DOI: https://doi.org/10.26565/2304-6201-2025-65-01

УДК (UDC) 004.62

Бакуменко Ніна Станіславівна *к.т.н., доцент, доцент кафедри комп'ютерних систем та робототехніки, Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 4, Харків-22, Україна, 61022; e-mail: n.bakumenko@karazin.ua;* **https://orcid.org/0000-0003-3496-7167**

Румянцев Данило Максимович *студент; Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 4, Харків-22, Україна, 61022; e-mail:* danylo.rumiantsev@gmail.com*;* **https://orcid.org/0009-0001-7502-8636**

Аналіз та прогнозування злочинності за допомогою методів машинного навчання

Актуальність. У зв'язку з розвитком галузі штучного інтелекту та збільшенням потужності комп'ютерів, з'являється інтерес щодо використання методів машинного навчання для вирішення складних для людей задач. Однією із цих задач є прогнозування злочинності, яке має великий потенціал для покращення людського життя. Завдяки алгоритмам машинного навчання, таким як дерева рішень або випадкові ліси, можна визначати тенденції розвитку злочинності, приховані закономірності та виявляти чинники злочинної діяльності.
Мета. Мета даної статті полягає в аналізі ефективності використання методів машинного навчання, таких як лінійна регресія, дерева рішень, алгоритм k-найближчих сусідів та нейронні мережі для аналізу та прогнозування злочинності.
Методи дослідження. Порівняльний аналіз, експеримент.
Результати. Проведено аналіз ефективності різних методів машинного навчання (лінійна регресія, регресія Ласо, гребнева регресія, регресія k-найближчих сусідів, дерева рішень та модель радіально-базисних нейронних мереж) для аналізу та прогнозування злочинності. Серед розглянутих методів машинного навчання найкращі характеристики показали регресія k-найближчих сусідів та модель радіально-базисних нейронних мереж.
Висновки. Проведений аналіз підтверджує необхідність здійснення довгострокового та оперативного аналізу статистичної інформації з подальшим прогнозуванням факторів та чинників, які впливають на показники злочинності, методами машинного навчання. Отримані результати можуть допомогти у вивченні проблеми аналізу впливу на злочинність соціальних, демографічних чинників, що дозволить планувати профілактичні заходи, розподіляти ресурси правоохоронних органів більш ефективно та ін.

Ключові слова: *методи машинного навчання, прогнозування злочинності, лінійна регресія, регресія Ласо, гребнева регресія, дерева рішень, метод k-найближчих сусідів, нейронні мережі.*

Як цитувати: Бакуменко Н. С., Румянцев Д. М. Аналіз та прогнозування злочинності за допомогою методів машинного навчання. *Вісник Харківського національного університету імені В. Н. Каразіна, серія Математичне моделювання. Інформаційні технології. Автоматизовані системи управління*. 2025. вип. 65. С.6-13. https://doi.org/10.26565/2304-6201-2025-65-01

How to quote: Bakumenko N. S., Rumiantsev D. M. "Crime analysis and prediction using machine learning methods". *Bulletin of V. N. Karazin Kharkiv National University, series Mathematical modeling. Information Technology. Automated control systems.* vol. 64. pp. 6-13. https://doi.org/10.26565/2304-6201-2025-65-01 [in Ukrainian]

1 Вступ

Прогнозування злочинності є важливим інструментом для правоохоронних органів, який дозволяє оцінювати ефективність роботи, планувати заходи запобігання злочинності, розробляти стратегії протидії тощо. Наявність великих обсягів даних, які надаються деякими державними органами у відкритий доступ, пригортає увагу дослідників до цієї галузі і спричиняє збільшення кількості публікацій на цю тему останніми роками [1]. Використання методів штучного інтелекту та машинного навчання для обробки статистичних, демографічних даних дозволяє аналізувати великі обсяги даних з метою виявлення закономірностей, виявлення чинників та передбачення злочинної діяльності [2].

