DOI: https://doi.org/10.26642/ten-2022-2(90)-78-85
УДК 004.089

Ю.О. Годлевський, Senior Software Engineer
Infopulse
Г.В. Марчук, ст. викладач
І.В. Панаріна, к.т.н., доц.
Державний університет «Житомирська політехніка»

Аналіз, моделювання та прогнозування ціни будинків залежно від їх розмірів

Мета аналізу даних – отримувати інформацію, яку непросто витлумачити, але яка, якщо її зрозуміти, допоможе правильно її використати. У статті описано новий підхід до використання інтелектуальних технологій для певних бізнес-рішень, а саме для дослідження цінової політики вартості будинків залежно від їх розмірів. Об'єктом дослідження є аналіз даних інформації про наявні в продажу будинки, їх розміри і вартість. Предметом дослідження є методи, алгоритми та засоби інтелектуального аналізу даних. У статті розглядається найбільш простий і зрозумілий, водночас часто використовуваний метод математичного програмування для вирішення завдань такого роду – метод градієнтного спуску (gradient descent). У роботі використана найбільш поширена функція втрат середньоквадратичної помилки. Похідна цієї функції показує обчислення градієнта для лінійної регресії. Використано лінійну регресію (Linear regression) – модель залежності змінних (вартості та площі будинку) з лінійною функцією залежності. Дані для аналізу були завантажені з декількох інтернет-ресурсів продажу та оренди житла. В статті представлено візуалізацію методу градієнтного спуску по функції помилки за двома параметрами. Крім 3D-графіків, у роботі представлені контурні графіки. Контурний графік – це метод представлення тривимірного зображення у двомірній площині, що добре представляє покрокову роботу методу градієнтного спуску. В результаті було спроєктовано модель, де можна побачити, як модель підлаштовується під тренувальні дані і готова виконувати своє завдання. Провівши тестування запропонованої моделі, можна виявити приблизну вартість будинку залежно від його розмірів. Функція помилки мінімізована і відповідає вимогам поставленого завдання. В подальшому планується збільшити кількість вхідних даних для аналізу, вказавши місце розташування, кількість кімнат, величину прилеглої території.

Ключові слова: *аналіз даних; модель лінійної регресії; машинне навчання; штучний інтелект.*

Актуальність теми. На сьогодні існує велика кількість інформації, яка потребує певного аналізу, завдяки якому можна робити багато корисних висновків для різних сфер діяльності людини. Для інтелектуального аналізу цих даних і розробки відповідних інтелектуальних й автоматизованих додатків ключем є знання штучного інтелекту. Актуальність теми обумовлена впровадженням інтелектуальних технологій у такі сфери людства, як соціальні мережі, фінансові послуги, медицина, транспорт, уряд, торгівля, пошук корисних копалин тощо. Бізнес-аналітики часто потребують різних зрізів даних, які є необхідними для певних бізнес-рішень, які приводять бізнес до успіху. Однією з ключових частин ринку продажу є інформація та актуальні прогнози стосовно вартості житла на ринку.

Аналіз останніх досліджень та публікацій, на які спираються автори. Аналіз даних увійшов у різні сфери людського існування. У [1] досліджуються алгоритми інтелектуального аналізу даних, які на основі правил і обчислень дозволяють створити модель, що аналізує дані, здійснюючи пошук певних закономірностей і тенденцій. Автори демонструють практичне застосування алгоритмів для підвищення якості медичної допомоги пацієнтам. Для проведення аналізу даних застосовують різні інструменти. Автори статті [2] пропонують новий інструмент аналізу даних під назвою «модальна лінійна регресія для дослідження багатовимірних даних». У роботі проведено дослідження зі змодельованими та реальними даними. Автори демонструють, що запропонована модальна регресія дає коротші прогностичні інтервали, ніж середня лінійна регресія, медіанна лінійна регресія та ММ-оцінки.

У наукових роботах часто застосовують кластерний аналіз для вивчення диференціації соціально-економічного розвитку регіонів. Цим питанням займаються як вітчизняні, так і зарубіжні автори. Опмане І. [3] проводить кластерний аналіз регіонів Латвії. Різним напрямам побудови кластерів серед регіонів України присвячено праці [4–9]. Вибір туристичних кластерів та їх ранжування є складним завданням у сфері аналізу даних, оскільки не існує єдиного зведеного показника інвестиційної привабливості. У [10] наведено опис процесу та результатів впровадження алгоритму kmeans в аналітичній платформі Loginom для задачі кластеризації регіонів України за рівнем інвестиційної привабливості у сфері туризму.

