

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ПЕРЕДБАЧЕННЯ ВАРТОСТІ БУДИНКУ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

У роботі виконано попередню підготовку та розвідувальний аналіз даних для створення інформаційної технології передбачення вартості житлової нерухомості за допомогою методів машинного навчання. Аналізовано набір даних Housing Prices Dataset, який містить технічні та соціально-економічні характеристики будинків. Виділено основні фактори, що впливають на ціну, зокрема площу, кількість кімнат, наявність кондиціонера, підвалу тощо.

Ключові слова: машинне навчання, ціна будинку, аналіз даних, нерухомість, передбачення, ознаки, розвідувальний аналіз.

Abstract

The paper presents the preliminary preparation and exploratory data analysis for building an information technology to predict residential real estate prices using machine learning methods. The Housing Prices Dataset, which includes technical and socio-economic characteristics of houses, was analyzed. Key features affecting price were identified, such as area, number of rooms, presence of air conditioning, basement, and others.

Keywords: machine learning, house price, data analysis, real estate, prediction, features, exploratory analysis.

Вступ

У сучасних умовах ринку нерухомості точне передбачення вартості житла має велике практичне значення як для покупців, так і для продавців. Завдяки розвитку інформаційних технологій і методів машинного навчання, з'являється можливість створювати моделі, здатні автоматично оцінювати вартість житла на основі набору параметрів.

У цьому дослідженні здійснено розвідувальний аналіз даних, що містять характеристики будинків та їхню фактичну ціну. Метою аналізу є виявлення ключових ознак, які мають найбільший вплив на формування ціни, з подальшим використанням цих даних для побудови моделей передбачення.

Розвідувальний аналіз

Для проведення аналізу було обрано набір даних під назвою Housing Prices Dataset, який має відкритий доступ для загального використання на платформі Kaggle [1]. Даний датасет містить інформацію про вартість будинків, а також низку матеріальних показників, що потенційно впливають на цю вартість. Загалом датасет містить 545 записів про житлову нерухомість та має 13 ключових ознак, що використовуються для розрахунку вартості будинку (рис. 1).

Основні атрибути датасету:

- Price — ціна будинку;
- Area — площа будинку (в квадратних футах);
- Bedrooms, Bathrooms, Guestroom, Stories — кількість спальних кімнат, ванних кімнат, наявність гостьових кімнат та кількість поверхів;
- Mainroad, Bbasement, Hotwaterheating, Airconditioning, Prefarea — наявність інфраструктурних переваг, як от сполученість з головною дорогою, наявність гарячої води, наявність кондиціонерів, та чи будинок розташований у престижному районі;
- Parking, Basement — кількість місць для паркування;
- Furnishingstatus — рівень меблювання (мебльовані, напів-мебльовані, не мебльовані);

Цей набір даних надає достатньо можливостей для проведення глибокого аналізу, визначення основних факторів впливу на вартість будинку та побудови моделей передбачення із застосуванням методів машинного навчання.

data

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	par
0	13300000	7420	4	2	3	yes	no	no	no	yes	
1	12250000	8960	4	4	4	yes	no	no	no	yes	
2	12250000	9960	3	2	2	yes	no	yes	no	no	
3	12215000	7500	4	2	2	yes	no	yes	no	yes	
4	11410000	7420	4	1	2	yes	yes	yes	no	yes	
...	...	...	...	...	...	...	...	...	...	...	...
540	1820000	3000	2	1	1	yes	no	yes	no	no	
541	1767150	2400	3	1	1	no	no	no	no	no	
542	1750000	3620	2	1	1	yes	no	no	no	no	
543	1750000	2910	3	1	1	no	no	no	no	no	
544	1750000	3850	3	1	2	yes	no	no	no	no	

545 rows × 13 columns

Рис. 1. Приклад даних із датасету «Housing Prices Dataset»

Було виконано попереднє очищення та об’єднання даних та проведено розвідувальний аналіз даних, в ході якого було створено деякі візуалізації, наведені нижче (рис. 2–6).

