Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
https://nv.nltu.edu.ua
https://doi.org/10.36930/40320513
Article received 09.10.2022 p.
Article accepted 27.10.2022 p.
UDC 004.[6+8]:342.951:351
ISSN 1994-7836 (print)
ISSN 2519-2477 (online)
@ Correspondence author
Kh. V. Lipianina-Honcharenko
xrustya.com@gmail.com

Х. В. Ліп'яніна-Гончаренко, М. П. Комар, А. О. Саченко, Т. В. Лендюк
Західноукраїнський національний університет, м. Тернопіль, Україна

МЕТОД ВИЯВЛЕННЯ ФІКТИВНИХ ПІДПРИЄМСТВ НА ПІДСТАВІ ГАУСОВОГО НАЇВНОГО КЛАСИФІКАТОРА БАЙЄСА

Розроблено метод виявлення фіктивних підприємств на підставі машинного навчання за допомогою Гаусового наївного класифікатора Байєса, що є корисним для працівників державного сектору із запобігання економічним злочинам. Встановлено, що фіктивне підприємництво, як самостійний злочин, одночасно є своєрідним засобом вчинення цілої низки інших кримінальних правопорушень у сфері економіки. Це можуть бути суб'єкти господарювання, які мають ознаки фіктивності, а саме використання неправдивої інформації щодо засновників, адміністрації, місцезнаходження. Тому виявлення таких підприємств є актуальним питанням для будь-якої держави. Розслідування економічного злочину потребує багато часу для працівників правоохоронних органів і додаткових коштів. З огляду на це, розроблення інструменту розпізнавання фіктивного підприємства на підставі класичного методу машинного навчання є одним із перспективних напрямів зі швидкого виявлення економічних злочинів. Під час дослідження робіт у сфері діяльності фіктивних підприємств виявлено, що вони не описують саме виявлення фіктивних підприємств за допомогою інформаційних технологій. Тому потрібно розробити метод виявлення фіктивного підприємства на підставі машинного навчання за допомогою Гаусового наївного класифікатора Байєса, що надалі дасть змогу розробити програмне середовище для працівників державного сектору із запобігання економічних злочинів. У роботі визначено основні типи фіктивних підприємств, зокрема за призначенням та способом створення. На підставі цього запропоновано алгоритм виявлення фіктивного підприємства на підставі класичного методу машинного навчання, такого як Гаусовий наївний класифікатор Байєса, що уможливлює відстежування фіктивного підприємства. Для побудови методу використано дані 1100 компаній, що здійснювали економічну діяльність в Україні. Виконано розподіл ймовірності, за допомогою оцінки щільності ядра KDE (англ. *Kernel Density Estimation*). Побудовано діаграму кореляційної матриці, встановлено дуже малі коефіцієнти кореляції між більшістю ознак. Виведено гістограми відмінностей середніх значень і дисперсії вибірки для двох класів. Для машинного навчання моделі поєднано квантильний перетворювач і Гаусовий наївний класифікатор Байєса.

Ключові слова: суб'єкти господарювання; економічні злочини; класифікація; машинне навчання.

Вступ/ Introduction

Від несприятливих умов ведення господарської діяльності, непрозорості податкової системи, неефективної регуляторної політики економіка держави зазнає непоправних втрат. Збільшується її тіньова сфера (близько 50 %) та кількість економічних злочинів, пов'язаних з використанням фіктивних суб'єктів підприємництва, що становить істотну загрозу економічній безпеці держави.

Суб'єкти господарювання з ознаками фіктивності створено з використанням неправдивої інформації щодо засновників, адміністрації, місцезнаходження (за підробленими, викраденими документами, на підставних осіб, на осіб без постійного місця проживання), або шляхом передачі легально зареєстрованого підприємства у володіння чи управління підставним, померлим, безвісти зниклим особам, щоб використати такі суб'єкти господарювання як засоби вчинення або приховування злочинів.