У даній статті методи машинного навчання, зокрема регресійні моделі різних типів, розглядаються з метою порівняння ефективності цих методів для аналізу та прогнозування злочинності за статистичними та демографічними показниками.

2 Аналіз літературних джерел та постановка задачі

Методи машинного навчання та інтелектуального аналізу даних активно використовується для обробки великих масивів даних в різних галузях людської практики, зокрема в аналізі та прогнозуванні злочинності [3]. В роботі [4] наведений детальний огляд проблем та факторів, які виникають при прогнозуванні злочинів, і запропонований алгоритм K-найближчих сусідів для прогнозування злочинності у Ванкувері. Geetha Vadav та ін. в роботі [5] серед алгоритмів на основі дерев рішень, SVM, K-найближчих сусідів та випадкових лісів показали, що останній є найбільш ефективним в задачах прогнозування злочинності. В роботі [6] для прогнозування злочинності в Індії було запропоновано поєднання різних методів машинного навчання. Прикладом використання нейронних мереж для прогнозування злочинності може стати стаття [7], у якій одна з моделей нейронних мереж визначала місце злочину в межах сусіднього поштового індексу у 31,2% випадків, коли місце злочину було одразу невідомо. Наприкінці статті автор спонукає до подальшого вивчення використання нейронних мереж для прогнозу злочинності, особливо при використанні певних додаткових ознак для покращення прогнозів. В роботі [8] розглядається використання нейронних мереж в таких задачах, але з акцентом на необхідність можливості прозорої інтерпретації результатів моделювання в сфері аналізу злочинності.

В розглянутих джерелах аналіз та прогнозування злочинності було здійснено на підставі моделей класифікації. В даній роботі досліджується можливість побудови моделі для прогнозування показників злочинності у вигляді кількісних показників, яка дозволяла б оцінювати вплив окремих чинників на результат та надавала б просту інтерпретацію результатів, в класі регресійних моделей. Регресійний аналіз є фундаментальною концепцією в галузі машинного навчання і належить до методів контрольованого навчання, у якому алгоритм навчається як з вхідними значеннями, так і з вихідними мітками. Регресія в машинному навчанні складається з математичних методів, які дозволяють дослідникам даних передбачити безперервний результат (y) на основі значення однієї або кількох змінних предиктора (x). Це допомагає встановити зв'язок між змінними, оцінюючи, як одна змінна впливає на іншу.

3 Використання регресійного аналізу для прогнозування злочинності

Регресійний аналіз – це статистичний метод, який використовується для вивчення зв'язку між залежною змінною та однією або кількома незалежними змінними [9]. Він використовується для різних цілей, зокрема:

- Прогнозування: регресійний аналіз можна використовувати для прогнозування майбутніх тенденцій на основі історичних даних.
- Перевірка гіпотез: регресійний аналіз можна використовувати для перевірки гіпотез про зв'язок між залежною та незалежною змінними.
- Контрольні змінні: регресійний аналіз можна використовувати для контролю інших змінних, які можуть впливати на зв'язок між залежними та незалежними змінними.

3.1 Лінійна регресія

Лінійна регресія надає рівняння залежності змінної-результату від пояснювальних змінних у вигляді лінійної моделі:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p$$

де $x_1, x_2, \ldots, x_p$ – незалежні змінні (вхідні змінні),

y – залежна змінна (вихід),

$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ – коефіцієнти рівняння регресії, параметри моделі.

Лінійні моделі припускають лінійний зв'язок між залежною змінною та незалежними змінними. Це припущення може бути обмеженим, особливо коли зв'язок між змінними не є лінійним. Нелінійні зв'язки часто є складнішими і можуть вимагати більш складних моделей, щоб вловити їхні нюанси.