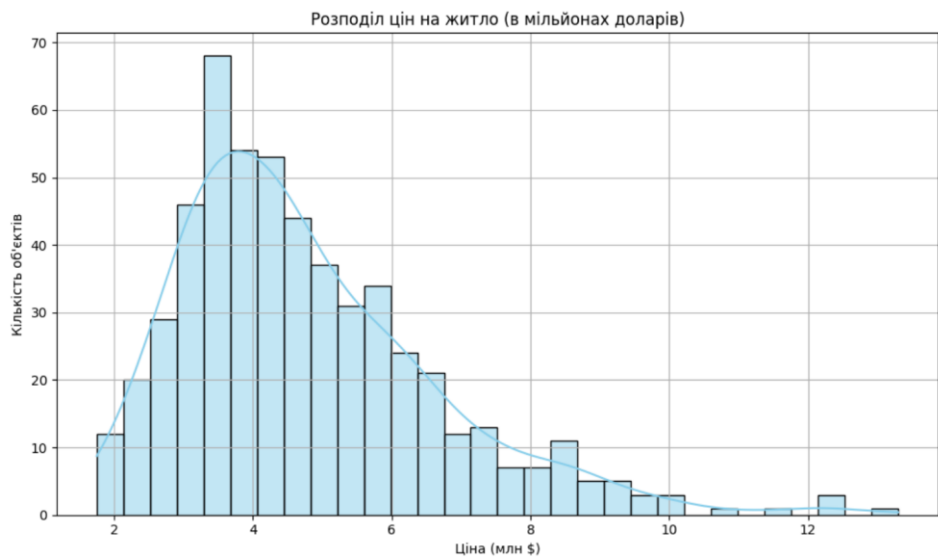


Рис. 2. Гістограма розподілу цін на житло (в млн. доларах)