Зазвичай фіктивне підприємництво є допоміжним злочином, або початковим етапом для вчинення інших основних злочинів у сфері економіки, серед яких: незаконні операції з фінансовими ресурсами; привласнення бюджетних коштів; шахрайство; фіктивне банкрутство; нецільове використання або неповернення отриманих кредитів; ухилення від сплати податків, зборів та інших обов'язкових платежів; здійснення незаконних валютних операцій; приховування за кордоном валютної виручки; різні види розкрадань; легалізація доходів, отриманих злочинним шляхом, а також торгівля людьми,

Інформація про авторів:
Ліп'яніна-Гончаренко Христина Володимирівна, канд. техн. наук, доцент, кафедра інформаційно-обчислювальних систем і управління. **Email:** xrustya.com@gmail.com; https://orcid.org/0000-0002-2441-6292
Комар Мирослав Петрович, д-р техн. наук, доцент, кафедра інформаційно-обчислювальних систем і управління. **Email:** mko@wunu.edu.ua; https://orcid.org/0000-0001-6541-0359
Саченко Анатолій Олексійович, д-р техн. наук, професор, кафедра інформаційно-обчислювальних систем і управління. **Email:** as@wunu.edu.ua; https://orcid.org/0000-0002-0907-3682
Лендюк Тарас Васильович, канд. техн. наук, доцент, кафедра інформаційно-обчислювальних систем і управління. **Email:** tl@wunu.edu.ua; https://orcid.org/0000-0001-9484-8333

Цитування за ДСТУ: Ліп'яніна-Гончаренко Х. В., Комар М. П., Саченко А. О., Лендюк Т. В. Метод виявлення фіктивних підприємств на підставі Гаусового наївного класифікатора Байєса. Науковий вісник НЛТУ України. 2022, т. 32, № 5. С. 92–96.
Citation APA: Lipianina-Honcharenko, Kh. V., Komar, M. P., Sachenko, A. O., & Lendiuk T. V. (2022). Identification method of fictitious enterprises based on Gaussian naive Bayes. *Scientific Bulletin of UNFU*, 32(5), 92–96. https://doi.org/10.36930/40320513

зброєю, наркотиками. Різновидом кримінального бізнесу, що здійснюється шляхом використання фіктивних підприємств, є незаконне переведення безготівкових грошових коштів у готівку та незаконна їх конвертація в іноземну валюту.

Розслідування економічних злочинів потребує багато часу для працівників правоохоронних органів і додаткових коштів. З огляду на це, розроблення інструменту розпізнавання фіктивного підприємства на підставі класичного методу машинного навчання є одним із перспективних напрямів швидкого виявлення економічних злочинів.

Об'єкт дослідження – процеси економічної діяльності.

Предмет дослідження – методи та засоби виявлення фіктивних підприємств з використанням машинного навчання.

Мета роботи – розробити метод виявлення фіктивного підприємства на підставі машинного навчання за допомогою Гаусового наївного класифікатора Байєса, що дасть змогу працівникам державного сектору проводити їх ідентифікацію, аналіз та запроваджувати заходи із запобігання економічним злочинам.

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

- проаналізувати дані компаній, що провадять економічну діяльність в Україні на підставі Гаусового наївного класифікатора Байєса;
- розробити метод виявлення фіктивного підприємства на підставі Гаусового наївного класифікатора Байєса, який дасть змогу працівникам державного сектору здійснювати їх ідентифікацію, аналіз та запроваджувати заходи із запобігання економічним злочинам;
- провести навчання моделі з використанням квантильного перетворювача і Гаусового наївного класифікатора Байєса.

Аналіз останніх досліджень та публікацій. Машинне навчання використовують у багатьох сферах людської діяльності, проте сьогодні воно є надзвичайно актуальним саме у сфері економіки [7]. Роботи [1, 2, 6] спрямовані на огляд сучасних наукових досліджень з питань машинного навчання для виявлення підозрілих операцій з відмивання грошей. У дослідженні [10] розглянуто використання статистичних методів для виявлення фактів відмивання грошей. Статті [3, 8, 9, 11, 12] висвітлюють удосконалені підходи до Гаусового наївного класифікатора Байєса. У роботі [5] розроблено підхід до класифікації на підставі машинного навчання, що визначає законність операції, використовуючи дані з відмивання грошей. У роботі [4] розроблено, описано і протестовано моделі машинного навчання для встановлення пріоритетів щодо того, які фінансові операції варто досліджувати на предмет можливого відмивання грошей.