- Моделі лінійної регресії чутливі до викидів. Викиди можуть впливати на нахил і перетин лінії регресії, що призводить до неточних прогнозів, тому важливо ідентифікувати викиди та відповідним чином обробляти їх.
- Моделі лінійної регресії припускають, що зв'язок між залежною змінною та незалежними змінними є лінійним. У деяких випадках це припущення може не відповідати дійсності, що призводить до неточних прогнозів.
- Моделі лінійної регресії схильні до перенавчання, особливо коли кількість незалежних змінних велика порівняно з розміром вибірки.
- Моделі регресії припускають, що незалежні змінні не сильно корельовані одна з одною.
- Моделі регресії не можуть безпосередньо обробляти категоріальні змінні.
- Моделі регресії припускають, що дисперсія залежної змінної постійна на всіх рівнях незалежних змінних. На практиці це припущення може не відповідати дійсності, що призводить до неточних прогнозів.

Загалом, регресійний аналіз є корисним інструментом для аналізу та розуміння зв'язку між змінними, а також для прогнозування та прийняття обґрунтованих рішень на основі цього зв'язку, але слід пам'ятати про його недоліки при побудові моделей.

3.2 Рідж та Ласо регресія

Регресія ласо, також відома як регулярізація L1, є формою регуляризації моделей лінійної регресії. Регулярізація – це статистичний метод для зменшення помилок, викликаних перенавчанням. Регресія Лассо додаває до функції втрат при навчання моделі штрафну змінну – суму модулей коефіцієнтів, до залишкової суми квадратів (RSS), яка потім множиться на параметр регуляризації λ, який контролює ступінь застосованої регулярізації:

$$S(\hat{\beta}) = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip} - y_i)^2 + \lambda \sum_{j=1}^{p}\left|\hat{\beta}_j\right|$$

Більші значення параметра збільшують штраф, скорочуючи більше коефіцієнтів до нуля; що може зменшити важливість (або взагалі виключити) деякі чинники з моделі. І навпаки, менші значення параметра регуляризації λ зменшують ефект штрафу, зберігаючи більше функцій у моделі.

Гребнева регресія в якості штрафу використовує суму квадратів коефіцієнтів моделі:

$$S(\hat{\beta}) = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip} - y_i)^2 + \lambda \sum_{j=1}^{p}\hat{\beta}_j^2$$

Тут λ також є параметром регуляризації,

3.3 Метод K-найближчих сусідів (КНС)

Регресія K-найближчих сусідів (КНС) – це непараметричний алгоритм машинного навчання, який можна використовувати як для завдань класифікації, так і для регресії [10]. Цей алгоритм робить прогнози, знаходячи k найближчих точок даних до заданого вхідного сигналу та усереднюючи їхні цільові значення (для числової регресії).

Ключовими параметрами методу КНС є:

1 Кількість сусідів (k): кількість найближчих сусідів k, які використовуватимуться для створення прогнозів. Маленьке k може призвести до шумних прогнозів, тоді як велике k може призвести до надмірно згладжених прогнозів.
2 Метрика відстані: КНС покладається на метрику відстані для вимірювання подібності між точками даних. Залежно від природи даних можна використовувати різні показники відстані.

Для кожної нової точки вхідних даних КНС обчислює відстань між цією точкою та всіма іншими точками даних у наборі даних. Потім він вибирає k точок даних з найменшими відстанями. Прогнозом регресії є середнє значення цільової функції k найближчих сусідів. Це може бути просте середнє арифметичне.

Регресію КНС легко реалізувати та зрозуміти, але вона може бути дорогою з точки зору обчислень, особливо для великих наборів даних, оскільки вимагає обчислення відстані між

новою точкою даних і всіма існуючими точками даних. Крім того, вибір правильного значення k і відповідної метрики відстані може вплинути на якість прогнозів.

3.4 Регресія дерева рішень (Decision Tree Regression)

Регресія дерева рішень – це метод машинного навчання, який створює деревоподібну модель для прогнозування безперервних числових значень [11]. Дерева рішень можуть фіксувати нелінійні зв'язки між змінними, не вимагаючи явних перетворень або припущень. На дерева рішень менше впливають викиди порівняно з іншими моделями регресії, а також можуть обробляти відсутні значення шляхом автоматичного розгляду альтернативних гілок на основі доступних даних.

3.5 Радіально-базисні нейронні мережі.

