https://doi.org/10.31891/2219-9365-2024-78-45
УДК 004.8:004.6

ВАЙС Тімея
ДВНЗ «Ужгородський національний університет»
https://orcid.org/0009-0005-4520-858X
vais.timeia@student.uzhnu.edu.ua
ОНИЩАК Назарій
ДВНЗ «Ужгородський національний університет»
https://orcid.org/0000-0002-3321-5986
nazarii.onyshchak@uzhnu.edu.ua
ПОЛОВКО Іван
ДВНЗ «Ужгородський національний університет»
https://orcid.org/0000-0001-7128-4846
polovko.ivan@uzhnu.edu.ua
ШАРКАДІ Маріанна
ДВНЗ «Ужгородський національний університет»
https://orcid.org/0000-0002-1850-996X
marianna.sharkadi@uzhnu.edu.ua

ВИКОРИСТАННЯ ГЕНЕРАТИВНОГО ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ АНАЛІЗУ ДАНИХ

У 2022 році розпочався новий етап розвитку штучного інтелекту. Багато компаній почали публікувати свої напрацювання та методології у цій сфері. Бурхливий розвиток дав можливість автоматизувати багато процесів рутинних справ у різних сферах життя та діяльності. Не стала винятком і сфера науки про дані та їхній аналіз. Актуальність задачі полягає у тому, щоб показати, що штучний інтелект є корисним у використанні для завдань аналізу даних. Завдяки цьому інструменту можна поглянути на задачу з іншої точки зору. В ході роботи завдяки генеративному штучному інтелекту було обрано набір даних, який зміг підібрати датасет так, щоб можна було розкрити метод кластеризації, прогнозування, а також ансамблевий метод машинного навчання. Дане дослідження показує, що використання штучного інтелекту є корисним, допомагає та дає поради щодо створення та ходу розробки коду.

Ключові слова: ChatGPT, штучний інтелект, метод ліктя, сегментація, кластеризація, лінійна регресія

VAIS Timeia, ONYSHCHAK Nazarii, POLOVKO Ivan, SHARKADI Marianna
Uzhhorod National University

USING GENERATIVE ARTIFICIAL INTELLIGENCE FOR DATA ANALYSIS

The optimization and development of artificial intelligence algorithms open up endless possibilities for improving the processes of machine learning and data analytics. In this context, ChatGPT, based on advanced deep learning technologies, acts as a key tool. Its powerful potential in understanding and generating natural language makes it an effective tool for a variety of data mining tasks.

Reflecting the rapid advancement of technology, machine learning and data analytics are becoming not only important but integral components in today's world. They allow companies to use vast amounts of data to obtain valuable information and make informed decisions. In this context, intelligent models such as ChatGPT become a crystal ball, providing the ability to understand and generate text in natural language, which opens new horizons for performing analysis and interaction with data.

In 2022, a new stage of artificial intelligence development began. Many companies started publishing their developments and methodologies in this area. The rapid development has made it possible to automate many routine processes in various spheres of life and activity. The field of data science and analysis is no exception. The task is to show that artificial intelligence is useful for data analysis tasks. Thanks to this tool, you can look at the task from a different perspective. In the course of the work, thanks to generative artificial intelligence, a dataset was selected that was able to select a dataset so that the clustering method, forecasting, and ensemble machine learning method could be revealed. This study shows that the use of artificial intelligence is useful, helps and gives advice on the creation and progress of code development.

Keywords: ChatGPT, artificial intelligence, elbow method, segmentation, clustering, linear regression.

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Оптимізація та розвиток алгоритмів штучного інтелекту відкривають безмежні можливості для вдосконалення процесів машинного навчання та аналітики даних. У цьому контексті ChatGPT, що базується на передових технологіях глибинного навчання, виступає як ключовий інструмент. Його потужний потенціал у розумінні та генерації природної мови робить його ефективним інструментом для різноманітних завдань у сфері аналізу даних.

