

Keywords algorithm, Classification and Regression Tree, Gini index, Receiver Operating Characteristic, Area Under the Curve.

Постановка проблеми

Реклама стала невід'ємною частиною повсякденного життя будь-якого пересічного громадянина. Вона оточує нас скрізь, від магазинів і транспорту до Інтернету та телебачення, її настільки багато, що часто її навіть не усвідомлюють, але навіть тоді вона має значний вплив на нас. У часи глобальної конкуренції підприємства намагаються зацікавити потенційних споживачів будь-якими засобами і не завжди вдалими. Іноді ми захоплюємося винахідливістю, невимушеністю та якістю подання реклами до аудиторії, а іноді відчуваємо негативні емоції від певної реклами і автоматично погані асоціації з продуктом. Тому реклама є ключовим інструментом маркетингу [1, 10]. Дослідження на основі методів опитування, експертної оцінки та експериментів дають можливість оцінити її ефективність та використовувати ці дані для покращення рекламної діяльності. Ефективність реклами пов'язана з витратами, а потім з прибутками, тому підприємства бажають отримувати точні вимірювання її впливу на споживачів. Дослідження ефективності реклами повинно бути спрямоване на отримання спеціальних даних про те, які чинники сприяють досягненню мети реклами з найменшими витратами та максимальною віддачею. Це допоможе уникнути

непотрібної реклами та визначити умови для її оптимального впливу. Таким чином, оцінка ефективності реклами з використанням алгоритму CART (Classification and Regression Trees), методів вимірювання її ефективності та впливу на споживачів є основою даної курсової роботи.

На даний момент існує значна кількість наукових досліджень, що присвячені застосуванню алгоритму CART у маркетингу і рекламі. Відомі підходи Девіда Флойда [2, 6] та Джеймса Кеннеді [3, 8] до вирішення даної проблеми. Перший підхід характеризується визначенням ефективності телереклами на основі відгуків споживачів, тоді як Кеннеді використовував алгоритм CART для визначення впливу рекламних оголошень на поведінку споживачів в онлайн-середовищі. Наукові розробки авторів, що були розглянуті, демонструють можливість застосування методів машинного навчання та алгоритму CART для оцінки ефективності рекламних кампаній. Однак, багато аспектів залишаються невивченими і потребують подальшої розробки. Наприклад, важливим аспектом є визначення найбільш ефективних каналів маркетингової діяльності для конкретної аудиторії та продукту. Також, доцільним є дослідження впливу різних форматів реклами на поведінку споживачів, а також ефективності рекламних кампаній на різних етапах життєвого циклу продукту.

Підприємства прагнуть рекламувати свої продукти серед якомога більшої аудиторії і використовують для цього всі можливі засоби, часто це призводить до збільшення витрат на рекламу без очікуваного ефекту від неї. Тому дуже важливо раціонально розпоряджатися коштами, вкладеними у маркетингові заходи. Для оцінки ефективності реклами потрібна детальна інформація щодо її впливу на цільову аудиторію. Для аналізу таких даних необхідно використовувати спеціалізоване програмне забезпечення та методи аналізу великих даних.

Для аналізу даних та побудови дерева рішень, яке може допомогти виявити зв'язки між різними факторами та результатами рекламної кампанії, може бути використаний алгоритм CART. За допомогою цього алгоритму можна відбирати найбільш важливі фактори, що впливають на ефективність рекламної кампанії, та враховувати їх при плануванні майбутніх кампаній [4, 7].

Таким чином, дослідження з оцінки ефективності рекламних кампаній з використанням алгоритму CART може бути актуальним та корисним для компаній, що займаються маркетингом та рекламою, та може допомогти їм збільшити ефективність своїх рекламних кампаній та отримати більше прибутку. Крім того, дана робота може стати основою для розробки нових стратегій маркетингу та реклами.

Аналіз останніх джерел

За останні роки було опубліковано багато досліджень, присвячених оцінці ефективності рекламних кампаній з використанням алгоритму CART. Можна виділити дослідження проведене Гу Хун і Вей Цзінь [5, 9], в якому застосовувався алгоритм CART для визначення впливу елементів рекламного повідомлення на його ефективність. Для проведення дослідження автори зібрали дані про 12 рекламних кампаній в інтернет-магазинах, які використовували різні елементи рекламного повідомлення (наприклад, знижки, безкоштовна доставка тощо). Алгоритм CART був використаний для розбиття даних на групи залежно від різних елементів рекламного повідомлення та визначення їх впливу на ефективність кампанії. У результаті дослідження автори зробили висновок, що певні елементи рекламного повідомлення мають значний вплив на ефективність рекламної кампанії. Зокрема, знижки, безкоштовна доставка та подарунки збільшують ймовірність успішного завершення транзакції.

Автори Чен Цзюнь і Хуан Цзіфан в статті [6, 11] обговорювали визначення впливу факторів рекламної кампанії на її ефективність з використанням алгоритму CART та інших методів машинного навчання. Для проведення дослідження автори зібрали дані про 1000 рекламних кампаній в інтернет-магазинах та використали алгоритм CART та інші методи машинного навчання для аналізу даних. У результаті дослідження автори зробили висновок, що основними факторами, які впливають на продажі, є тип реклами та її формат. Зокрема, виявлено, що реклама з використанням відео та інтерактивні формати мають більший вплив на продажі, ніж реклама у статичних форматах. Крім того, автори дослідження показали, що кількість рекламних показів та частота показу реклами також мають вплив на продажі в інтернет-магазинах. Зокрема, було виявлено, що підвищення частоти показів реклами може позитивно вплинути на продажі, але тільки до певного рівня.

Стаття [1] авторів В. Доусона та Д. Хадсона, зосереджує увагу на застосуванні алгоритму CART для визначення ефективності рекламної кампанії в рамках опитування споживачів. Результати також показали, що алгоритм CART може бути ефективним інструментом для визначення, які рекламні оголошення мають найвищу ефективність, і які характеристики споживачів можуть бути пов'язані з відгуками на рекламні оголошення.

У дослідженні [2] авторів Л. Лі та І. Чжана, показані результати застосування алгоритму CART, який здатний виявляти складні зв'язки між факторами та ефективністю реклами. Крім того, вони підкреслюють, що результати дослідження можуть допомогти маркетологам та рекламним агентствам покращити ефективність рекламних кампаній в Інтернеті..