Матеріали та методи дослідження. Суб'єкти господарювання з ознаками фіктивності можна класифікувати за певними критеріями (рис. 1).

Рис. 1. Класифікація фіктивних підприємств / Classification of fictitious enterprises

Для того, щоб легко визначити чи веде підприємство злочинний бізнес чи ні, розроблено метод виявлення фіктивних підприємств на підставі наївного класифікатора Байєса, який подано алгоритмом (рис. 2) та такими кроками:

Крок 1. Запит на виявлення фіктивного підприємства (блок 1). На цьому етапі проводиться запит на визначення, чи підприємство фіктивне, чи ні.

Крок 2. Введення даних (блок 2). Проводиться введення даних, що стосуються компанії: код підприєм-

ства, назва компанії, її фізична та юридична адреси, КВЕД, реквізити власника компанії та фото компанії з геолокаційними даними.

Крок 3. Збір даних (блок 3). Усі дані збираються автоматично з мережі інтернет.

Крок 4. Порівняння параметрів (блок 4). Проводиться порівняння параметрів на відповідність: тих, які були завантаженні як вхідні, і тих, які були опрацьовані системою та зібрані з мережі інтернет.

Крок 5. Зберігання даних (блок 5) та перетворення їх в бінарні логічні значення (блок 6). За наявності значення параметра ставиться 1, за відсутності – 0.

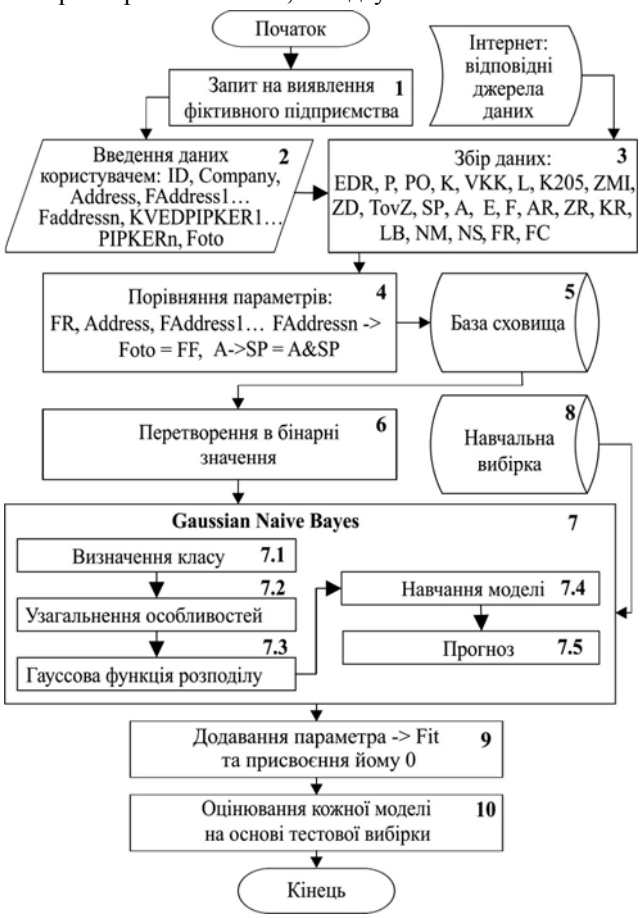

Рис. 2. Алгоритм виявлення фіктивного підприємства на підставі Гаусового наївного класифікатора Байєса / Gaussian Naive Bayes classification algorithm for detecting a fictitious enterprise

Крок 6. Проведення класифікації (блок 7), використовуючи готову навчальну вибірку (блок 8). Розглянемо детальніше принцип класифікації методом наївного Байєса (блок 7). Наївний Байєс – це простий класифікатор, а саме модель, яка присвоює мітки класів у вигляді векторів значень ознак. Існує сімейство алгоритмів підготовки таких класифікаторів, заснованих на загальному принципі: усі наївні класифікатори Байєса допускають, що значення певної ознаки не залежить від значення будь-якої іншої функції, заданої змінної класу.