Віддзеркалюючи стрімкий прогрес технологій, машинне навчання та аналітика даних стають не лише важливими, але й невід'ємними компонентами у сучасному світі. Вони дозволяють компаніям використовувати величезні обсяги даних для отримання цінної інформації та прийняття обґрунтованих

рішень. У цьому контексті інтелектуальні моделі, такі як ChatGPT, стають кришталевою кулею, забезпечуючи здатність розуміти та генерувати текст природною мовою, що відкриває нові горизонти для здійснення аналізу та взаємодії з даними.

АНАЛІЗ ДОСЛІДЖЕНЬ ТА ПУБЛІКАЦІЙ

Наразі існує велика кількість наукових статей, присвячених аналізу даних та машинному навчанню із застосуванням мови програмування Python.

У праці [1] розглядається принцип роботи алгоритмів методу "random forest", пояснюється його функціонування і причини такого підходу, а також проводиться аналіз складності цього методу.

Стаття [2] містить огляд алгоритмів машинного навчання, зокрема лінійної регресії. Автори визначають, для яких задач використовуються ці алгоритми, і наводять приклади популярних методів, таких як лінійна регресія, логістична регресія, кластеризація методом k-means та метод опорних векторів. Лінійній регресії приділяється окрема увага, детально пояснюються незалежні та залежні змінні.

Дослідження у роботі [3] представляє новий підхід до методу k-means, що дозволяє виконувати кластеризацію без початкової ініціалізації параметрів, автоматично знаходячи оптимальну кількість кластерів. Також проаналізовано обчислювальну складність цього підходу.

У статті [4] розглядаються різні методи науки про дані у контексті бізнесу. Автори висвітлюють процес від постановки проблеми до використання результатів для оптимізації прийняття рішень, а також обговорюються методи та прийоми, які застосовуються у бізнес-аналітиці.

Нарешті, праця [5] пропонує детальний математичний аналіз методу лінійної регресії, супроводжений візуалізаціями для кращого розуміння його роботи.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

У роботі запропоновано використання ChatGPT для аналітики даних, зокрема для вибору датасету та розробки коду.

На початку дослідження важливо було обрати правильний датасет. За допомогою ChatGPT було обрано потрібний набір даних з кількох пропонованих.

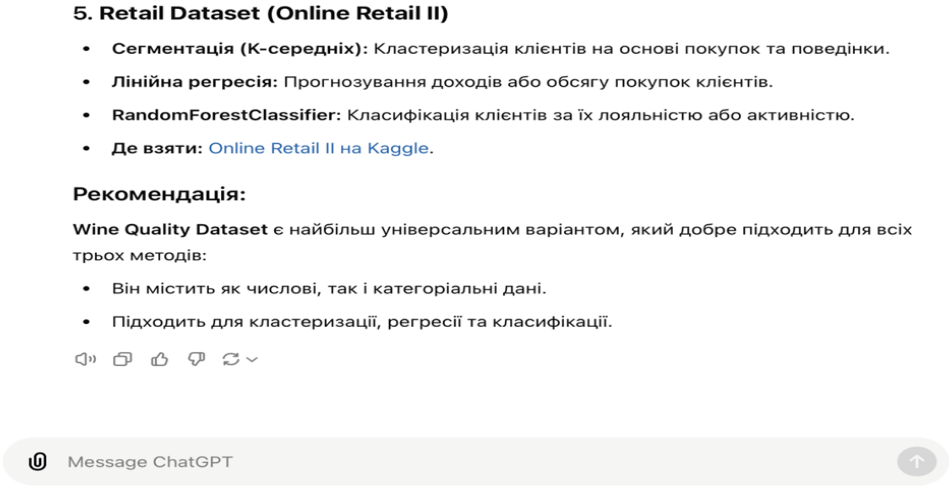


Рис.1. Рекомендації ChatGPT для підбору датасету

На рисунку 1 можна побачити результат роботи зі штучним інтелектом. Після введеного запиту ChatGPT надав кілька варіантів та рекомендації щодо вибору конкретного датасету. Було вирішено використати вказаний набір даних.

Отже, було використано дані з https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009/data. Цей набір даних стосується червоних сортів португальського вина «Vinho Verde».

