УДК 519.7
DOI https://doi.org/10.24144/2616-7700.2023.42(1).201-207

Д. П. Сабов[1], М. М. Шаркаді[2]

[1] ДВНЗ «Ужгородський національний університет»,
магістр
szabodora20@outlook.hu
ORCID: https://orcid.org/0009-0004-0176-8211

[2] ДВНЗ «Ужгородський національний університет»,
доцент кафедри кібернетики і прикладної математики,
кандидат економічних наук, доцент
marianna.sharkadi@uzhnu.edu.ua
ORCID: https://orcid.org/0000-0002-1850-996X

ПІДХОДИ ЩОДО КЛАСТЕРИЗАЦІЇ КРИПТОВАЛЮТ

Криптовалюти еволюціонували з цифрової новинки до технологій на трильйон доларів, що можуть за кілька років викликати значний вплив на глобальну фінансову систему. Біткоїн та сотні інших криптовалют стають все більш популярними як інвестиційний інструмент, а також використовуються для оплати товарів та послуг, від програмного забезпечення до нерухомості [1].

В межах даної наукової роботи проведено кластеризацію криптовалют з використанням різних методів. Для проведення дослідження використано реальні дані із сервісу CryptoCompare. На першому етапі набір даних нормалізовано та стандартизовано. Далі проведено зменшення розмірності даних. На наступних етапах визначено оптимальну кількість кластерів та проведено поділ криптовалют на відповідні кластери. Для досягнення поставленої мети використано наступні методи: EDA, PCA, t-SNE, k-means, метод ліктя та силуетний метод.

Ключові слова: кластеризація, ефективність, ризик, аналіз даних, кореляція.

Список умовних позначень:
P2P — Peer-to-peer;
ЄЦБ — Європейський центральний банк;
PoW — proof-of-work;
PoS — proof-of-stak;
NFT — Non fungible token (незамінні токени);
PCA — Principal component analysis (метод головних компонентів);
t-SNE — t-distributed stochastic neighbor embedding (T-розподілене вкладення стохастичної близькості);
EDA — дослідницький аналіз даних;
K-means — метод k-середніх.

1. Вступ. Завдяки прогресу в криптографії з'явилась можливість безпечних комунікацій та електронних платежів через Інтернет. Використання кредитних карток є формою електронної готівки, яка покладається на довірену третю сторону для запобігання перевитратам або подвійним витратам. Протокол біткоїн вніс значний внесок у створення чистої однорангової (P2P) децентралізованої валюти, усунувши потребу в довірених третіх сторонах [3].

Наші національні валюти покладаються на центральні банки, які мають повноваження регулювати грошову масу. Ці центральні органи влади не завжди

є державними установами, як Федеральна резервна система. В останнє десятиліття Європейський центральний банк (ЄЦБ) взяв під контроль національні центральні банки Європейського Союзу. ЄЦБ заявляє про свою незалежність від національних урядів, щоб визначати монетарну політику. На жаль, роль ЄЦБ не є політично нейтральною, як це видно в нещодавній грецькій кризі. Обраний грецький уряд зазнав тиску з боку ЄЦБ, коли позбавив грецькі банки ліквідності. Той, хто володіє ключем від друкарського верстата, має величезну економічну та політичну владу [2].

Прихильники криптовалюти бачать її як демократичний інструмент, який усуває контроль центральних банків та фінансових установ над створенням та регулюванням грошей. У свою чергу, критики стверджують, що нова технологія є нерегульованою і може дати можливість злочинним та терористичним організаціям та неприйнятним режимам здійснювати фінансові операції. Вони також підкреслюють, що енергоємний майнінг криптовалют негативно впливає на навколишнє середовище [6].

Зараз фінансові регулятори реагують на швидкий розвиток криптовалют. Норми щодо їх використання значно відрізняються по всьому світу: деякі уряди вітають криптовалюти, а інші обмежують або забороняють їх використання. У своєму бажанні конкурувати з криптовалютними технологіями, центральні банки з усього світу, включаючи Федеральну резервну систему США, розглядають можливість створення власних цифрових валют.