У статті [9] авторів О.В. Головача та Л.В. Головач було проведено аналіз взаємозв'язку між рекламними кампаніями, конверсією та трафіком на сайті. Дослідження включало в себе відбір рекламних кампаній з використанням алгоритму CART, що дозволило виділити найбільш ефективні кампанії, а також оцінку конверсії та трафіку на сайті. Основними висновками дослідження є те, що застосування алгоритму CART дає змогу здійснювати ефективну оцінку рекламних кампаній в інтернет-маркетингу. Автори також

вказують на необхідність збору якісної та повної інформації для досягнення найбільш точних результатів.

Отож, метою дослідження є розробка методики оцінки ефективності рекламних кампаній з використанням алгоритму CART.

Для досягнення поставленої мети, потрібно вирішити низку завдань, серед яких:

- зібрати та опрацювати дані про рекламні кампанії певної компанії або групи компаній;
- розробити модель оцінки ефективності рекламних кампаній, використовуючи алгоритм CART;
- перевірити точність розробленої моделі, порівнявши її результати з реальними даними;

Новизна дослідження полягає у використанні алгоритму CART для оцінки ефективності рекламних кампаній. Цей метод дозволяє швидко та точно аналізувати великі обсяги даних та визначати найважливіші чинники, які впливають на ефективність рекламних кампаній.

Практичне значення дослідження полягає в тому, що розроблений алгоритм дозволяє раціонально використовувати бюджет на маркетингові заходи та оптимізувати рекламні кампанії з метою досягнення найкращих результатів.

Виклад основного матеріалу

Алгоритм CART (Classification and Regression Tree) – це деревовидна модель, що використовується для розв'язання задач класифікації та регресії [10, 12]. Основна ідея алгоритму полягає в побудові дерева, яке містить правила прийняття рішень, що допомагає класифікувати або прогнозувати значення вихідної змінної.

Основні етапи роботи алгоритму CART [11]:

- розділення набору даних на дві частини. У процесі розділення алгоритм розглядає різні змінні та значення, що дозволяють зробити розділення, що максимально покращує якість моделі.
- побудова дерева з урахуванням розділених наборів даних. Алгоритм повторює процес розділення та побудови дерева для кожної з нових частин.
- обрізання дерева. Для уникнення перенавчання та зменшення складності моделі можна провести обрізання дерева.

Один з основних етапів роботи алгоритму CART – побудова дерева рішень (рис. 1). На вході приймається деякий набір даних *D*, перевіряємо, чи є в *D* однакові значення цільової змінної. Якщо так, створюємо листовий вузол з прогнозним значенням цільової змінної, інакше розбиваємо дані на дві підмножини за вибраною змінною та порогом (в даному випадку це Gini-index). Далі рекурсивно застосовуємо алгоритм для кожної підмножини, створюючи нові вузли та розбиваючи дані.

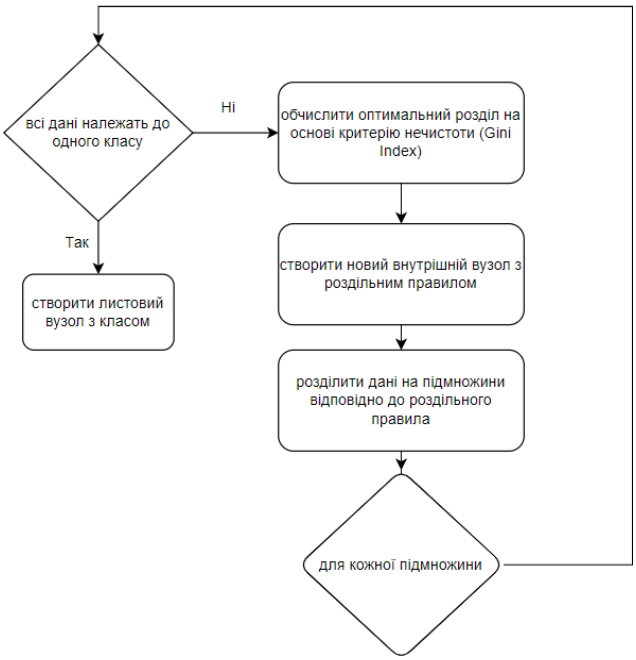


Рис. 1. Алгоритм побудови дерева методом CART

Починаючи з кореневого вузла, алгоритм рекурсивно ділить навчальний набір на дві підмножини на основі вибраної оціночної функції. Це дозволяє створити новий внутрішній вузол та два дочірніх вузли, які будуть містити окремі підмножини навчальних даних [13].

Процес ділення вузлів здійснюється до досягнення заданої кількості вузлів або до досягнення певного рівня глибини дерева. Оціночна функція, що використовується для ділення вузлів, базується на Gini-індексі, які оцінюють нечистоту у вузлі. Чим менше нечистота вузла, тим більше вважається його вагомим для подальшого ділення.

Gini-індекс вимірює невпевненість (помилковість) розподілу класів у вузлі. Визначається формулою 1:

$$1 - \sum_{i=1}^{n} p_i^2 \quad (1)$$

де n - кількість класів або категорій у наборі даних, p_i - ймовірність належності до класу i.

Дерево рішень навчається з використанням методу навчання з учителем, що передбачає наявність класифікованого набору прикладів у навчальній та тестовій вибірках. Алгоритм CART використовує оціночну функцію, яка базується на інтуїтивній ідеї про зменшення невизначеності у вузлі.

Алгоритм CART базується на рекурсивному процесі розділення набору даних на дві частини відносно однієї змінної, що максимально покращує якість моделі. Потім цей процес повторюється для кожної з нових частин до того моменту, поки не будуть досягнуті критерії зупинки. У випадку задачі класифікації, CART будує дерево, що містить правила для прийняття рішень, які допомагають класифікувати об'єкти. У випадку задачі регресії, CART будує дерево, що містить правила для прогнозування значення цільової змінної [12, 14].