Даний датасет містить хімічні та сенсорні характеристики червоного вина, а також оцінку його якості, що робить його популярним для аналізу:

1. fixed acidity (фіксована кислотність): Основна частка кислот у вині, вимірюється в г/дм³.
2. volatile acidity (летюча кислотність): Рівень оцтової кислоти у вині, що впливає на його смак, в г/дм³.
3. citric acid (лимонна кислота): Вміст лимонної кислоти, яка додає свіжості вину, в г/дм³.
4. residual sugar (залишковий цукор): Цукор, що залишився після ферментації, вимірюється в г/дм³.
5. chlorides (хлориди): Кількість солі у вині, в г/дм³.
6. free sulfur dioxide (вільний діоксид сірки): Вільні форми SO_2, що запобігають росту мікроорганізмів та окисленню вина, в мг/дм³.

7. total sulfur dioxide (загальний діоксид сірки): Сума всіх форм діоксиду сірки у вині, в мг/дм³.
8. density (густина): Густина вина, вимірюється як маса на об'єм, зазвичай близька до густини води.
9. pH: Рівень кислотності/лужності вина (шкала pH).
10. sulphates (сульфати): Додають гостроту і консерваційні властивості, в г/дм³.
11. alcohol (алкоголь): Відсотковий вміст алкоголю у вині.
12. quality (якість): Оцінка якості вина, від 0 до 10.

У рамках даного дослідження було проведено прогнозування методом лінійної регресії та методом random forest. Окремим етапом дослідження є визначення оптимальної кількості кластерів та проведення кластеризації.

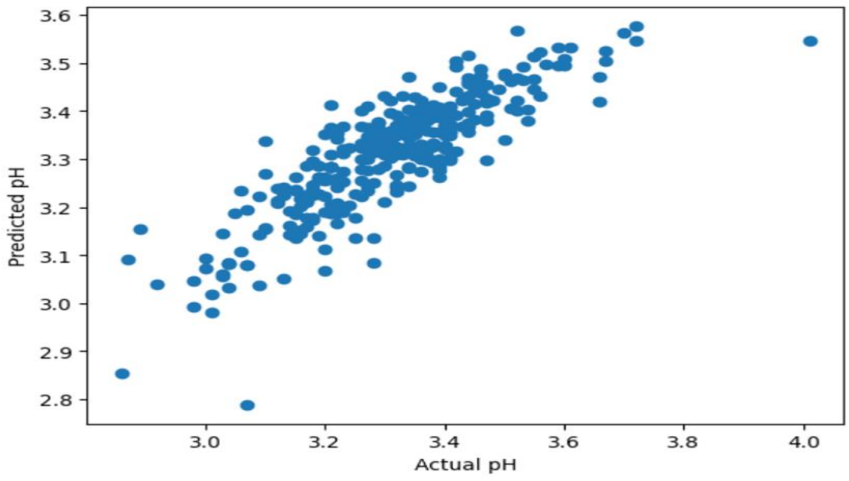
Для досягнення поставленої мети застосовано наступні методологічні підходи. Лінійна регресія – це алгоритм, який використовується для прогнозування або візуалізації зв'язку між двома різними ознаками/змінними. У задачах лінійної регресії досліджуються два типи змінних: залежна змінна і незалежна змінна. Незалежна змінна – це змінна, яка є самостійною, на яку не впливає інша змінна. Коли незалежна змінна коригується, рівні залежної змінної будуть коливатися. Залежна змінна – це змінна, яка вивчається, і це те, що регресійна модель розв'язує/намагається передбачити. У задачах лінійної регресії кожне спостереження/випадок складається як зі значення залежної змінної, так і зі значення незалежної змінної.

Кластеризація K-Means - це алгоритм навчання без вчителя, який групує немаркований набір даних у різні кластери. Тут K визначає кількість заздалегідь визначених кластерів, які потрібно створити в процесі. Це ітераційний алгоритм, який розділяє немаркований набір даних на k різних кластерів таким чином, що кожен набір даних належить лише до однієї групи, яка має схожі властивості. Даний алгоритм побудований на основі центроїдів, де кожен кластер асоціюється з центроїдом. Основною метою цього алгоритму є мінімізація суми відстаней між точками даних та відповідними кластерами.