Під час роботи з безперервними даними типовим припущенням є те, що безперервні значення, пов'язані з кожним класом, розподіляються відповідно до нормального (або Гаусового) розподілу. Дані навчальної вибірки містять постійний атрибут x. Спочатку потрібно сегментувати дані за класом, а потім обчислити середнє значення та дисперсію x у кожному класі. Нехай μ_k є середнім значенням x, асоційованим із класом C_k, і нехай σ_k^2 буде відкоригованою дисперсією Бесселя значень x, асоційованих із класом C_k. Якщо є деяке значення спостереження v, тоді розподіл ймовірностей має клас C_k, $p(x=v\,|\,C_k)$, який можна обчислити використовуючи v у рівнянні для нормального розподілу, параметризованого як і σ_k^2. Отже,

$$p(x=v\,|\,C_k)=\frac{1}{\sqrt{2\pi\sigma_k^2}}Exp\left(-\frac{(v-\mu_k)^2}{2\sigma_k^2}\right).$$

Перевагами цього методу є: класифікація, зокрема багатокласова, виконується легко і швидко; краще працює з нормальним розподілом, що є досить сильним результатом. Негативною рисою є: значення прогнозованих ймовірностей не завжди є достатньо точними; метод краще працює з повністю незалежними ознаками, що в реальних прикладах трапляються вкрай рідко.

Гаусовий наївний класифікатор Байєса має такі кроки:

Крок 7.1. Визначення класу (блок 7.1) за ймовірнісним значенням для кожної ознаки кожного класу.

Крок 7.2. Узагальнення особливостей (блок 7.2), обчислюється на підставі середнього та стандартного відхилення для кожної ознаки кожного класу.

Крок 7.3. Гаусова функція розподілу (блок 7.3). Функція розподілу Гауса (GDF) обчислюється для кожної ознаки кожного класу.

Крок 7.4. Навчання моделі (блок 7.4) за допомогою Гаусового наївного класифікатора Байєса – містить процедуру обчислення середнього та стандартного відхилення для кожної ознаки кожного класу.

Крок 7.5. Передбачення (блок 7.5) класу. Для передбачення класу потрібно розрахувати попередню ймовірність для кожного класу. Клас, який має найбільшу попередню ймовірність, є передбачуваним класом.

Крок 8. Додавання (блок 9) параметра Fit та присвоюємо йому значення 0.

Крок 9. Далі (блок 10), щоб перевірити ефективність отриманої моделі, потрібно розділити кількість правильних передбачень на загальну кількість передбачень, отримані результати і будуть точністю моделі.

Результати дослідження та їх обговорення / Research results and their discussion

Для реалізації методу виявлення фіктивних підприємств обрано мову Python, адже вона найкраще працює з аналізом даних на підставі машинного навчання. Для аналізу використано такі бібліотеки: pandas, numpy, sklearn.naive_bayes, sklearn.metrics.

Для реалізації методу використано дані 1100 компаній, що проводили економічну діяльність в Україні. Дані подано в логічних бінарних значеннях, 355 з яких є фіктивними підприємствами.

Розглянемо розподіл ймовірностей, побудувавши KDE (рис. 3). Метод KDE заснований на схожості, тому він стає повільнішим із збільшенням кількості даних.

Перевіримо залежності між показниками, для цього прорахуємо коефіцієнти кореляції (рис. 4). Діаграма кореляційної матриці показує дуже малі коефіцієнти кореляції між більшістю ознак. Відповідно, визначити залежність між параметрами складно, тому варто використати методи машинного навчання, які дадуть змогу простежити навіть незначні залежності між параметрами.