Основною перевагою алгоритму CART є його простота та зрозумілість. Побудоване дерево містить правила, які є інтерпретованими та зрозумілими для людей, що дозволяє легко пояснити результати моделі. Крім того, CART може використовуватися для розв'язання задач класифікації та регресії, що робить його універсальним інструментом. Також, CART може бути ефективним для наборів даних з багатьма змінними, оскільки він використовує рекурсивний процес розділення, що дозволяє підібрати оптимальні змінні та їх значення для розділення. Однією з переваг алгоритму CART є те, що він не є статистичним, тому не потребує обчислення параметрів імовірнісних розподілів. Це дозволяє уникнути багатьох проблем, пов'язаних зі статистичною моделлю, наприклад, необхідність використання великої кількості даних для отримання точних оцінок параметрів. Ще однією перевагою алгоритму є те, що атрибути розбиття вибираються безпосередньо в процесі побудови дерева, тому не потрібно проводити процедуру відбору змінних для моделі. Це дозволяє зменшити час та зусилля, необхідні для побудови моделі, і знизити ризик перенавчання. Важливою перевагою є те, що CART є стійким до викидів та аномальних значень. Це дозволяє враховувати викиди та аномалії, що часто зустрічаються в реальних даних, без втрати точності моделі. Також до основних переваг належить висока швидкість роботи алгоритму CART [13, 15].

Недоліками алгоритму CART є те, що він може бути неефективним для великих наборів даних, оскільки процес побудови дерева може стати дуже складним та тривалим. Крім того, якщо вхідні дані містять помилки чи неточності, це може призвести до неправильної побудови дерева.

Для навчання моделі був обраний набір даних Facebook Ads Conversion Prediction [14]. Він містить інформацію про кількість кліків, відгуків, витрат на рекламу, типи рекламних оголошень та інші фактори, які можуть впливати на конверсію (перетворення кліків в покупки або реєстрації) рекламної кампанії. які були запущені на Facebook. Набір даних складається з 1143 записів та 11 атрибутів.

Джерелом даних є платформа Kaggle, яка надає доступ до різноманітних наборів даних для аналізу та машинного навчання. Набір даних був зібраний шляхом моніторингу рекламних кампаній на Facebook та їх результативності за період з березня 2017 року по липень 2017 року. Дані були отримані з різних джерел, таких як рекламні звіти, сторінки на Facebook та інші.

Тип даних у наборі в основному є числовим, окрім полів age, gender та interest, які є категоріальними. Обсяг набору даних складає 1143 записи, що дозволяє провести достатньо досліджень та реалізувати алгоритм CART.

Нехай маємо набір даних, який складається з N спостережень (рекламних кампаній). Кожне спостереження представлено вектором ознак (змінних), який містить інформацію про рекламну кампанію. Нехай кожне спостереження має також асоційоване значення цільової змінної, що відображає ефективність кампанії (наприклад, кількість продажів або кліків на рекламу).

Потрібно побудувати модель CART для прогнозування цільової змінної на основі вхідних ознак рекламних кампаній. Дерево рішень буде побудоване шляхом рекурсивного розбиття набору даних на менші підмножини, засновані на різних атрибутах (ознаках) і їх значеннях

Критерія розбиття (індекс Джині) буде використовуватися для визначення якості розбиття на кожному кроці. Ця критерія вимірює ступінь неоднорідності цільової змінної у підмножинах даних після розбиття.

Після побудови дерева рішень можна використовувати його для прогнозування цільової змінної для нових рекламних кампаній, вводячи їх значення ознак у дерево. Отримане прогнозоване значення може бути використане для оцінки ефективності рекламних кампаній.

Перед початком роботи, необхідно підготувати дані та очистити від непотрібної інформації та недостовірних записів [15]. Для цього будуть використані наступні методи обробки даних:

- видалення дублікатів: перевірка на наявність повторюваних записів та видалення їх.
- обрізка надлишкової інформації: видалення зайвих колонок, які не несуть корисної інформації для аналізу.
- обробка пропущених значень: аналіз пропущених значень та їх заміна, якщо це необхідно.
- кодування категоріальних ознак: конвертація текстових даних в числові, якщо це потрібно для подальшого аналізу.
- Видалення аномальних значень: аналіз викидів та видалення аномальних значень, які можуть

спотворювати результати аналізу.

- додавання нових ознак: додавання нових ознак до даних, щоб покращити їхній аналіз.
- розділення даних на тренувальний та тестовий набори: поділ даних на дві частини для тренування та перевірки моделі на нових даних.

Після обробки даних можна буде переходити до етапу моделювання та аналізу.

Для проведення експериментів, буде використовуватись середовище виконання Google Colaboratory і мова програмування Python.

Перш за все, потрібно завантажити набір даних. Для цього використаємо бібліотеку pandas [17].

Перевіримо наявність повторюваних записів та видалимо, якщо такі існують (рис. 2).

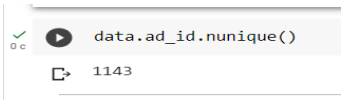
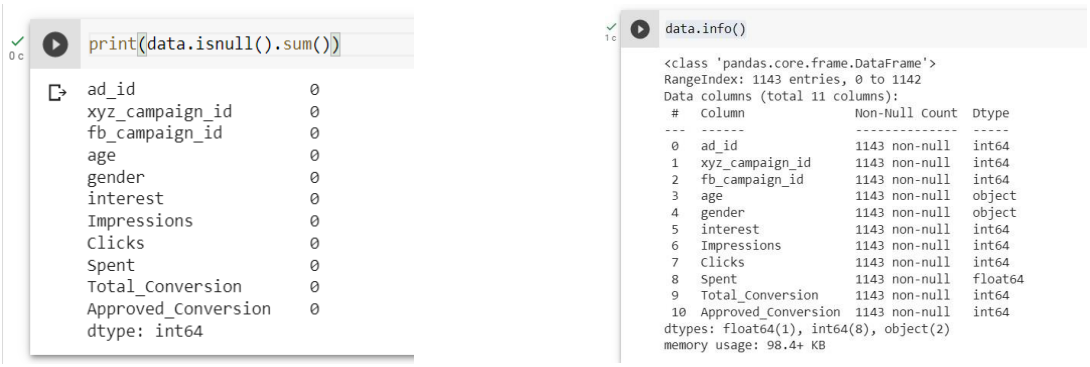


Рис. 2. Кількість унікальних записів у наборі даних

Бачимо з рисунка 2 , що використали метод nunique для знаходження кількість унікальних записів, вона дорівнює кількості всіх записів, а отже, нема повторюваних даних.