Мотивація творців біткойн полягає у створенні форми «електронного золота», чия цілісність і не фальсифікованість покладаються на математичні властивості, а не на фізичні властивості золота або віру в центральні банки для фіатних грошей [5]. Як це взагалі можливо, коли цифровий токен можна копіювати точно, нескінченно, безкоштовно? Це було б як мати можливість легко виробляти золото, ставлячи під загрозу дефіцитність і властивості не фальсифікації, які роблять його цінним [4]. Але з іншого боку, його електронна природа робить його ідеальним для зберігання та транспортування. Основна перешкода полягає в тому, щоб запобігти можливості «подвійних витрат», тобто одночасного використання одного і того ж токена для різних платежів. Спочатку «децентралізація» та «електронність» здаються несумісними цілями. «Проблема подвійних витрат» є основною складністю створення децентралізованих електронних грошей.

2. Основний результат. У роботі пропонується процес підготовки набору даних криптовалют на основі кластеризації. Кластеризація є одним із найпоширеніших методів дослідницького аналізу даних, який використовується для отримання інформації щодо структури даних.

Для проведення дослідження було використано набір даних криптовалют на основі фінансових даних, який можна використовувати для машинного навчання та експериментів. Ці дані були зібрані з CryptoCompare, сервісу, що відстежує вартість та обсяги торгівлі різними криптовалютами на фінансових ринках. Розглядаються дані, які складаються з шести атрибутів (стовпців):

1) CoinName — назва монети (string).
2) Algorithm — тип алгоритму (string).
3) IsTrading — чи криптовалюта зараз торгується (boolean).
4) ProofType — тип доказу (string).

5) TotalCoinsMined — загальна кількість видобутих монет (int).
6) TotalCoinSupply — загальний запас монет (int).

У рамках даного дослідження були проведені операції з нормалізації та стандартизації набору даних. Для зменшення розміру набору даних були використані методи головних компонент та t-SNE. Окремим етапом дослідження є визначення оптимальної кількості кластерів у наборі даних криптовалют за допомогою підходів кластеризації.

Для досягнення мети використані наступні методи:

– EDA (англ. Exploratory Data Analysis) – це метод аналізу даних, який дозволяє зрозуміти їх структуру та відповідність загальним шаблонам. Цей метод може допомогти у прийнятті рішень про те, які типи моделей можуть бути прийнятними для подальшого моделювання даних.
– PCA (англ. Principal Component Analysis) є методом зменшення розмірності набору даних з метою спрощення та полегшення їхнього розуміння.
– t-SNE (англ. t-distributed Stochastic Neighbor Embedding) – використовується для візуалізації високовимірних наборів даних у вигляді дво- або тривимірних діаграм розсіювання, які легше інтерпретувати, ніж традиційні графіки. Це також корисно для візуалізації кластерів у більших наборах даних, оскільки дозволяє ідентифікувати викиди, не втрачаючи з поля зору загальні тенденції в наборі даних.
– K-means – це алгоритм машинного навчання для кластеризації даних, який створює k кластерів, групуючи схожі елементи разом. Алгоритм працює на основі розрахунку відстані між кожним елементом та центроїдом кластера, який визначається як середнє значення всіх елементів у кластері.
– Метод ліктя – це метод оцінки кількості кластерів у наборі даних, який працює шляхом аналізу відстаней між кожною точкою в кластері та його центроїдом – середньою точкою координат усіх його елементів. Після обчислення цих відстаней для кожного елемента в кожному кластері, метод ліктя визначає, де більшість точок припадає на лінію між їхніми центроїдами та середнім значенням їхнього кластера (або середнім, якщо є лише один), що дозволяє визначити оптимальну кількість кластерів для моделювання.
– Метод силуету – це метод оцінки якості кластеризації, який враховує схожість кожної точки з точками у своєму власному кластері порівняно з іншими кластерами. Використовуючи метрики подібності, такі як взаємна інформація або коефіцієнти кореляції, метод силуету визначає ступінь схожості кожної точки з її кластером та іншими кластерами. Це дозволяє оцінити якість кластеризації та вибрати найоптимальнішу кількість кластерів.