У [11] запропоновано програмну реалізацію з використанням методів штучного інтелекту для пошуку вільного місця на парковці. Розроблена система надає змогу користувачам авто здійснювати пошук

вільних місць, витрачаючи при цьому мінімум часу. У [12] описано основні алгоритми аналізу потоку кадрів відеоданих, шо надходять з камер міста. Основною метою дослідження є мінімізація часу на пошук вільного місця для паркування автомобіля. Месюра В. та Гранік М. [13] застосовують методи штучного інтелекту для аналізу новин та доводять ефективність у вирішенні проблеми виявлення фейкових новин. У [14] визначається вплив нематеріальних активів на ринкову вартість європейських компаній (Німеччина, Франція та Велика Британія) за допомогою інтелектуального аналізу даних. Було виявлено зв'язок між нематеріальними активами та ринковою вартістю компаній в аналізованих країнах Європи.

Метою статті є дослідження даних будинків та виявлення відповідних цін залежно від їхнього розміру. Досягнення поставленої мети передбачає вирішення таких завдань:

– аналіз проблематики, методів та засобів аналізу статистики;

– реалізація алгоритмів інтелектуального аналізу даних;

– аналіз результатів.

Викладення основного матеріалу. Варто зазначити, що аналітика у питаннях оренди та купівлі житла може заощадити колосальні кошти та час як продавцям, так і покупцям. Підприємці здатні збирати величезні обсяги даних і шукати найкращі стратегії їхнього використання для поліпшення та прискорення процесу підбору житла, тим паче з прогресуванням сучасних технологій це робити дедалі простіше та дешевше. Розглянемо як аналітичний підхід може покращити процеси продажу житла.

Для знаходження певної інформації потрібно використати алгоритми пошуку, які працюють у певній структурі даних або в проблемній області, з дискретними або безперервними значеннями. Завдяки різноманітним алгоритмам пошуку, користувач має змогу знайти варіанти житла за різними фільтрами, у тому числі і за локацією, що дає змогу вибрати житло, яке буде знаходитися саме у потрібному місці, з потрібною кількістю кімнат та потрібними розмірами.

Для чого потрібна повна фотоінформація. Завдяки цьому користувач може вибрати житло у будь-якому місці без потреби виїзду та оцінки візуального стану, завдяки онлайн-формату це можна зробити з дому.

За допомогою великих даних можна провести аналіз та рекомендувати житло. Завдяки аналізу користувачам можна підбирати рекомендації за його минулими пошуками та вподобаннями, що дозволяє швидше знаходити варіанти, які ймовірніше йому сподобаються.

Отже, аналіз великих даних у справах оренди та купівлі житла може допомогти у пришвидшенні та оптимізації пошуку. Використовуючи різні алгоритми рекомендацій, можна створити додаток, який буде автоматично прогнозувати вартість житла, що буде корисним як продавцям, так і покупцям. Методи й алгоритми для аналізу даних, які використовуються в цій роботі, базуються на лінійній регресії та використанні методу градієнтного спуску для оптимізації моделі.

Градієнтний спуск – це алгоритм оптимізації, який використовується для мінімізації певної функції шляхом ітераційного переміщення в напрямку найкрутішого спуску, визначеного негативним значенням градієнта. У машинному навчанні градієнтний спуск використовується для оновлення параметрів моделі.

Лінійна регресія – це метод моделювання залежності між скалярною змінною Y та векторною (у загальному випадку) змінною X. У разі, якщо змінна X також є скаляром, регресію називають простою. Наприклад, користувач вводить розміри будинків X і завдяки навченій моделі отримує приблизну вартість Y. Для тренування моделі використовується набір даних з розмірами будинків X і відповідною вартістю Y, завдяки градієнтному спуску модель адаптує нахил і зрушення лінії лінійної регресії (рис. 1), що допомагає знайти залежність ціни від розмірів будинку. Метод містить математичну функцію f із входом X і виходом Y.