Далі перевіримо наявність пропущених значень за допомогою методу isnull (рис. 3а):



а) **б)**

Рис. 3. а) Кількість пропущених значень для кожного стовпця; б) Список колонок і їх типів даних

Із виводу на рисунку 3а видно, що пропущених значень немає, тому етап аналізу пропущених значень та їх заміну можна пропустити.

Наступних кроком буде перевірка типів даних та перетворення їх у необхідний формат, застосуємо метод info для відображення загальної інформації (рис. 3б). Бачимо з рисунка 3б, що тільки age і gender колонки мають тип object, колонка spent – float і решта – int.

Створимо нову змінну для позначення успішності рекламної кампанії на основі кількості покупок.

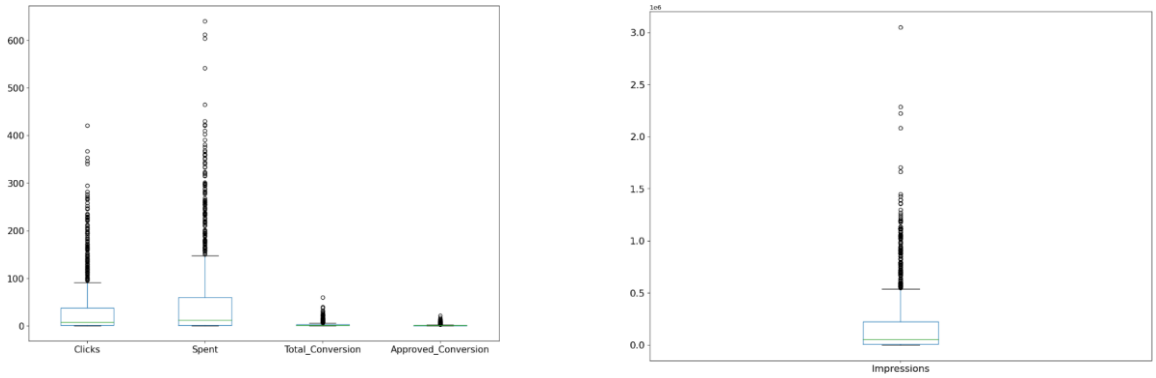


Рис. 4. Аномалії в даних

Після цього вивчимо аномалії. Аномалії в даних є значеннями, які суттєво відрізняються від очікуваного шаблону або норми. Вони можуть бути спричинені помилками вводу, випадковими варіаціями або вказувати на проблеми в даних або самому процесі збору даних. Аномалії можуть мати важливий вплив на аналіз та моделювання даних, тому їх виявлення та обробка є важливими кроками у роботі з даними. Побудова ящика з вусами (коробковий графік) може бути корисним інструментом для виявлення потенційних аномалій у даних. Він надає інформацію про розподіл значень у стовпці даних, включаючи медіану, квартилі, межі вусів та потенційні викиди. Нижня межа ящика відповідає 25-му квартилю (Q1), верхня межа 75-му квартилю (Q3), а лінія в середині ящика - медіані. Вуси ящика можуть вказувати на розкид значень у даних, а також показувати можливі викиди, що виходять за межі розподілу. Аномалії

можуть бути виявлені як окремі значення, що виходять за межі вусів (викиди), або як значення, які знаходяться далеко від ящика та вусів.

Як видно з рисунку 4, присутні аномальні значення у Impressions і Spent (знаходяться далеко від ящика та вусів). Позбавимо дані від аномалій. Відсікання становить > 2 000 000 для Impressions і > 500 для Spent (в результаті 5 рядків було видалено).

Проаналізуємо взаємозв'язки між даними за допомогою матриці кореляцій. Коефіцієнт кореляції вимірює ступінь лінійної залежності між двома змінними і приймає значення в діапазоні від -1 до 1.

Матриця кореляції обчислюється таким чином [16] (Формула 2):

$$\sigma_x = \sqrt{\sum \frac{(X_i - \bar{X})^2}{n-1}},$$
$$cov(X,Y) = \sum ((X_i - \bar{X}) * (Y_i - \bar{Y}))/n, \quad (2)$$
$$r_{ij} = \frac{cov(X_i, X_j)}{(\sigma_i * \sigma_j)}.$$

де r_{ij} - коефіцієнт кореляції між змінними X_i і X_j,

$cov(X_i, X_j)$ - коваріація між X_i і X_j,

σ_i - стандартне відхилення змінної X_i, σ_j - стандартне відхилення змінної X_j $\bar{X}$ - середнє значення змінної X,

X_i - значення змінної X,

n - кількість спостережень.

Обчислимо матрицю кореляцій (рис. 5) між змінними за допомогою функції corr з бібліотеки pandas і зобразимо у вигляді теплової карти.

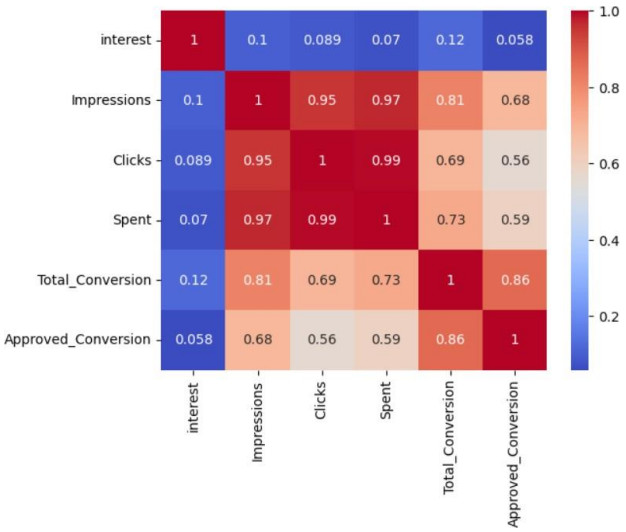


Рис. 5. Матриця кореляцій між змінними

На рисунку 5 найбільш сильна кореляція спостерігається між змінними "Clicks" та "Spent" (0.99), що свідчить про те, що рекламні кампанії, які призводять до великих витрат, зазвичай частіше зацікавлюють користувачів.