Після проведення процесу очищення даних, наступним кроком є зменшення їх розмірності. Для досягнення цієї мети використані два методи: метод аналізу головних компонент (PCA) та t-SNE.

Використання методу аналізу головних компонент дозволило виявити, що за умови використання 74 головних компонент, ми можемо зберегти більше 90% інформації, яка була присутня в оригінальному наборі даних. Це вказує на високу ефективність методу PCA при зменшенні розмірності даних.

Крім того, було використано метод t-SNE для подальшого зменшення розмірності даних до двох. Цей метод є потужним інструментом, що дозволяє візу-

алізувати високорозмірні дані у вигляді двовимірної картинки, зберігаючи при цьому інформацію про структуру та взаємозв'язки даних (рис. 1).

Аналіз матриці кореляції показав, що метод PCA продовжує використовувати значну кількість ознак (74), що змушує показувати значення кореляції для всіх змінних та їхніх скоригованих оцінок. Недоліком такого підходу є те, що надмірна кількість ознак може спричинити шум та ускладнити обробку даних. Для поліпшення роботи алгоритмів з даними, можливо застосувати додаткові методи зменшення розмірності, які дозволяють зменшити кількість ознак та зосередитися на найбільш значущих змінних.

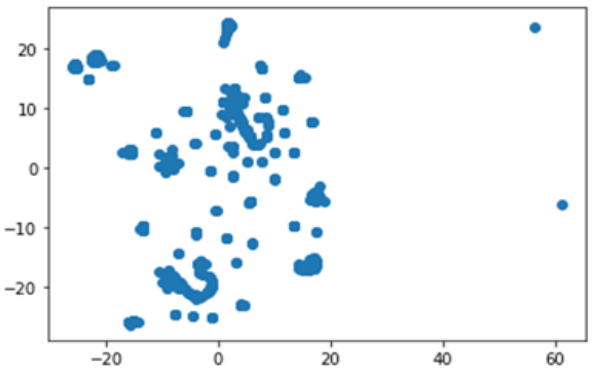

Рис. 1. Побудова результатів методу t-SNE.

Після успішного застосування методів зменшення розмірності, доцільно перейти до визначення оптимальної кількості кластерів за допомогою методу ліктя. На графіку (рис. 2) можна спостерігати точку (чітке ліктя), яка відокремлює область, де інерція змінюється помітно від іншої. У цьому випадку можливий поділ криптовалют на 7 або 8 кластерів. Однак, для точнішого визначення кількості кластерів, необхідно використовувати інші методи, такі як силуетна оцінка, який дозволяє знайти оптимальну кількість кластерів на основі даних, що допоможе у побудові більш точної класифікації криптовалют.

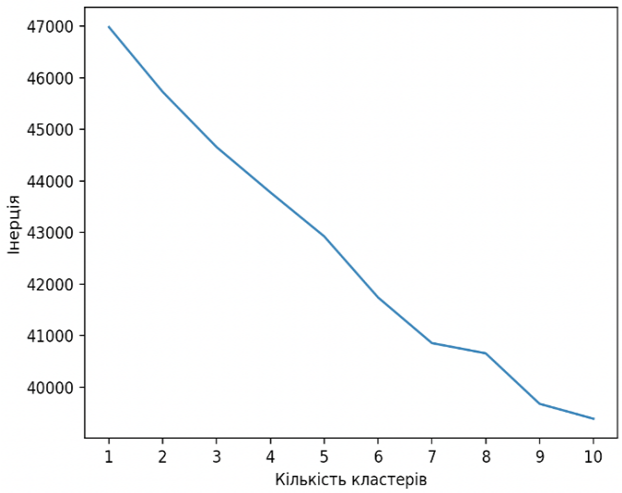

Рис. 2. Створення кривої ліктя.

У цьому дослідженні застосовано метод аналізу силуетів для визначення оптимальної кількості кластерів у кластеризації за допомогою методу k-середніх (рис. 3). З використанням оцінки Silhouette були проаналізовані кластери зі значеннями $K = 2$, $K = 3$, $K = 4$ та $K = 6$. Після ретельного вивчення результатів, було прийнято рішення використовувати $K = 6$ в подальшому аналізі даних.