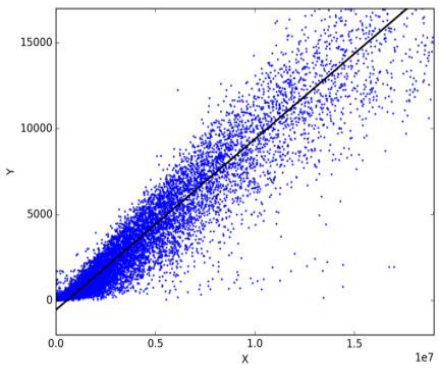

Рис. 1. Приклад навченої моделі лінійної регресії

Для тренування моделі лінійної регресії використовується метод градієнтного спуску. Суть методу полягає у тому, що ми обираємо початкову точку і від неї крок за кроком рухаємося за антиградієнтом, якщо нас цікавить локальний мінімум функції, або ж за градієнтом, якщо нас цікавить максимум функції.

Для того щоб визначити напрямок руху та розмір кроку, потрібно обчислити похідну функції у даній точці, фактично вона дорівнюватиме відношенню приросту функції до приросту значення аргументу.

Отримавши результат, розуміємо (залежно від того чи позитивний він, чи негативний) напрям, та величину, чим крутіший нахил дотичної у цій точці, тим більше значення будемо отримувати. Тут потрібно враховувати проблематику затухання та вибухання градієнта, для цього, щоб реалізувати правильну роботу методу, вводимо поняття «навчального рейтингу».

Навчальний рейтинг – певне число, на яке будемо множити отриманий результат з похідної, для того щоб чим ближче ми були до мінімуму функції, тим менші кроки виходили у результаті.

Також зауважимо, що чим ближче ми до мінімуму функції, тим ближче, за значенням, до нуля буде похідна, і тим менший буде нахил дотичної.

Завдяки цьому алгоритму ми можемо знайти мінімум певної функції (це може бути функція втрат) і отримати, наприклад, точку (значення певного параметра), у якій значення функції втрат, для певного алгоритму машинного навчання, є мінімальним, що дозволить оптимізувати алгоритм.

Найбільш поширена функція втрат середньоквадратичної помилки. У цьому випадку мінімізуватися буде функція втрат, яка розраховується за формулою (1).

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m} \quad (h_0(x^{(i)}) - y^{(i)})^2. \tag{1}$$

У параметрах цієї функції є нахил на зрушення лінії лінійної регресії, яка випливає з формули лінії (2).

$$h_0(x) = \theta_0 + \theta_1 x. \tag{2}$$

Фактично рахуємо суму різниці квадрата передбачення моделі та тренувального значення. Після цього ділимо цю суму на кількість прикладів, помножену на два, множимо на два для зменшення числа результату і зручності роботи з ним у подальших розрахунках.

Завдяки двом параметрам можемо реалізовувати візуалізацію роботи градієнтного спуску по функції втрат та візуально оцінити результат навчання моделі, що зображено на рисунку 2.

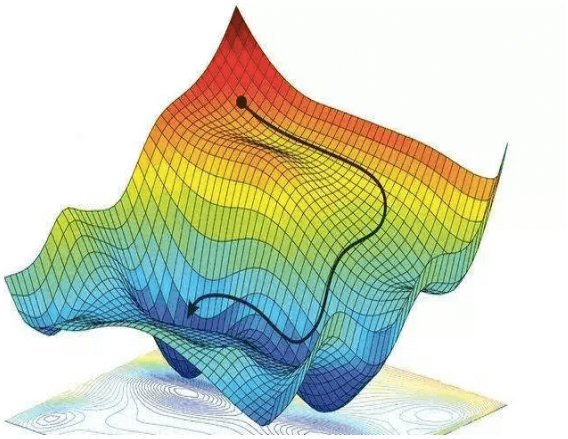

Рис. 2. Візуалізація градієнтного спуску по функції помилки за двома параметрами

Характеристика джерела даних для проведення аналізу. Дані були завантажені з доступних інтернет-ресурсів продажу та оренди житла. Завантажимо та переглянемо набір даних.

size in feet²	price in $1000's
2104	400
1416	232
1534	315
852	178
...	...
3210	870

Примітки: стовпець «size in feet²» – розмір будинку; стовпець «price in $1000's» – вартість.