Радіально-базисні нейронні мережі (РБНМ) – це особливий клас прямої нейронної мережі, що складається з трьох рівнів: вхідного, прихованого та вихідного [12]. Перетворення від вхідного шару до прихованого є нелінійним, а від прихованого до вихідного – лінійним. З кожним прихованим нейроном пов'язана радіально-базисна виду

$$\phi_i\left(\|x-u_i\|\right)=\exp\left(-\frac{1}{2\sigma_i^2}\|x-u_i\|^2\right)$$

де u_i – центр функції, σ_i – її ширина.

Вихід нейронної мережі розраховується за формулою

$$y=\sum_{i=1}^{p} w_i\phi_i\left(\|x-u_i\|\right)$$

де w_i – ваги вихідного шару. центр функції.

Архітектура радіально-базисної нейронної мережі зображена на рис. 1.

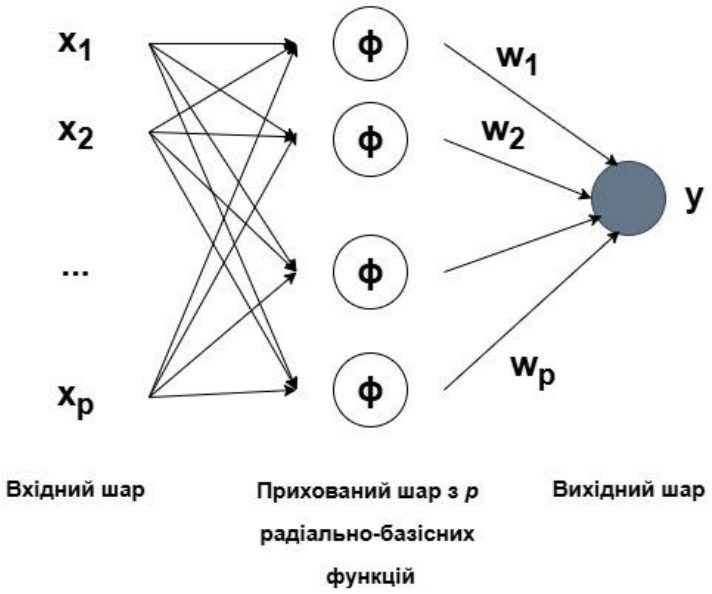

Рисунок 1. Архітектура РБНМ.
Fig. 1. Archtecture of RBNN (Radial Basis NeuralNetwork)

Вхідний рівень складається з одного нейрона для кожної змінної предиктора. Вхідні нейрони передають значення кожному нейрону прихованого шару. Прихований шар містить змінну кількість нейронів (ідеальна кількість визначається процесом навчання). Кожен нейрон містить радіальну базисну функцію з центром у точці. Коли вектор x вхідних значень подається з вхідного рівня, прихований нейрон обчислює евклідову відстань між тестовим випадком і центральною точкою нейрона. Потім він застосовує функцію ядра (радіально-базисну функцію).

Отримане значення надходить до рівня підсумовування: значення, отримане з прихованого шару, множиться на вагу, пов'язану з нейроном. Отримана сума є виходом мережі.

4 Тестовий набір даних

Для тестування роботи моделей було використано набір даних «Communities in the US» [13], який містить у собі соціально-економічні дані з перепису 1990 року, дані правоохоронних органів з опитування щодо управління правоохоронними органами та адміністративної статистики 1990 року та дані злочинності з ФБР 1995 року. Цей набір даних містить 147 атрибутів і 2216 екземплярів і також має у собі пропущені значення. Розглянемо цільові атрибути даного датасету:

Таблиця 1. Цільові атрибути Communities in the US
Table 1. Target attributes of Communities in the US