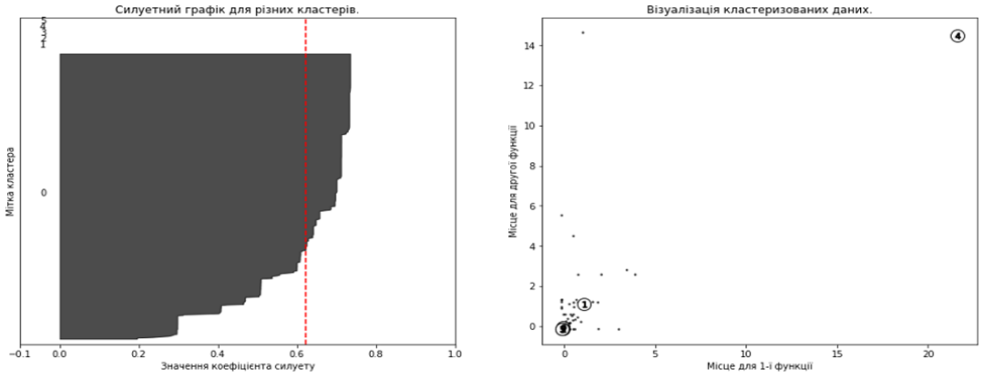

Рис. 3. Аналіз силуетів для кластеризації k-середніх на вибіркових даних з n_clusters $= 6$.

Після прогнозування кластерів за допомогою методу k-середніх при $k = 6$, отримано чітко виділені шість груп (рис. 4). Варто зазначити, що головні компоненти, які створюються за методом головних компонент (PCA), розташовуються в порядку кількості варіацій, які вони охоплюють. Зокрема, перша головна компонента (PC1) фіксує найбільшу кількість варіацій, друга головна компонента (PC2) — другу за кількістю, і так далі. Кожна головна компонента надає певну інформацію про дані, і в методі головних компонент кількість компонент визначається кількістю характеристик у вихідних даних.

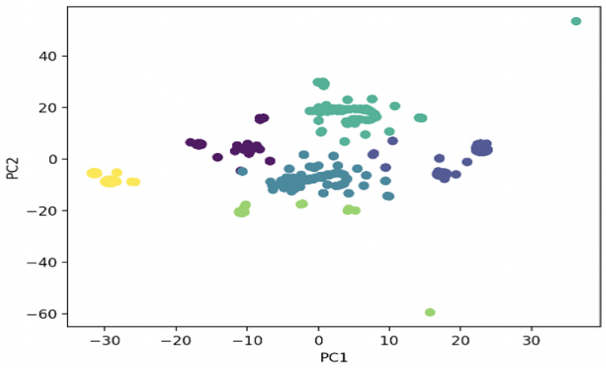

Рис. 4. Прогнозування кластерів при $k = 6$.

3. Висновки та перспективи подальших досліджень. Результати даного дослідження вказують на можливість поділу криптовалют на шість кластерів. Для підтвердження зв'язку між результатами кластеризації та іншими описовими властивостями криптовалют (загальна пропозиція монет, загальна кількість видобутих монет) було проведено різні аналізи в рамках профілювання ринку криптовалют. Цей аналіз дозволив визначити, чи пов'язані пев-

ні характеристики з конкретними фінансовими показниками, встановленими за допомогою алгоритмів кластеризації. Крім того, було виявлено сильний зв'язок між технологічними характеристиками та поведінкою ринку, що відкрило нові можливості для майбутніх досліджень з метою уточнення їхніх висновків.

Це дослідження підтвердило, що дані мають фундаментальну структуру, яка була підтверджена після детального аналізу даних протягом тривалого періоду часу. Кожен алгоритм кластеризації зробив свій внесок у відкриття різних особливостей ринку криптовалют. Комбінування результатів кластеризації дозволило ефективно виявляти основні закономірності на ринку біткойнів. Сегменти кластерів, їх прототипи та опис надають зрозумілу зведену інформацію про загальні фінансові умови на ринку. Оцінюючи перетин кластерів, можна створити більш детальні профілі. Крім того, дослідники або інвестори можуть знайти окремі криптовалюти та встановити, до яких кластерів вони належать.