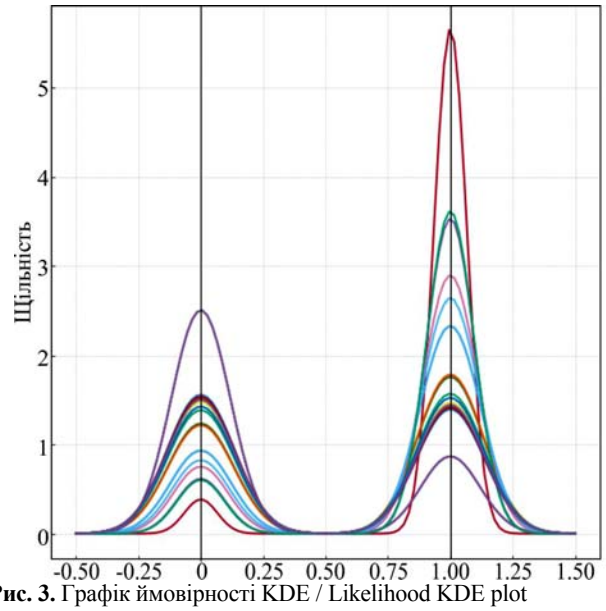

Рис. 3. Графік ймовірності KDE / Likelihood KDE plot

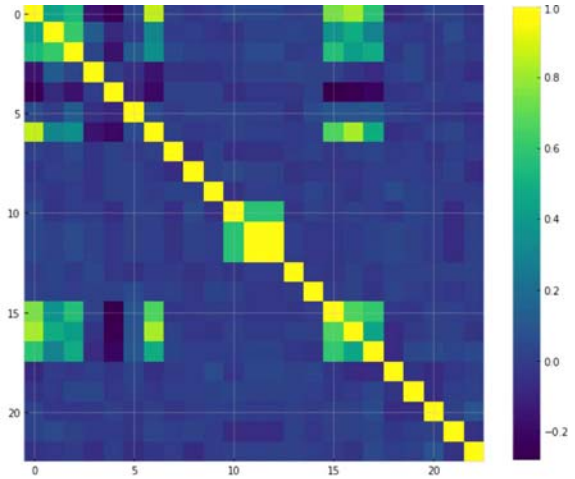

Рис. 4. Графік кореляційної матриці ознак / Correlation matrix plot of the features

Гістограми відмінностей середніх значень та дисперсії вибірки для двох класів подано на рис. 5.

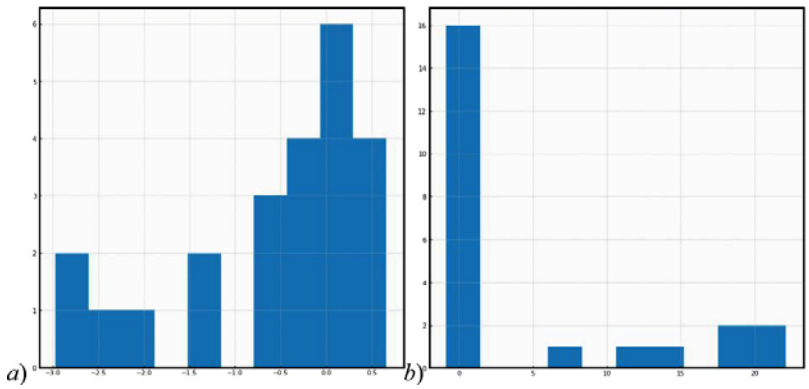

Рис. 5. Гістограми відмінностей середніх значень та дисперсії вибірки для двох класів / Histograms of mean and variance for two classes

Тепер проведемо навчання моделі на підставі поєднання квантильного перетворювача і Гаусового наївного класифікатора Байєса. На рис. 6 подано перші 20 результатів. Як видно з рис. 6, деякі значення не рівні 1, а 0,99. Таких значень у результативній вибірці є невелика кількість, і саме ці значення впливають на точність отриманих результатів, яка загалом дорівнює 0,99.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
fit	1	1	0,9999	0,9999	1	1	0,9999	1	1	1	1	1	1	1	1	0,9999	1	1	1	

Рис. 6. Отримані результати / Obtained results

Отже, для побудови методу використано дані 1100 компаній, що проводили економічну діяльність в Україні на підставі класичного методу машинного навчання, такого як Гаусовий наївний класифікатор Байєса. Виконано розподіл ймовірності, за допомогою KDE. Побудовано діаграму кореляційної матриці, що показує дуже малі коефіцієнти кореляції між більшістю ознак. Для навчання моделі поєднано квантильний перетворювач і Гаусовий наївний класифікатор Байєса.