Назва змінної	Тип	Пропущені змінні	Опис
population	Integer	ні	Громадське населення
agePct12t29	Continuous	ні	Відсоток населення віком 12-29 років
numbUrban	Integer	ні	Кількість людей, які проживають у районах, класифікованих як міські
pctUrban	Continuous	ні	Відсоток населення, яке проживає у районах, класифікованих як міські
medIncome	Integer	ні	Середній дохід домогосподарства
pctWInvInc	Continuous	ні	Відсоток домогосподарств з інвестиційним/рентним доходом у 1989 році
pctWSocSec	Continuous	ні	Відсоток домогосподарств із доходом соціального забезпечення у 1989 році
pctWPubAsst	Continuous	ні	Відсоток домогосподарств із доходом від державної допомоги у 1989 році
medFamInc	Integer	ні	Середній дохід сім'ї
NumUnderPov	Integer	ні	Кількість людей, які перебувають за межею бідності
PctPopUnderPov	Continuous	ні	Відсоток населення, яке перебуває за межею бідності
PctLess9thGrade	Continuous	ні	Відсоток людей віком від 25 років, які мають освіту менше 9 класів
PctNotHSGrad	Continuous	ні	Відсоток людей віком від 25 років, які не є випускниками середньої школи
PctBSorMore	Continuous	ні	Відсоток людей віком від 25 років, які мають ступінь бакалавра або вищу освіту
PctUnemployed	Continuous	ні	Відсоток людей віком від 16 років, які входять до робочої сили та безробітних
PctEmploy	Continuous	ні	Відсоток працевлаштованих осіб віком від 16 років
PctOccupManu	Continuous	ні	Відсоток людей віком від 16 років, які зайняті у виробництві
PctOccupMgmtProf	Continuous	ні	Відсоток людей віком від 16 років, які зайняті на керівних або професійних посадах
Murders	Integer	ні	Кількість вбивств у 1995 р.
murdPerPop	Continuous	ні	Кількість вбивств на 100 тис. населення
Rapes	Integer	так	Кількість зґвалтувань у 1995 р.
rapesPerPop	Continuous	так	Кількість зґвалтувань на 100 тис. населення
Robberies	Integer	так	Кількість пограбувань у 1995 р.
robbbPerPop	Continuous	так	Кількість пограбувань на 100 тис. населення
Assaults	Integer	так	Кількість нападів у 1995 році
assaultPerPop	Continuous	так	Кількість нападів на 100 тис. населення

Burglaries	Integer	так	Кількість крадіжок зі зломом у 1995 році
burglPerPop	Continuous	так	Кількість крадіжок зі зломом на 100 тис. населення
Larcenies	Integer	так	Кількість крадіжок у 1995 році
larcPerPop	Continuous	так	Кількість крадіжок на 100 тис. населення
autoTheft	Integer	ні	Кількість крадіжок автомобілів у 1995 році
autoTheftPerPop	Continuous	ні	Кількість крадіжок автомобілів на 100 тис. населення
Arsons	Integer	так	Кількість підпалів у 1995 р.
arsonsPerPop	Continuous	так	Кількість підпалів на 100 тис. населення
violentPerPop	Continuous	так	Загальна кількість насильницьких злочинів на 100 тис. населення
nonViolPerPop	Continuous	так	Загальна кількість ненасильницьких злочинів на 100 тис. населення

Оскільки набір даних містив пропущені значення, була виконана обробка пропущених значень шляхом заміни їх на медіанні значення. Також була здійснена нормалізація змінних та поділ вибірки на тестову та навчальну в пропорції 70% – тренувальні дані, 30% – тестові.

Для побудови моделі залежності кількості вбивств *Y* вули використані змінні population, agePct12t29, numbUrban, pctUrban, medIncome, pctWInvInc, pctWSocSec, pctWPubAsst, medFamInc, NumUnderPov, PctPopUnderPov, PctLess9thGrade, PctNotHSGrad, PctBSorMore, PctUnemployed, PctEmploy, PctOccupManu, PctOccupMgmtProf, які показали найбільшу кореляцію з цільовою змінною.