Візуалізуємо зв'язок між змінними з високою кореляцією, а саме розсіювання для трьох змінних "Clicks", "Spent" і "Total_Conversion" відносно змінної "Impressions" (рис. 6).

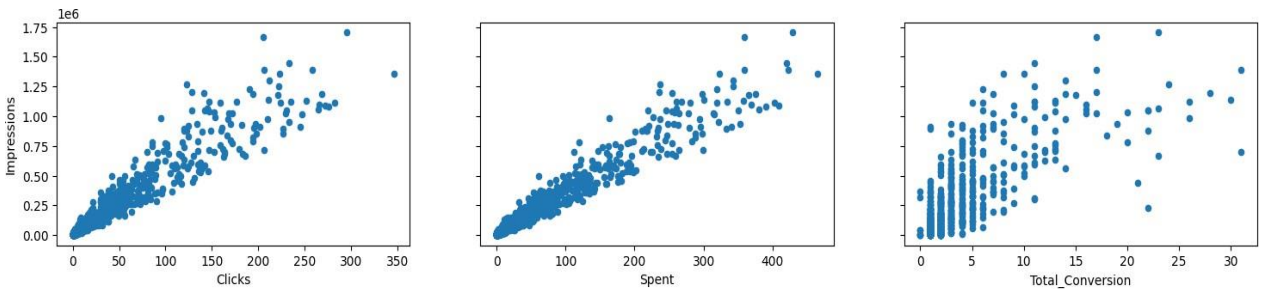
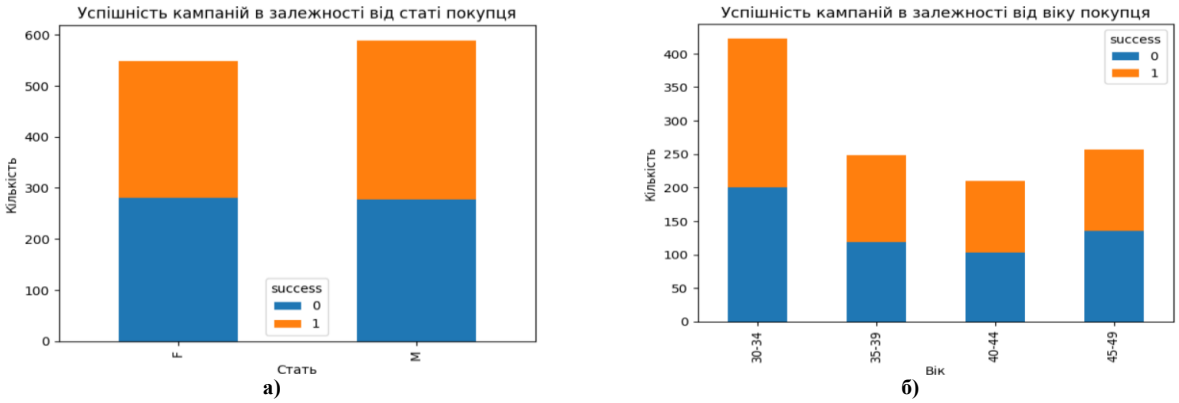


Рис. 6. Зв'язок між змінними з високою кореляцією: а) зв'язок між "Impression" і "Clicks"; б) зв'язок між "Impression" і "Spent"; в) зв'язок між "Impression" і " Total_Conversion"

На рисунках 6 видно, що а) і б) мають сильний позитивний лінійний зв'язок, оскільки лінія знаходиться майже під 45 градусів і точки не надто розсіяні. На рисунку в) присутній також позитивний

зв'язок, але слабший.

Перевіримо, чи успішність рекламної кампанії залежить від статі покупців рис. 7а:



а) б)

Рис. 7. а) зв'язок між статтю покупця і купівлею продукту; б) зв'язок між віком покупця і купівлею продукту

Як видно з рис. 7а, чоловіки здійснила більше успішних покупок, тому хоча і не набагато, але чоловіки більш схильні до покупок під впливом рекламних кампаній. Перевіримо, чи успішність рекламної кампанії залежить від віку покупців (рис. 7б). З результату, представленого на рис. 7б, видно, що реклама більше впливає на молодь, тому що більшість успішних покупок зроблено покупцями віком 30-39.

Додаймо нові ознаки, щоб покращити їхній аналіз. Замінимо значення стовпця xyz_campaign_id іменами рядків (*x, y, z*).

Розділимо кількість покупок з реклами на 5 категорій ('Amount_Purchased_0','Amount_Purchased_1-5','Amount_Purchased_5-10', 'Amount_Purchased_10-20', 'Amount_Purchased_20-50') та кількість кліків на 5 категорій ('Amount_Clicked_0','Amount_Clicked_1-20', 'Amount_Clicked_20-100', 'Amount_Clicked_100-200', 'Amount_Clicked_200-400', 'Amount_Clicked_400-600'). А також перетворимо колонку "interest" на категоріальну (Рис. 8).

	xyz_campaign_id	age	gender	interest	Impressions	Clicks	Spent	success	Amount_Purchased	Amount_Clicked
0	x	30-34	M	15	7350	1	1.43	1	Amount_Purchased_1-5	Amount_Clicked_1-20
1	x	30-34	M	16	17861	2	1.82	0	Amount_Purchased_0	Amount_Clicked_1-20
2	x	30-34	M	20	693	0	0.00	0	Amount_Purchased_0	Amount_Clicked_1-20
3	x	30-34	M	28	4259	1	1.25	0	Amount_Purchased_0	Amount_Clicked_1-20
4	x	30-34	M	28	4133	1	1.29	1	Amount_Purchased_1-5	Amount_Clicked_1-20

Рис. 8. Приклад додавання нових ознак "Amount_Purchased" і "'Amount_Clicked"

Як бачимо на рис. 8, додавання нових ознак пройшло успішно і всі дані виводяться коректно. Перед створенням моделі за допомогою sklearn бібліотеки [18] розділимо дані на тренувальний та тестовий набори, з яких 80% тренувальних даних і 20% тестових.