Рис. 3. Перегляд набору даних

Побудова моделі аналізу і практична реалізація. Для початку відобразимо дані про будинки та їхню ціну на графіку (рис. 4).

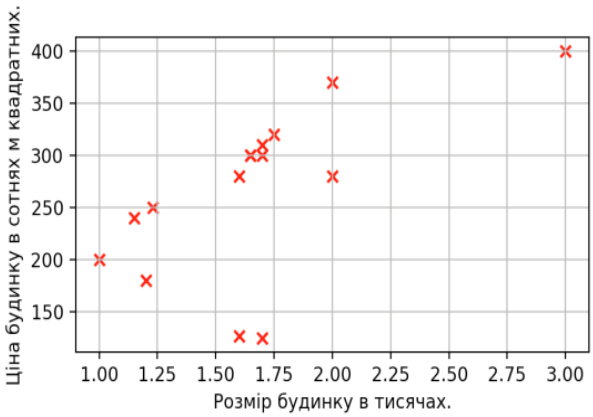

Рис. 4. Відображення набору даних

Для тренування моделі лінійної регресії будемо використовувати метод градієнтного спуску. Але спершу побудуємо довільну лінію моделі лінійної регресії та порахуємо функцію втрат, а також відобразимо лінію моделі у довільних значення нахилу та зрушення на рисунку 5. Для спрощення було взято нульові значення.

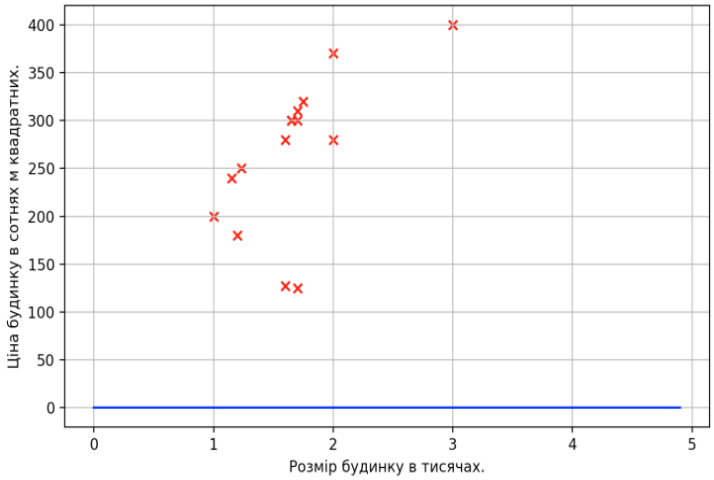

Рис. 5. Відображення довільної лінії моделі лінійної регресії

На рисунку 6 представлено 3D-візуалізацію функції похибки та її значення в точці з проєкції нахилу та зрушення.

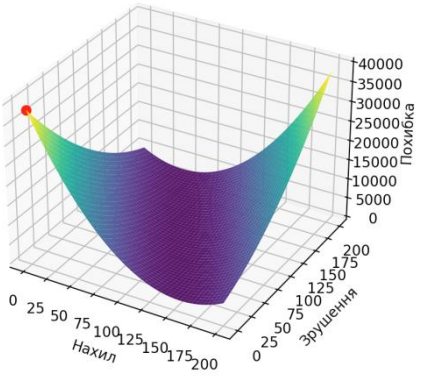

Рис. 6. Відображення функції похибки та значення в точці з проєкцією нахилу та зрушення

Відобразимо проєкцію функції помилки та значення помилки в точці при цьому нахилу та зрушення лінії на рисунку 7.