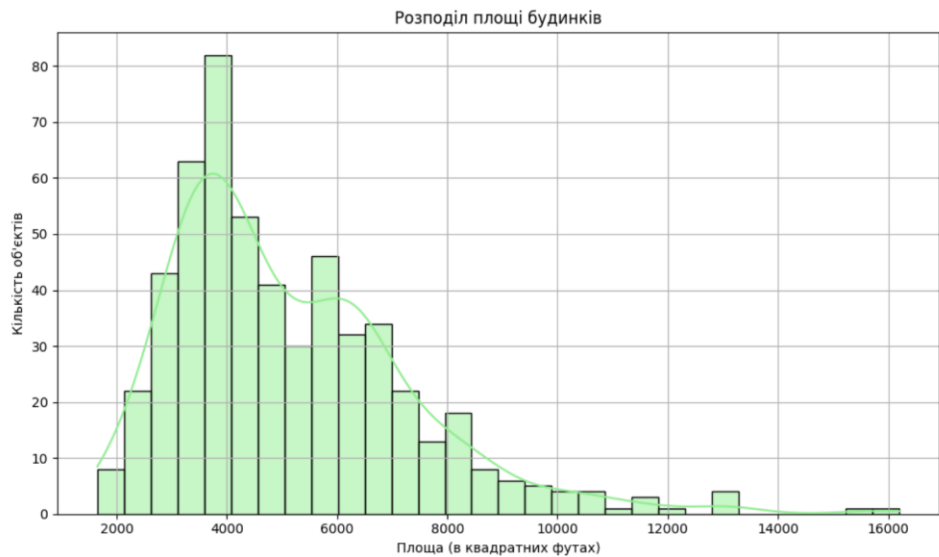


Рис. 3. Гістограма розподілу площі будинків

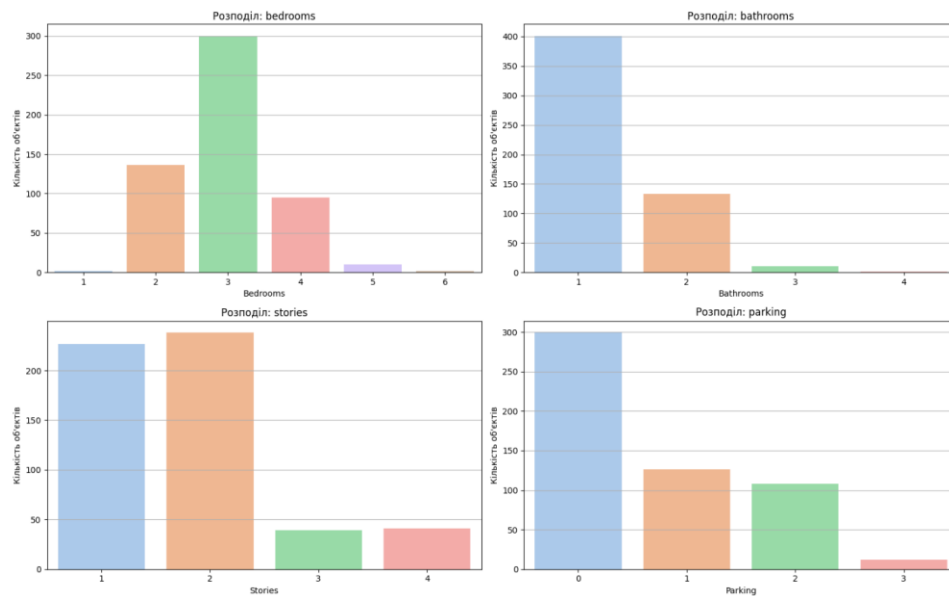


Рис. 4. Розподіл числових ознак

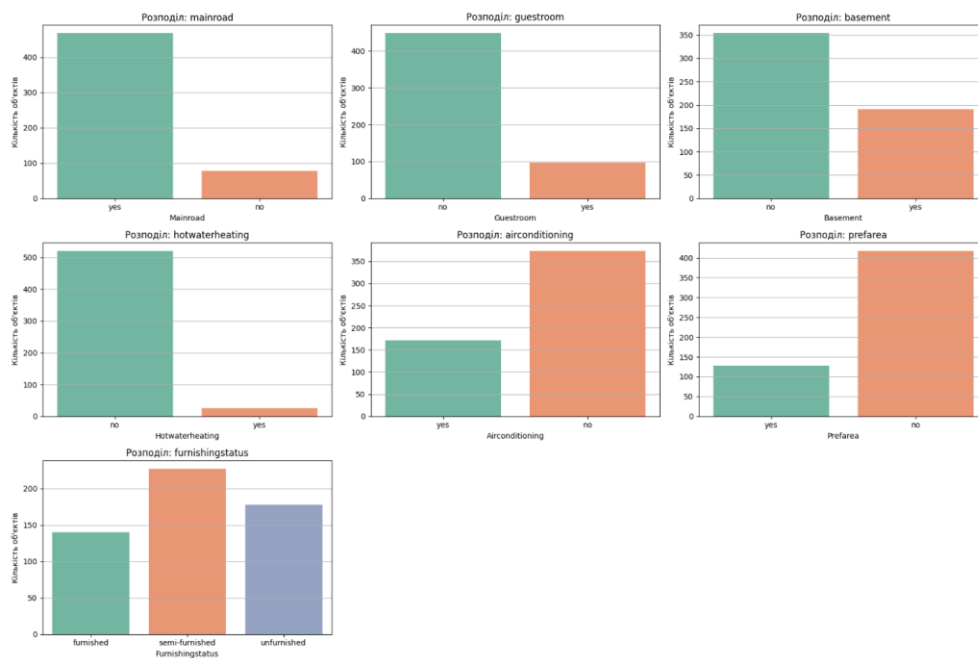


Рис. 5. Розподіл категоріальних ознак

Наведені гістограми надають візуальне уявлення про розподіл ключових факторів, що впливають на вартість будинків.