Дослідження, що здійснено з використанням методів кластеризації для класифікації криптовалют, відкрило нові можливості для аналізу ринку цифрових валют. Проте, наступні етапи дослідження повинні відповісти на багато важливих питань.

Однією з перспектив може стати дослідження взаємодії між криптовалютами та традиційними фінансовими інструментами. Крім того, можна провести дослідження зв'язку між кластеризацією криптовалют та їх технологічними характеристиками, такими як швидкість обробки транзакцій та розмір блоку. Під час таких досліджень можна виявити спільні зв'язки та визначити нові тенденції на ринку криптовалют, що сприятиме подальшому розвитку цього сектору.

Список використаної літератури

1. Nakamoto S. Bitcoin: A peer-to-peer electronic cash system. 2008. 9 p.
2. Pérez-Marco R. Bitcoin and decentralized trust protocols. *European Mathematical Society.* 2016. No. 100. P. 31–38. DOI: https://doi.org/10.48550/arXiv.1601.05254
3. Grunspan C., Pérez-Marco R. Satoshi risk tables. *CoRR.* 2017. DOI: https://doi.org/10.48550/arXiv.1702.04421
4. Giogladis E., Zeilberger D. A combinatorial-probabilistic analysis of bitcoin attacks. *Journal of Difference Equations and Applications.* 2019. Vol. 25, No. 1. P. 56–63. DOI: https://doi.org/10.1080/10236198.2018.1555247
5. Eyal I., Sirer E. G. Majority is not enough: Bitcoin mining is vulnerable. *Communications of the ACM.* 2014. Vol. 61, No. 7. P. 95–102. DOI: https://doi.org/10.48550/arXiv.1311.0243
6. Grunspan C, Pérez-Marco R. The mathematics of Bitcoin. *Newsletter of the European Mathematical Society.* 2020. Vol. 115. P. 31–37. DOI: https://doi.org/10.48550/arXiv.2003.00001

Sabov D. P., Sharkadi M. M. Approaches to clusterization of cryptocurrencies.

Cryptocurrencies have evolved from being a novel digital concept to a trillion-dollar technology that has the potential to significantly impact the global financial system in the coming years. Bitcoin and hundreds of other cryptocurrencies are gaining popularity as investment tools and are increasingly used to pay for goods and services ranging from software to real estate.

Within the context of this scientific work, cryptocurrencies were clustered using various methods. Real data from the CryptoCompare service was used for the study. At the first stage, the data set is normalized and standardized. Data dimensionality reduction was then performed. At the next stages, the optimal number of clusters was determined and cryptocurrencies were divided into the corresponding clusters. To achieve the goal, the following methods were used: EDA, PCA, t-SNE, k-means, the elbow method, and the

silhouette method.

Keywords: clustering, efficiency, risk, data analysis, correlation.

References

1. Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system.*
2. Pérez-Marco, R. (2016). Bitcoin and decentralized trust protocols. *European Mathematical Society,* 100, 31–38. https://doi.org/10.48550/arXiv.1601.05254
3. Grunspan, C., & Pérez-Marco, R. (2017). Satoshi risk tables. *CoRR.* https://doi.org/10.48550/arXiv.1702.04421
4. Giogladis, E., & Zeilberger, D. (2019). A combinatorial-probabilistic analysis of bitcoin attacks. *Journal of Difference Equations and Applications,* 25(1), 56–63. https://doi.org/10.1080/10236198.2018.1555247
5. Eyal, I., & Sirer, E. G. (2014). Majority is not enough: Bitcoin mining is vulnerable. *Communications of the ACM,* 61(7), 95–102. https://doi.org/10.48550/arXiv.1311.0243
6. Grunspan, C, & Pérez-Marco, R. (2020). The mathematics of Bitcoin. *Newsletter of the European Mathematical Society,* 115, 31–37. https://doi.org/10.48550/arXiv.2003.00001

Одержано 31.04.2023