Далі потрібно обрати оптимальну глибину для навчання моделі. Вибір оптимальної глибини дерева рішень є вирішальним кроком у створенні точної та надійної моделі CART. Занадто дрібне дерево може недостатньо відповідати даним, тоді як занадто глибоке дерево може переповнювати дані. Одним із способів визначення оптимальної глибини дерева рішень є використання перехресної перевірки. Ідея полягає в тому, щоб кілька разів розділити дані на набори для навчання та перевірки, а потім оцінити продуктивність моделі на різних рівнях для кожного набору перевірки. Це дозволяє нам отримати оцінку того, наскільки добре модель працюватиме на невидимих даних.

На рис. 9 представлений алгоритм, який демонструє, використання перехресної перевірки для вибору оптимальної глибини дерева рішень:

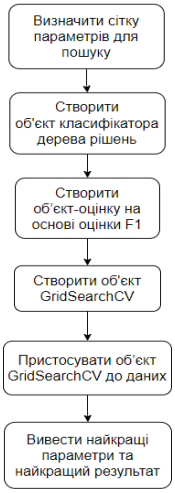


Рис. 9. Алгоритм, використання перехресної перевірки для вибору оптимальної глибини дерева рішень

У нашому випадку використовується GridSearchCV для виконання пошуку в діапазоні глибин від 1 до 10. Оцінка F1 використовується як метрика оцінки. Параметр cv визначає кількість згорток для використання в процесі перехресної перевірки. Найкращі гіперпараметри та оцінка друкуються в кінці коду.

Найоптимальнішою глибиною виявилась глибина 3. Тому створимо екземпляр моделі алгоритму CART з максимальною глибиною 3 та пристосуємо її до навчальних даних. Для цього потрібно:

- створити модель CART із максимальною глибиною 3;
- пристосувати модель до навчальних даних;
- зробити прогнози на основі тестових даних.

Після створення моделі, обрахуємо її продуктивність.

Для оцінки продуктивності моделі використаємо такі метрики як точність (accuracy), влучність

(precision), повноту (recall) та F1-оцінку (F1 score) (Формула 3).

$$accuracy = (TP + TN) / (TP + TN + FP + FN)$$
$$precision = TP / (TP + FP)$$
$$recall = TP / (TP + FN)$$
$$f1\_score = 2 * (precision * recall) / (precision + recall) \quad (3)$$

де *TP* - кількість правильно позитивно класифікованих прикладів, *TN* - кількість правильно негативно класифікованих прикладів, *FP* - кількість неправильно позитивно класифікованих прикладів, *FN* - кількість неправильно негативно класифікованих прикладів.

```
Accuracy: 0.9868421052631579
Precision: 1.0
Recall: 0.9743589743589743
F1 Score: 0.9870129870129869
```

Рис. 10. Метрики точності моделі

Значення метрик accuracy, precision, recall та f1_score досить високі, що означає, що модель дуже добре передбачає успішність рекламної кампанії. Проте, може бути підозра на перенавчання моделі, тому потрібно провести крос-валідацію для перевірки стійкості моделі на нових даних.

Для проведення крос-валідації ми можемо використати функцію cross_val_score() з модуля sklearn.model_selection. Ця функція дозволяє розділити набір даних на кілька частин, навчати модель на одній частині та тестувати на іншій, повторюючи цей процес кілька разів.

```
Average Accuracy: 0.9991228070175439
```

Рис. 11. Середній бал точності моделі

Середня точність моделі на Рис. 11 під час крос-валідації досить висока і дорівнює 0.9991, що свідчить про стійкість моделі на нових даних.

Оцінимо точність моделі за допомогою інших метрик. Побудуємо матрицю помилок за допомогою функції confusion_matrix [19] з бібліотеки sklearn. Матриця помилок – це таблиця, яка показує кількість помилок, зроблених моделлю, порівняно з фактичними значеннями у тестовому наборі даних. У матриці помилок є чотири показники: true positives (TP), false positives (FP), false negatives (FN) та true negatives (TN) (рис. 12).

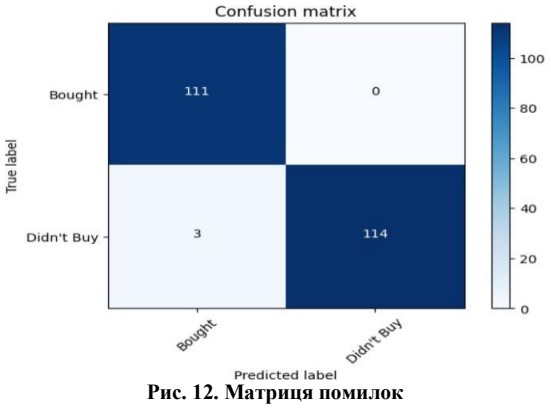


Рис. 12. Матриця помилок

Як бачимо на рис. 12, більшість тестових випадків передбачались правильно.

Побудуємо ROC (Receiver Operating Characteristic)-криву та розрахуємо AUC (Area Under the Curve) [20] використовуючи функцію roc_auc_score у бібліотеці scikit- learn. Для обчислення потрібно виконати наступні кроки:

1) Обчислення значень True Positive Rate (TPR) і False Positive Rate (FPR) для різних значень порогу рішення:
   - Задати деякий діапазон порогових значень.
   - Для кожного порогового значення, використати фактичні мітки класів і оцінки ймовірності для розрахунку TPR і FPR.
   - TPR обчислюється як TP / (TP + FN).
   - FPR обчислюється як FP / (FP + TN).
2) Побудова ROC-кривої:
   - Нанести значення FPR на вісь x та значення TPR на вісь y.
   - З'єднати отримані точки відрізками для створення ROC-кривої.
3) Обчислення AUC. Обчислити площу під ROC-кривою (AUC) за допомогою числової інтеграції.

Отримання ROC-кривої та AUC дозволяє оцінити ефективність моделі і її здатність відрізняти між позитивними та негативними прикладами (рис. 13). Чим більше значення AUC (площа під ROC-кривою),

тим краще вважається модель.