Обговорення результатів дослідження. Серед проаналізованих робіт [3, 7, 6, 9, 10] не описано виявлення фіктивних підприємств за допомогою інформаційних технологій. Відповідно, метою цієї роботи є розроблення методу виявлення фіктивного підприємства на підставі машинного навчання, а саме класифікатора Гаусового наївного класифікатора Байєса, який дасть змогу надалі розробити програмне середовище для працівників державного сектору, щоб запобігти економічним злочинам та відстежити фіктивні підприємства.

Існує кілька близьких аналогів мети дослідження [4, 5], але в цих роботах проводиться тільки виявлення протиправних дій компаній, а не всі можливі варіанти загалом.

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

Наукова новизна отриманих результатів дослідження – розроблено метод виявлення фіктивного підприємства на підставі машинного навчання за допомогою Гаусового наївного класифікатора Байєса, придатного для ідентифікації, аналізу та запобігання економічним злочинам.

Практична значущість результатів дослідження – розроблене програмне середовище для працівників державного сектору із запобігання економічним злочинам можна використати для відстежування фіктивних підприємств.

Висновки / Conclusions

Запропоновано метод виявлення фіктивного підприємства на підставі класичного методу машинного навчання за допомогою Гаусового наївного класифікатора Байєса, що дасть змогу працівникам державного сектору проводити їх ідентифікацію, аналіз та запроваджувати заходи із запобігання економічним злочинам.

Для побудови запропонованого методу використано дані 1100 компаній, що проводили економічну діяльність в Україні. Виконано розподіл ймовірності за до-

помогою KDE. Побудовано діаграму кореляційної матриці, що показує дуже малі коефіцієнти кореляції між більшістю ознак. Проведено навчання моделі, для цього поєднано квантильний перетворювач і Гаусовий наївний класифікатор Байєса. Оцінка точності моделі становить 0,99.

У наступних наукових дослідженнях плануємо виконати детальніший аналіз виявлення фіктивних підприємств на підставі методу логістичної регресії, зокрема: застосування алгоритму пошуку відповідних значень із джерел інформації, що підходять для виявлення фіктивних підприємств; використання алгоритму розпізнавання образів обладнання підприємств з обробленням геолокаційних даних та перетворення їх у бінарні значення.