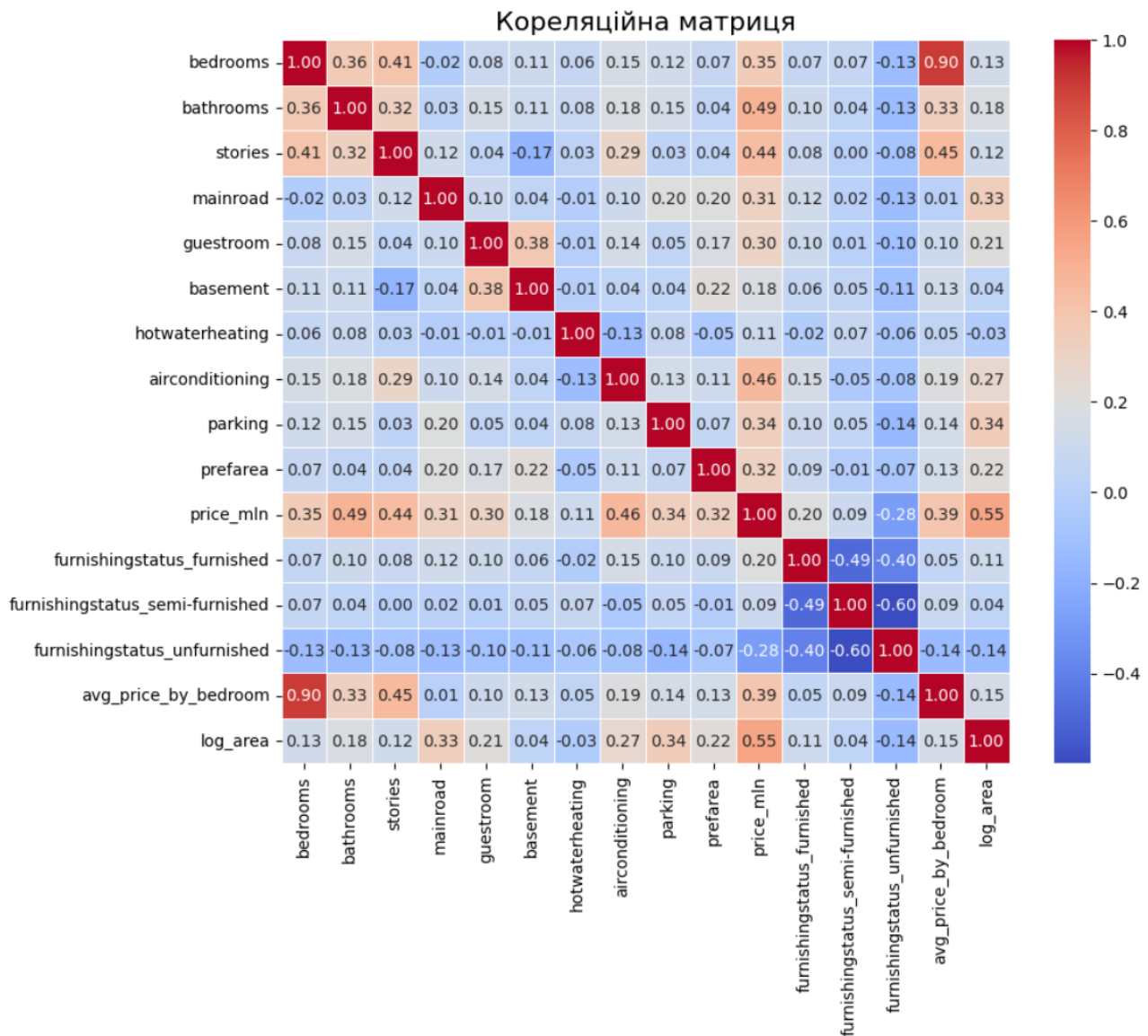


Рис. 6. Матриця кореляції

З наведеної матриці кореляції можна зробити висновок, що вартість будинку найбільше корелює з площею будинку, кількістю ванних кімнат, наявністю кондиціонерів та мебльованості. Кількість поверхів, наявність підвалу, виходу до головної дороги також відіграють важливу, але меншу роль. Кількість спалень, наявність гостьової кімнати та кількість паркомісць демонструють найслабший зв'язок з іншими досліджуваними показниками.

Висновки

При розвідувальному аналізі набору даних «Housing Prices Dataset», що містить інформацію про характеристики будинків та їх вартість, було досліджено вплив рівня особливості кожного будинку на його вартість.

Було досліджено розподіл усіх необхідних параметрів та гістограм (Рис. 2-5). Аналіз гістограм показав, що розподіл вартості будинку напряму залежить від площі будинку.

Також було побудовано матрицю кореляції, яка показала залежність між різними особливостями кожного будинку, його матеріальними ознаками та вартістю будинку. Аналізом матриці встановлено, що вартість будинку формується під впливом сукупності як технічних характеристик (площа, кількість поверхів, кількість санвузлів), так і соціальних (наявність меблів, розташування, зручності). На кореляційній матриці чітко простежується тенденція високої залежності вартості будинку від площі будинку, кількості ванних кімнат, наявності кондиціонера та статусу меблювання. Отримані результати дозволяють сформулювати основу для побудови моделі машинного навчання для прогнозування вартості нерухомості.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Housing Prices Dataset [Електронний ресурс]. – Режим доступу: <https://www.kaggle.com/datasets/yasserh/housing-prices-dataset/data>
2. Pandas Getting started. 2024 [Електронний ресурс] – Режим доступу: https://pandas.pydata.org/docs/getting_started/index.html
3. Matplotlib Pyplot Documentation. 2024 [Електронний ресурс]. – Режим доступу: https://matplotlib.org/3.5.3/api/as_gen/matplotlib.pyplot.html
4. Seaborn Tutorial. 2023 [Електронний ресурс]. – Режим доступу: <https://seaborn.pydata.org/tutorial.html>

Янковчук Михайло Ігорович – студент групи СА-23б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: mykhailoyanki@gmail.com.

Жуков Сергій Олександрович – к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, e-mail: sazhukov@gmail.com.

Yankovchuk Mykhailo I. – student of Faculty of Intelligent Information Technology and Automation, SA-23b, Vinnytsia National Technical University, Vinnytsia, e-mail: mykhailoyanki@gmail.com.

Zhukov Serhii O. - Ph.D., Assistant Professor of the Department of Systems Analysis and Information Technology, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@gmail.com.