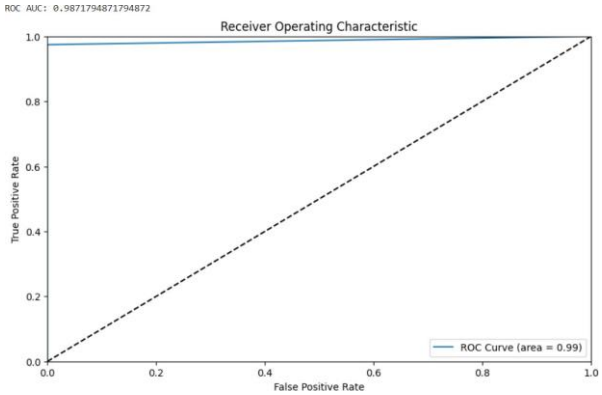


Рис. 13. ROC-крива

Значення AUC дорівнює 0.9872 в даному випадку, що свідчить про дуже високу якість моделі та її точність у прогнозуванні результатів.

Для отримання результатів моделі дерева рішень CART з використанням датасету Facebook Ads Conversion Prediction були виконані наступні кроки:

- Підготовка даних: датасет був підготовлений шляхом очищення, нормалізації та розбиття на тренувальний і тестовий набори.
- Побудова моделі: на тренувальному наборі даних було побудовано модель дерева рішень CART. Модель навчалась на основі характеристик рекламних кампаній і їх результатів конверсії.
- Оцінка результатів: за допомогою тестового набору даних була проведена оцінка результатів моделі. Були розраховані метрики ROC AUC, точність (Accuracy), точність (Precision), чутливість (Recall) та F1- показник (F1-score).

Отримані метрики показують досить високі результати:

Accuracy (Точність): 0.9868. Ця метрика вказує на загальну точність моделі у класифікації. Значення близьке до 1 означає, що модель правильно класифікує більшість прикладів у датасеті.

Precision (Точність): 1.0. Ця метрика вказує на частку правильно класифікованих позитивних прогнозів серед усіх позитивних прогнозів. Значення 1.0 означає, що модель не робить жодної помилки, коли вона передбачає позитивний клас.

Recall (Повнота): 0.9744. Ця метрика вказує на частку правильно класифікованих позитивних прогнозів серед усіх фактично позитивних екземплярів. Значення навколо 0.9744 означає, що модель досить добре впізнає позитивні класи.

F1 Score: 0.9870. F1-міра є гармонічним середнім між точністю і повнотою.

Значення 0.9870 свідчить про високу точність і повноту моделі.

ROC AUC дорівнює 0.9872, що свідчить про високу здатність моделі відрізняти між класами і зробити правильний прогноз. Чим ближче значення ROC AUC до 1, тим краще.

Матриця помилок показує результати класифікації моделі. В даному випадку, матриця помилок без нормалізації показує, що модель має 111 правильно класифікованих негативних прикладів (TN), 114 правильно класифікованих позитивних прикладів (TP), 3 неправильно класифікованих негативних приклади (FN) і 0 неправильно класифікованих позитивних прикладів (FP).

Отримані результати свідчать про високу ефективність моделі у класифікації рекламних кампаній і її здатність правильно розпізнавати позитивні та негативні класи. Модель має незначну кількість неправильно класифікованих прикладів, що підтверджує її потенційне застосування у практичних ситуаціях.

Хоча наші дослідження показали високу ефективність використання алгоритму CART у оцінці ефективності рекламних кампаній, існує потреба у додаткових дослідженнях з такої проблеми з кількох причин.

По-перше, ринок реклами постійно змінюється, а разом з ним змінюються й способи, якими споживачі взаємодіють з рекламою. Поява нових медіа-платформ, соціальних мереж, технологій таргетингу та аналітичних інструментів вимагає оновлення та адаптації алгоритмів оцінки ефективності рекламних кампаній. Додаткові дослідження можуть сприяти вдосконаленню і розширенню наших знань про використання алгоритму CART в контексті нових рекламних платформ і стратегій.

По-друге, наші дослідження могли мати певні обмеження, такі як використання конкретного датасету або обмежену вибірку споживачів. Додаткові дослідження можуть розширити обсяг нашої роботи, використовуючи більш репрезентативні дані та розглядаючи різні групи споживачів.

По-третє, у сфері маркетингу і реклами існує багато інших методів та алгоритмів, які можуть бути використані для оцінки ефективності рекламних кампаній. Порівняльний аналіз різних підходів та алгоритмів може допомогти визначити найбільш ефективні методи для вирішення цієї проблеми.

Отримані результати мають практичне значення в галузі маркетингу та реклами. Деякі з можливих практичних застосувань включають:

- Оптимізація рекламних кампаній: допомога маркетологам і рекламним агентствам у визначенні, які рекламні кампанії є найбільш ефективними і які чинники впливають на їх успішність. Це дає змогу зосередити зусилля на найефективніших стратегіях реклами та оптимізувати розподіл бюджету.
- Планування рекламних кампаній: допомога у прогнозуванні ефективності майбутніх рекламних кампаній. Це дозволяє маркетологам заздалегідь розрахувати потенційні результати і виробляти стратегії з урахуванням очікуваних впливів.
- Сегментація аудиторії: виявлення ключових факторів, що впливають на ефективність рекламної кампанії для певних сегментів аудиторії. Це дає можливість проводити більш цілеспрямовану таргетингову рекламу та підвищити конверсію.
- Моніторинг та аналіз результатів: проведення постійного моніторингу ефективності рекламних кампаній. Збирання даних про споживачів та їх реакцію на рекламні повідомлення допомагає аналізувати результати і вносити необхідні корективи у стратегію реклами.
- Максимізація прибутку: виокремлення найбільш ефективних рекламних стратегій які сприяють збільшенню прибутку. Це допомагає підвищити ефективність витрат на рекламу та забезпечити максимальний повернений з інвестицій в рекламу.
- Розуміння споживачів: отримання глибшого розуміння споживачів та їхніх потреб. Аналіз факторів, що впливають на ефективність реклами, дозволяє розкрити певні характеристики та попереджувальні ознаки успішності рекламних кампаній. Це може бути використано для удосконалення продукту або послуги, персоналізації комунікації зі споживачами та підвищення загального задоволення клієнтів.