Random forest додає додаткову випадковість до моделі, вирощуючи дерева. Замість того, щоб шукати найважливішу ознаку (тобто предиктор) при розбитті вузла, він шукає найкращу ознаку серед випадкової підмножини ознак. Тобто, алгоритм розбиття вузла бере до уваги лише випадкову підмножину ознак. Це призводить до широкого розмаїття, що, як правило, призводить до кращої моделі. Наприклад, якщо серед набору предикторів є сильний предиктор, дерево рішень зазвичай покладається на цей конкретний предиктор для прогнозування та побудови дерев. Однак випадкові ліси змушують при кожному розбитті враховувати лише набір предикторів, що призводить до створення дерев, які використовують не лише сильний предиктор, але й інші предиктори, які помірно корелюють з результуючою змінною.

У процесі побудови прогнозних моделей із застосуванням описаних методик ключовим аспектом був правильний вибір набору даних. Це забезпечило отримання очікуваних результатів під час виконання коду, підтверджуючи коректність вибраного набору даних. Даний факт ілюструється на рисунках 2, 3, 4 та 5.

```
plt.scatter(y_test,pred)
plt.xlabel("Actual pH")
plt.ylabel("Predicted pH")
plt.show()
```




```
[ ] mae = metrics.mean_absolute_error(y_test, pred)
    print('Mean Absolute Error:', mae)
```

```
Mean Absolute Error: 0.06275668708161745
```

Рис. 2 Метод лінійної регресії

ISSN 2219-9365

На рисунку 2 чітко продемонстровано, що модель лінійної регресії забезпечує високу точність у прогнозуванні значень pH, характеризуючись низьким значенням середньої абсолютної помилки (MAE). Близькість прогнозованих значень до фактичних свідчить про ефективне навчання моделі на представлених даних.

```
[14] model = RandomForestClassifier()

[15] model.fit(X_train,y_train)

     RandomForestClassifier
     RandomForestClassifier()

[16] pred = model.predict(X_test)

     accuracy_score(y_test, pred)

     0.984375
```

Рис.3 Результати Random forest

На рисунку 3 представлено результати прогнозування за допомогою методу Random Forest. Значення показника точності (accuracy score) для моделі Random Forest Classifier становить 0.984375, що свідчить про високу ефективність моделі, яка коректно класифікувала приблизно 98.44% зразків.