5 Оцінка точності прогнозуваннння для різних методів регресії

Модель лінійної регресії показала точність 0,96 для тренувального набору і 0,93 для тестового набору. Майже такі ж самі результати були отримані за допомогою гребеневої регресії (точність 0,959 для тренувального набору і 0,927 для тестового набору) і лассо регресії (0,958 для тренувального набору і 0,923 для тестового набору). Для моделі k-найближчих сусідів було отримано ь 0,705 для тренувального набору і 0,74 для тестового набору. Модель випадкових лісів показала кращі результати - точність 1,0 для тренувального набору і 0,871 для тестового набору. Результати на тренувальному наборі можуть свідчити про перенавчання моделі, хоча на тестових точність більш реалістична. Після проведених експериментів було проведено налаштування параметрів моделей за допомогою крос-валідації. Порівняння точності прогнозування для різних моделей наведено в таблиці 2.

Таблиця 2. Порівняння результатів роботи моделей
Table 2. Comparison of models performances

Модель	Точність на тренувальному наборі	Точність на тестовому наборі
Linear Regression (default)	0,96	0,93
Ridge Regression (default) ($\lambda = 1.0$)	0,959	0,927
Ridge Regression (best) ($\lambda = 0.001$)	0,96	0,93
Lasso Regression (default) ($\lambda = 1.0$, max_iter=1000)	0,953	0,928
Lasso Regression (best) ($\lambda = 0.1$, max_iter=10000)	0,959	0,93
K-nn Regression (default) (n_neighbors=5)	0,705	0,74
K-nn Regression (best) (n_neighbors=2)	0,934	0,832
РБНМ	-0,06	0,03

6 Висновки

У даному дослідженні проведено аналіз ефективності використання різних регресійних методів, зокрема лінійна регресія, дерева рішень, випадкові ліси, алгоритм k-найближчих сусідів та радіально-базисні нейронні мережі для аналізу та прогнозування злочинності. Тестування моделей було виконано на наборі даних «Communities in the US».

Результати дослідження показали, що метод k-найближчих сусідів, після підбору найкращих параметрів шляхом крос-валідації, надає найбільш точний прогноз порівняно з іншими обраними методами. У той же час, модель радіально-базисних нейронних мереж показала низьку похибку при тестуванні, і, теоретично, у задачах коли необхідна велика гнучкість, може перевершити метод k-найближчих сусідів.

Таким чином, вибір методу машинного навчання для аналізу та прогнозування злочинності залежить від конкретних вимог від користувача. Якщо пріоритетом є точність прогнозу, k-найближчих сусідів надає найбільш реалістичні прогнози.

Отже, результати дослідження підтверджують доцільність використання методів машинного навчання для аналізу та прогнозування злочинності та дозволяють зробити обґрунтований вибір алгоритму залежно від специфічних вимог.

REFERENCES

1. Crime Analysis Through Machine Learning / S. Kim ta in. IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). Canada: Vancouver, BC. 2018. PP. 415–420. URL: https://doi.org/10.1109/IEMCON.2018.8614828.

2. Jenga K., Catal C., Kar G. Machine learning in crime prediction. Journal of Ambient Intelligence and Humanized Computing. 2023. Vol. 14. PP. 1–27. URL: https://www.researchgate.net/publication/368164162_Machine_learning_in_crime_prediction (access date: 19.03.2025).

3. Crime data mining: a general framework and some examples / H. Chen та ін. *IEEE Computer*. 2004. Vol 37, No 4, PP. 50–56. URL: https://ieeexplore.ieee.org/document/1297301 (дата звернення: 19.03.2025).

4. Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions / V. Mandalapu та ін. *IEEE Access*. 2023. Vol. 11. PP. 60153–60170. URL: https://ieeexplore.ieee.org/document/10151873 (access date: 19.03.2025)

5. The Role of Machine Learning in Crime Analysis and Prediction / M. Geetha Vadav et al. *2024 International Conference on Expert Clouds and Applications (ICOECA), Bengaluru, India,*. 2024. Vol. 2024 International Conference on Expert Clouds and Applications (ICOECA), Bengaluru, India, 2024. P. 885–890. URL: https://doi.org/10.1109/ICOECA62351.2024.0015 .