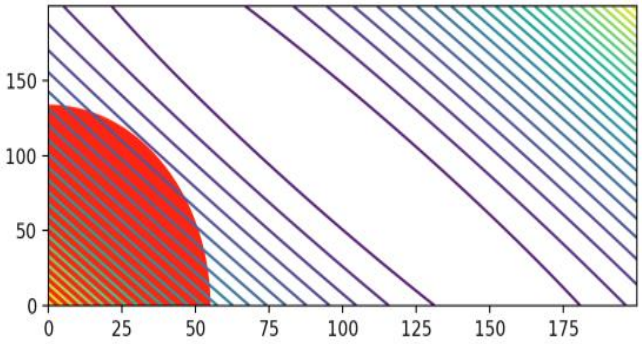

Рис. 7. Проєкції функції похибки та значення в точці з відображенням нахилу та зрушення лінії

Застосуємо метод градієнтного спуску і відобразимо результат його роботи. На рисунку 8 бачимо покрокову роботу методу градієнтного спуску та покрокову зміну нахилу та зрушення лінії лінійної регресії у бік покращення прогнозів моделі лінійної регресії.

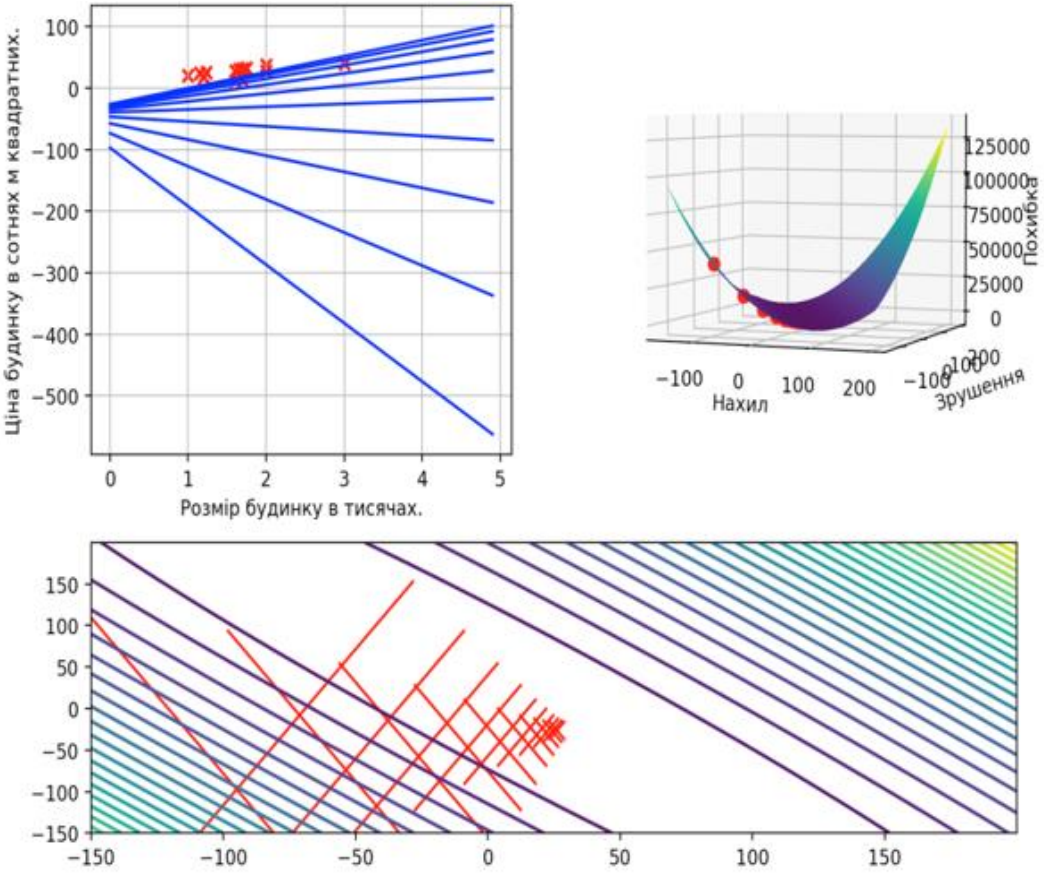

Рис. 8. Візуалізація покрокової роботи методу градієнтного спуску

Аналіз результатів. Результат роботи програми відображено на рисунку 9, де чітко видно, що модель підлаштувалася під тренувальні дані і готова виконувати своє завдання, протестувавши її, користувач зможе виявити приблизну вартість свого будинку залежно від його розмірів. Функція помилки мінімізована і відповідає вимогам поставленого завдання.