References

1. Canhoto, A. I. (2021). Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective. *Journal of Business Research*, 131, 441–452. https://doi.org/10.1016/j.jbusres.2020.10.012
2. Chen, Z., Van Khoa, L. D., Teoh, E. N., Nazir, A., Karuppiah, E. K. & Lam, K. S. (2018). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review. *Knowledge and Information Systems*, 57, 245–285. https://doi.org/10.1007/s10115-017-1144-z
3. Jahromi, A. H., & Taheri, M. (2017). A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. In *Proceedings of the 2017 Artificial Intelligence and Signal Processing Conference* (AISP), 2017, 209-212. https://doi.org/10.1109/AISP.2017.8324083
4. Jullum, M., Løland, A., Huseby, R.B., Ånonsen, G., & Lorentzen, J. (2020). Detecting money laundering transactions with machine learning. *Journal of Money Laundering Control*, 23(1), 173–186. https://doi.org/10.1108/JMLC-07-2019-0055
5. Kumar, A., Das, S., Tyagi, V., Shaw, R. N., & Ghosh, A. (2021). Analysis of Classifier Algorithms to Detect Anti-Money Laundering. In: Bansal, J. C., Paprzycki, M., Bianchini, M., Das, S. (Eds). *Computationally Intelligent Systems and their Applications. Studies in Computational Intelligence*, 950. Springer, Singapore, 143–152. https://doi.org/10.1007/978-981-16-0407-2_11
6. Kute, D. V., Pradhan, B., Shukla, N., & Alamri, A. (2021). Deep learning and explainable artificial intelligence techniques applied for detecting money laundering – A critical review. *IEEE Access*, 9, 82300–82317. https://doi.org/10.1109/ACCESS.2021.3086230
7. Lipyanina, H., Maksymovych, V., Sachenko, A., Lendyuk, T., Fomenko, A., & Kit, I. (2020). Assessing the investment risk of virtual IT company based on machine learning. In: Babichev, S., Peleshko, D., Vynokurova, O. (Eds). Data Stream Mining & Processing. DSMP 2020. *Communications in Computer and Information Science*, 1158, 167–187. Springer, Cham. https://doi.org/10.1007/978-3-030-61656-4_11
8. Ontivero-Ortega, M., Lage-Castellanos, A., Valente, G., Goebel, R., & Valdes-Sosa, M. (2017). Fast Gaussian Naive Bayes for searchlight classification analysis. *Neuroimage*, 163, 471–479. https://doi.org/10.1016/j.neuroimage.2017.09.001
9. Salma, D. F., Murfi, H., & Sarwinda, D. (2019). The performance of one dimensional Naive Bayes classifier for feature selection in predicting prospective car insurance buyers. In: *Tan, Y., Shi, Y. (Eds). Data Mining and Big Data, DMBD 2019. Communications in Computer and Information Science, 1071. Springer, Singapore*, 124–132. https://doi.org/10.1007/978-981-32-9563-6_13
10. Tiwari, M., Gepp, A., & Kumar, K. (2020). A review of money laundering literature: the state of research in key areas. *Pacific Accounting Review*, 32(2), 271–303. https://doi.org/10.1108/PAR-06-2019-0065
11. Valdiviezo-Diaz, P., Ortega, F., Cobos, E., & Lara-Cabrera, R. (2019). A collaborative filtering approach based on Naive Bayes Classifier. *IEEE Access*, 7, 108581-108592. https://doi.org/10.1109/ACCESS.2019.2933048
12. Yodnual, O., & Chaisricharoen, R. (2021). Optimized classification for organizational workload. In Proceedings of the 2021 IEEE Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, 313-317. https://doi.org/10.1109/ECTIDAMTNCON51128.2021.9425699

Kh. V. Lipianina-Honcharenko, M. P. Komar, A. O. Sachenko, T. V. Lendiuk

West Ukrainian Nationally University, Ternopil, Ukraine

IDENTIFICATION METHOD OF FICTITIOUS ENTERPRISES BASED ON GAUSSIAN NAIVE BAYES

A fictitious enterprise in Ukraine should be considered as a business entity that is registered in violation of the established order (legal norms) of registration with state bodies, whose constituent documents do not comply with current legislation, or for carrying out activities that are contrary to the law or constituent documents, or with violation of the procedure for keeping tax records and deadlines for submitting tax declarations and financial statements, or violation of the deadlines for submitting information to state bodies about changing a name, organizational form, form of ownership, and location. The investigation of an economic crime is often time-consuming for law enforcement officials, therefore, in this regard, the development of an algorithm for detecting a fictitious enterprise based on the classical method of machine learning, namely the Gaussian Naive Bayes classifier, will enable developing a single software environment that is the most promising way for quick detection of economic crimes. Therefore, the purpose of this paper is to develop a method for detecting fictitious enterprises based on the Gaussian Naive Bayes classifier, which will further allow the development of a software environment for public sector employees to prevent economic crimes and quickly monitor fictitious enterprises. To meet the specified purpose, the following tasks were performed: data analysis of companies performing economic activity in Ukraine based on the Gaussian Naive Bayes classifier was made; a method for detecting a fictitious enterprise based on the Gaussian Naive Bayes classifier was developed for operational tracking of fictitious enterprises; the model was developed using a quantile transformer and a Gaussian Naive Bayes classifier. Data from 1,100 companies conducting economic activity in Ukraine based on the Gaussian naive Bayes classifier were used to implement the method. In further research, it is planned to carry out a more detailed analysis of the detection of fictitious enterprises based on classification methods. In particular, it is planned to develop an algorithm for searching for relevant values from information sources suitable for detecting fictitious enterprises. We are also keen to develop an algorithm for recognizing images of enterprise equipment with geolocation data processing and converting them into binary values, and also to develop an appropriate software environment for detecting fictitious enterprises.

Keywords: business entities; economic crimes; classification; machine learning.