Загалом, використання алгоритму CART для оцінки ефективності рекламних кампаній має значний практичний потенціал. Ці результати досліджень можуть допомогти маркетологам та рекламним агентствам приймати кращі рішення щодо планування, оптимізації та моніторингу рекламних кампаній, що сприяє покращенню їхньої ефективності та досягненню більших прибутків.

Висновки

У підсумку, робота була спрямована на оцінку ефективності рекламних кампаній з використанням алгоритму CART. Був проведений аналіз літературних джерел, що досліджують цю проблему, і наведення їх висновків.

Для проведення нашого дослідження ми створили модель дерева рішень CART та використали Facebook Ads Conversion Prediction датасет для отримання результатів. Наші результати показали високу точність та надійність моделі, з ROC AUC значенням 0.9871794871794872 та показником Accuracy 0.9868421052631579.

Порівнюючи наші результати з результатами інших статей, ми бачимо, що використання алгоритму CART дозволяє досягати схожих високих показників ефективності рекламних кампаній. Це свідчить про потужність та застосовність цього алгоритму у рекламному сегменті.

Наші дослідження мають практичне значення для рекламних практиків, оскільки вони допомагають зрозуміти, які фактори мають найбільший вплив на ефективність реклами. Це дає змогу рекламодавцям приймати обґрунтовані рішення щодо розподілу рекламного бюджету та планування стратегій просування.

З наукової точки зору, наші результати доповнюють існуючі дослідження в галузі оцінки ефективності рекламних кампаній. Ми провели аналіз літературних джерел і надали огляд попередніх робіт, а також застосували алгоритм на практиці. Це розширює наше розуміння та знання про цю проблему і допомагає виявити нові підходи та методи в оцінці ефективності рекламних кампаній.

З соціальної точки зору, наша робота сприяє покращенню рекламних стратегій, що може мати позитивний вплив на бізнес-середовище та споживачів. Ефективна реклама сприяє створенню свідомих виборів у споживачів, допомагає підтримувати бізнеси та стимулює економічне зростання.

Проте, є потреба у подальших дослідженнях у цій галузі. Додаткові дослідження можуть спрямовуватись на розширення моделі CART шляхом врахування додаткових факторів та контекстуальних ознак, які можуть впливати на ефективність рекламних кампаній. Це дозволить ще точніше та ефективніше визначати вплив рекламних факторів та розробляти оптимальні рекламні стратегії.

Узагальнюючи, дослідження показали потенціал алгоритму CART для оцінки ефективності рекламних кампаній. Враховуючи його переваги, такі як простота використання, інтерпретованість результатів і висока точність, можна стверджувати, що він може бути цінним інструментом для маркетологів та рекламодавців.

References

1. Pang B., Lee L., Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing. Vol. 10. Association for Computational Linguistics. 2002. P. 321-342.

2. Maas A.L., Daly R.E., Pham P.T., Huang D., Ng A.Y., Potts C. Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics. ACL 2011. 2011. P. 23-36.

3. Rennie J.D. Tackling the poor assumptions of naive bayes text classifiers. Machine Learning-International Workshop then Conference. 2003. Vol. 20(2). P. 56-62.

4. Tseng C., Patel N., Paranjape H., Lin T. Y., Teoh S. Classifying twitter data with naive bayes classifier. IEEE International Conference on Granular Computing. 2012. P. 89-101.

5. Vitynskyi P., Tkachenko R., Izonin I., Kutucu H. Hybridization of the SGTM Neural-Like Structure Through Inputs Polynomial Extension. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). 2018. P. 386-391. doi: 10.1109/DSMP.2018.8478456.

6. Boyko N., Mochurad L., Andrusiak I., Drevnytskyi Yu. Organizational and Legal Aspects of Managing the Process of Recognition of Objects in the Image. Proceedings of the International Workshop on Cyber Hygiene (CybHyg-2019) co-located with 1st International Conference on Cyber Hygiene and Conflict Management in Global Information Networks (CyberConf 2019). Kyiv, Ukraine, November 30. 2019. P. 571-592.

7. Agrawal R., Gehrke J., Gunopulos D., Raghavan P. Automatic sub-space clustering of high dimensional data. Data mining knowledge discovery. 2005. Vol. 11(1). P. 5–33.

8. Estivill-Castro V., Lee I. Amoeba: Hierarchical clustering based on spatial proximity using Delaunay diagram. 9th Intern. Symp. on spatial data handling, Beijing, China. 2000. P. 26–41.

9. Boyko N., Shakhovska N. Prospects for Using Cloud Data Warehouses in Information Systems. IEEE 13th International scientific and technical conference on computer sciences and information technologies (CSIT). 2018. Vol. 2. DOI: 10.1109/STC-CSIT.2018.8526745

10. Guo D., Peuquet D.J., Gahegan M. ICEAGE: Interactive clustering and exploration of large and high-dimensional geodata. *Geoinformatica.* 2003. Vol. 3. N. 7. P. 229–253.

11. Harel D., Koren Y. Clustering spatial data using random walks. Proc. of the 7th ACM SIGKDD Intern. conf. on knowledge discovery and data mining, San Francisco, California. 2000. P. 281–286.

12. Boyko N., Pylypiv O., Peleshchak Yu., Kryvenchuk Yu., Campos J. Automated Document Analysis for Quick Personal Health Record Creation. The 2 nd International Workshop on Informatics & Data-Driven Medicine (IDDM 2019). Lviv, Ukraine, November 11-13. 2019. Vol. 1. P. 208-221.

13. Zhang C., Murayama Y. Testing local spatial autocorrelation using. Intern. J. of Geogr. Inform. Science. 2000. Vol. 14. P. 681–692.

14. Melnykova N., Melnykov V., Vasilevskis E. The personalized approach to the processing and analysis of patients' medical data. Proceedings of the 1st International workshop on informatics & Data-driven medicine (IDDM 2018), Lviv, Ukraine, November 28–30. 2018. Vol. 2255. P. 103-112.

15. Yakovyna V., Peleshchyshyn A., Albota S. Discussions of wikipedia talk pages: Manipulations detected by lingual-psychological analysis, CEUR Workshop Proceedings. 2019. Vol. 2392. P. 309-320.