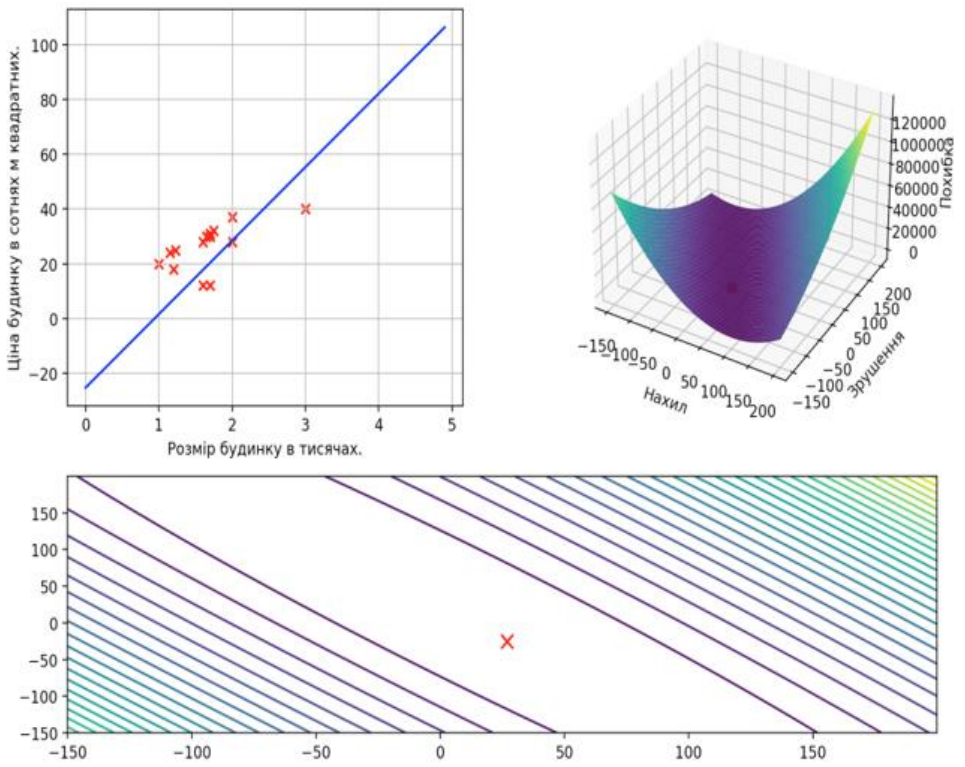

Рис. 9. Візуалізація результатів

Висновки та перспективи подальших досліджень. Для виконання поставлених завдань було обрано проведення інтелектуального аналізу шляхом методів та алгоритмів лінійної регресії та градієнта. Було проаналізовано вхідні дані про вартість та розміри будинків. Побудовано моделі аналізу для обраних алгоритмів та графіки прогнозу моделі. В результаті було розроблено модель аналізу вхідних даних про ціну будинків залежно від їх розміру. Досліджено взаємозв'язок ознак, що допоможе в майбутньому підібрати оптимальну ціну для будинку. В найближчій перспективі планується збільшити кількість вхідних даних для проведення аналізу. Як додаткові критерії можуть бути місце розташування, кількість кімнат, величина земельної ділянки.

Список використаної літератури:

1. *Levkivskyi V.* Research of algorithms of Data Mining. E3S Web of Conferences / *V.Levkivskyi, N.Lobanchykova, D.Marchuk* // The International Conference on Sustainable Futures: Environmental, Technological, Social and Economic Matters (ICSF 2020). – 2020. – Vol. 166. DOI: 10.1051/e3sconf/202016605007.
2. *Yao W.* A New Regression Model: Modal Linear Regression / *W.Yao, L.Li* // Scandinavian Journal of Statistics. – 2014. – № 41. – P. 656–671. DOI: 10.1111/sjos.12054.
3. *Opmane I.* Use of Cluster Analysis in Exploring Economic Indicator Differences among Regions: The Case of Latvia / *I.Opmane* // Journal of Economics, Business and Management. – 2013. – № 1 (1). – P. 42–45.
4. *Paianok T.* Cluster analysis of labor potential of Ukraine / *T.Paianok, Y.Vazhaliuk* // Economy and State. – 2019. – № 12. – P. 109–114.
5. *Behun S.* Application of cluster analysis to study the demographic situation in the region / *S.Behun* // Economic Journal of Lesya Ukrainka East European National University. – 2016. – № 2. – P. 122–128.
6. *Synytsia S.* Clustering of regions by level of economic potential / *S.Synytsia, O.Vakun* // Economy and society Mukachevo State University. – 2017. – № 12. – P. 776–784.
7. *Potapova N.* Clustering of economic regions of Ukraine in terms of innovation and research / *N.Potapova*. – Lviv : Polytechnic National University Institutional Repository, 2010. – P. 33–39.
8. *Riadno O.* Research of structure and dynamics of differentiation of social and economic development of regions of Ukraine on the basis of the cluster analysis / *O.Riadno, O.Berkut* // Economic bulletin of Donbass. – 2016. – № 1 (43). – P. 60–67.
9. *Zomchak L.* Clustering of regions of Ukraine by competitiveness / *L.Zomchak, Y.Dobrotii* // Administrative-territorial vs economicspatial borders of regions : proceedings of the International scientific-practical conference. – KNEU, 2020. – P. 328–332.
10. Cluster analysis of Ukrainian regions regarding the level of investment attractiveness in tourism / *G.Kharlamova, A.Roskladka, N.Roskladka and other* // Proceedings of the 17th International Conference on ICT in Education,

Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. II, 2021. – P. 401–416 [Electronic resource]. – Access mode : http://icteri.org/icteri-2021/proceedings/volume-2/202110401.pdf.
11. Алгоритмічно-програмне забезпечення обробки та аналізу потоку кадрів відеоданих, що надходять з камер міста : комп'ютерна програма / *В.Л. Левківський, Г.В. Марчук, В.В. Ципоренко, Д.К. Марчук.* – 2021 [Electronic resource]. – Access mode : http://eztuir.ztu.edu.ua/bitstream/handle/123456789/8019/109822.pdf?sequence=1&isAllowed=y.
12. Available parking places recognition system / *V.Levkivskyi, D.Marchuk, N.Lobanchykova and other* // CEUR Workshop Proceedings 4th Workshop for Young Scientists in Computer Science & Software Engineering. – 2022. – Vol. 3077. – P. 123–134 [Electronic resource]. – Access mode : http://ceur-ws.org/Vol-3077/paper07.pdf.
13. *Granik M.* Fake news detection using naive Bayes classifier / *M.Granik, V.Mesyura* // 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017. – P. 900–903. DOI: 10.1109/UKRCON.2017.8100379.
14. Analysis of the Impact of Intangible Assets on the Companies' market Value / *V.Ievdokymov, T.Ostapchuk, S.Lehenchuk and other* // Naukovyi Visnyk Natsional'nogo Hirnychogo Universytety. – 2020. – № 3. – P. 164–170.

References:

1. Levkivskyi, V., Lobanchykova, N. and Marchuk, D. (2020), «Research of algorithms of Data Mining. E3S Web of Conferences», *The International Conference on Sustainable Futures: Environmental, Technological, Social and Economic Matters (ICSF 2020)*, Vol. 166, doi: 10.1051/e3sconf/202016605007.
2. Yao, W. and Li, L. (2014), «A New Regression Model: Modal Linear Regression», *Scandinavian Journal of Statistics*, No. 41, pp. 656–671, doi: 10.1111/sjos.12054.
3. Opmane, I. (2013), «Use of Cluster Analysis in Exploring Economic Indicator Differences among Regions: The Case of Latvia», *Journal of Economics, Business and Management*, No. 1 (1), pp. 42–45.
4. Paianok, T. and Vazhaliuk, Y. (2019), «Cluster analysis of labor potential of Ukraine», *Economy and State*, No. 12, pp. 109–114.
5. Behun, S. (2016), «Application of cluster analysis to study the demographic situation in the region», *Economic Journal of Lesya Ukrainka East European National University*, No. 2, pp. 122–128.
6. Synytsia, S. and Vakun, O. (2017), «Clustering of regions by level of economic potential», *Economy and society Mukachevo State University*, No. 12, pp. 776–784.
7. Potapova, N. (2010), *Clustering of economic regions of Ukraine in terms of innovation and research*, Polytechnic National University Institutional Repository, Lviv, pp. 33–39.
8. Riadno, O. and Berkut, O. (2016), «Research of structure and dynamics of differentiation of social and economic development of regions of Ukraine on the basis of the cluster analysis», *Economic bulletin of Donbass*, No. 1 (43), pp. 60–67.
9. Zomchak, L. and Dobrotii, Y. (2020), «Clustering of regions of Ukraine by competitiveness», *Administrative-territorial vs economicspatial borders of regions*, proceedings of the International scientific-practical conference, KNEU, pp. 328–332.
10. Kharlamova, G., Roskladka, A., Roskladka, N. et al. (2021), «Cluster analysis of Ukrainian regions regarding the level of investment attractiveness in tourism», *Proceedings of the 17th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. II*, pp. 401–416, [Online], available at: http://icteri.org/icteri-2021/proceedings/volume-2/202110401.pdf
11. Levkivs'kyj, V.L., Marchuk, G.V., Cyporenko, V.V. and Marchuk, D.K. (2021), *Algorytmichno-programne zabezpechennja obrobky ta analizu potoku kadriv videodanyh, shho nadhodjat' z kamer mista*, komp'juterna programa, [Online], available at: http://eztuir.ztu.edu.ua/bitstream/handle/123456789/8019/109822.pdf?sequence=1&isAllowed=y
12. Levkivskyi, V., Marchuk, D., Lobanchykova, N. et al. (2022), «Available parking places recognition system», *CEUR Workshop Proceedings 4th Workshop for Young Scientists in Computer Science & Software Engineering*, Vol. 3077, pp. 123–134, [Online], available at: http://ceur-ws.org/Vol-3077/paper07.pdf
13. Granik, M. and Mesyura, V. (2017), «Fake news detection using naive Bayes classifier», *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pp. 900–903, doi: 10.1109/UKRCON.2017.8100379.
14. Ievdokymov, V., Ostapchuk, T., Lehenchuk, S. et al. (2020), «Analysis of the Impact of Intangible Assets on the Companies' market Value», *Naukovyi Visnyk Natsional'nogo Hirnychogo Universytety*, No. 3, pp. 164–170.

Годлевський Юрій Олександрович – Senior Software Engineer, компанія «Infopulse».
Наукові інтереси:
– аналіз даних.

Марчук Галина Вікторівна – старший викладач кафедри комп'ютерних наук Державного університету «Житомирська політехніка».
https://orcid.org/0000-0003-2954-1057.
Наукові інтереси:
– аналіз даних.

Панаріна Ірина Володимирівна – кандидат технічних наук, доцент кафедри комп'ютерних наук Державного університету «Житомирська політехніка».

https://orcid.org/0000-0003-4783-2587.

Наукові інтереси:

– аналіз даних.

Hodlevskyi Yu.O., Marchuk G.V., Panarina I.V.

Analysis, modeling and forecasting the price of houses, depending on their size

The goal of data analysis is to obtain information which is not easy to interpret, but will help to use it correctly. The paper describes a new approach to the use of intelligent technologies for certain business decisions, namely to research the pricing policy of the cost of houses based on their size. The object of the research is data analysis about available houses for sale, their size and value. The subject of the research are methods, algorithms and means of intelligent data analysis. The article considers the most simple and understandable, at the same time, widely used method of mathematical programming for solving problems of this kind – the gradient descent. The work uses the RMSE loss function. The derivative of this function shows the computation of the gradient for linear regression. The work uses linear regression i.e. a model of the dependence of variables (cost and area of the house) with a linear function of dependence. Data for analysis were downloaded from several Internet resources for the sale and rental of housing. The paper presents a visualization of the gradient descent method on the error function with two parameters. In addition to 3D plots, contour plots are presented. A contour plot is a method of representing a three-dimensional image in a two-dimensional plane, which well shows the step-by-step operation of the gradient descent method. As a result, a model was designed, where you can observe how the model adapts to the training data and is ready to perform its task. After testing the proposed model, you can determine the approximate cost of the house depending on its size. The error function is minimized and satisfies the requirements of the given task. In the future, it is planned to increase the amount of input data for analysis, indicating the location, the number of rooms, and the size of the surrounding area.

Keywords: data analysis; linear regression model; machine learning; artificial intelligence.

Стаття надійшла до редакції 27.09.2022.