6. Advancing Crime Analysis and Prediction: A Comprehensive Exploration of Machine Learning Applications in Criminal Justice / N. Thoiba Singh et al. *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India*. 2024. P. 1339–1343. URL: https://doi.org/10.1109/IDCIoT59759.2024.10467221 .

7. Walczak S. Predicting Crime and Other Uses of Neural Networks in Police Decision Making. *Frontiers in Psychology*. 2021. Vol. 12. P. 1–11. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.587943/full (access date: 19.05.2025).

8. Artificial Intelligence in Crime Prediction: A Survey With a Focus on Explainability / F. Ersöz et al. *IEEE Acces*. 2025. Vol. 13. P. 59646–59674. URL: https://unis.karabuk.edu.tr/yayin-detay/2_DJOqC3W_39/artificial-intelligence-in-crime-prediction-a-survey-with-a-focus-on-explainability (access date: 19.03.2025).

9. Bazhan T. O. Porivnialnyi analiz metodiv mashynnoho navchannia dlia pobudovy prohnoziv.. Suchasnyi zakhyst informatsii.. 2024. Vol. 4, iss. 60. P. 125–130. URL: https://doi.org/10.31673/2409-7292.2024.040013.

10. Cielen D., Arno D. B., Meysman M. A. Introducing Data Science Big data, machine learning, and more, using Python tools. Manning Publications, 2016. 320 p. ISBN 9781633430037. (access date: 09.02.2025).

11. Bishop C. M. Pattern Recognition and Machine Learning. New York : Springer-Verlag, 2016. 778 p. ISBN 978-0-387-31073-2.

12. Haykin S. S. Neural Networks: A Comprehensive Foundation. Hamilton, Ontario, Canada : Prentice Hall, 1999. 1104 p. ISBN 0-13-273350-1.

13. Redmond, M. (2009). Communities and Crime Unnormalized [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5PC8X

Bakumenko Nina *Candidate of Technical Sciences; Associate Professor of Computer Systems and Robotics Department;*
V. N. Karazin Kharkiv National University
Svobody Sq 4, Kharkiv, Ukraine, 61022
e-mail: n.bakumenko@karazin.ua;
https://orcid.org/0000-0003-3496-7167

Rumiantsev Danylo *student;*
V. N. Karazin Kharkiv National University
Svobody Sq 4, Kharkiv, Ukraine, 61022
e-mail: danylo.rumiantsev@gmail.com*;*
https://orcid.org/0009-0001-7502-8636

Crime analysis and prediction using machine learning methods

Relevance. As artificial intelligence continues to develop and computer power increases, there is growing interest in applying machine learning methods to tackle tasks that are challenging for humans. One such task is crime prediction, which has significant potential to enhance the effectiveness of law enforcement. Machine learning algorithms, including decision trees and random forests, can identify crime trends, uncover hidden patterns, and determine factors contributing to criminal activity.
Goal. The purpose of this article is to analyze the effectiveness of using machine learning methods, such as linear regression, decision trees, the k-nearest neighbors algorithm, and neural networks for crime analysis and prediction.
Research methods. Comparative analysis, experiment.
Results. The effectiveness of various machine learning methods (linear regression, Lasso regression, ridge regression, k-nearest neighbor regression, decision trees, and radial-basis neural network model) for crime analysis and prediction was analyzed. Among the considered machine learning methods, the k-nearest neighbor regression and radial-basis neural network model showed the best characteristics.
Conclusions. The analysis confirms the need for long-term and operational analysis of statistical information with subsequent prediction of factors and factors that affect crime rates using machine learning methods. The results obtained can help in studying the problem of analyzing the impact of social and demographic factors on crime, which will allow planning preventive measures, distributing law enforcement resources more effectively, etc.

Keywords: *machine learning methods, crime prediction, linear regression, Lasso regression, ridge regression, decision trees, k-nearest neighbor method, neural networks.*

Надійшла у першій редакції 12.02.2025, в останній - 19.03.2025.
The first version has been received on 12.02.2025, the final version - on 19.03.2025.