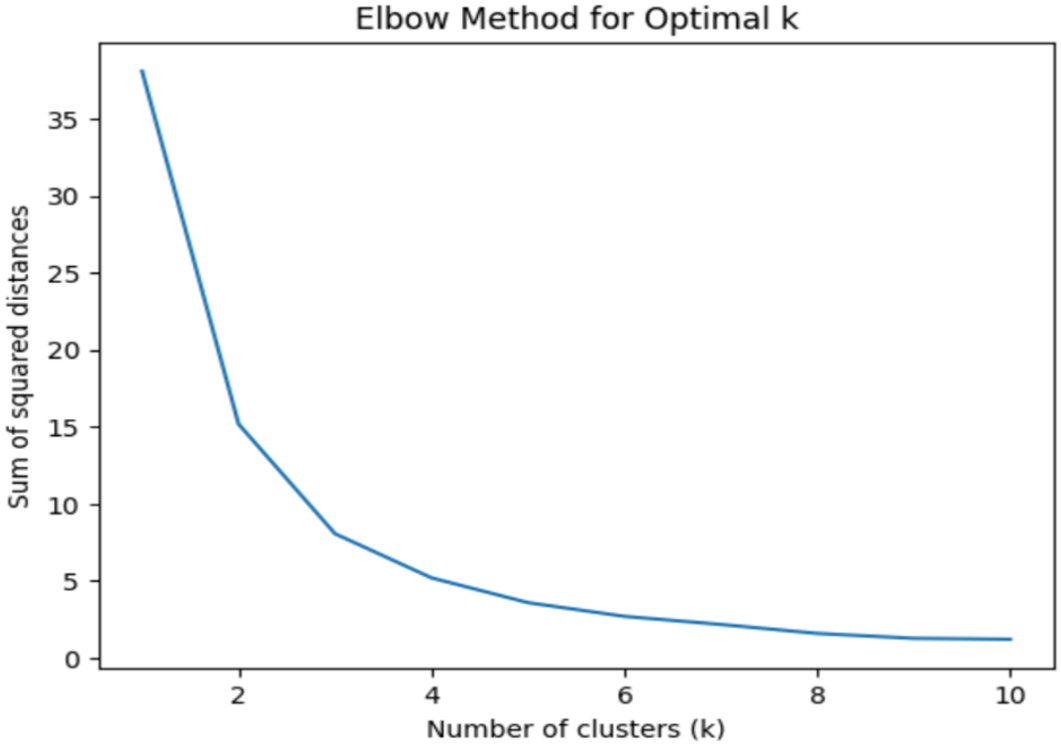


Рис.4 Метод ліктя

Для реалізації методу сегментації необхідно виконати попередній етап визначення оптимальної кількості кластерів. Для цього використано метод ліктя. Як показано на рисунку 4, оптимальним є вибір k = 5, що свідчить про необхідність поділу даних на п'ять кластерів.

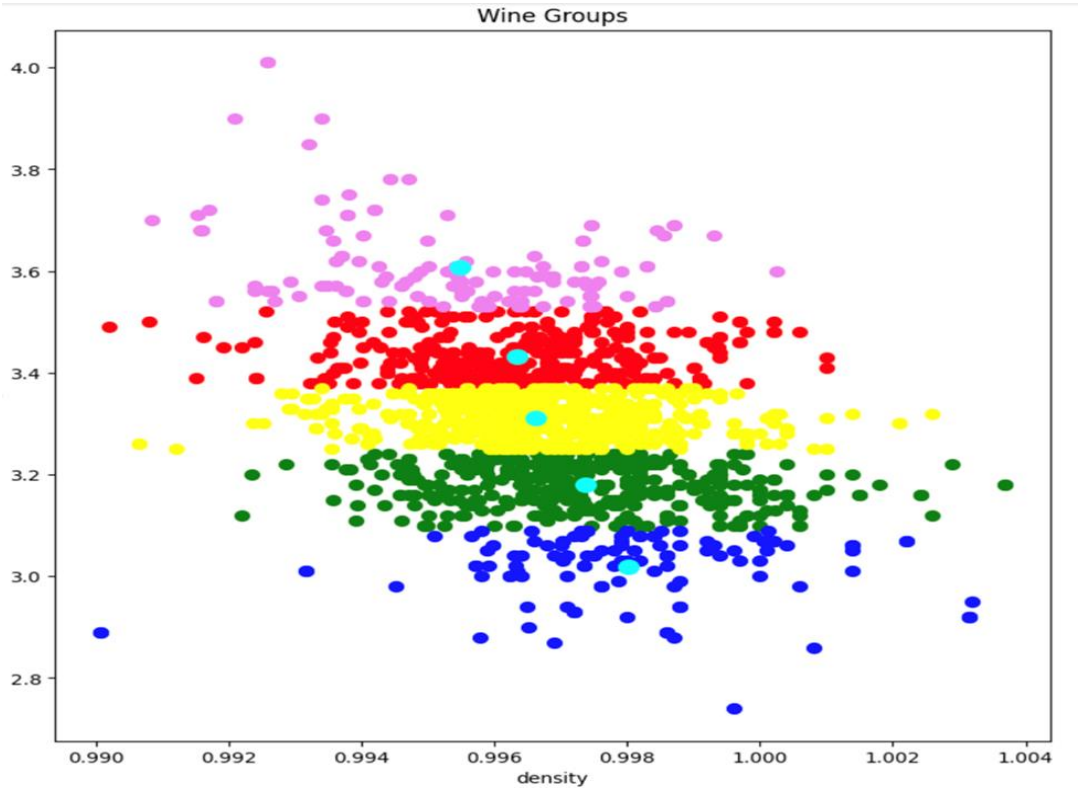


Рис.5 Результат методу k-means

На рисунку 5 продемонстровано результати роботи методу k-means. Виявлено, що кластеризація за допомогою цього методу ефективно сегментує дані за параметрами pH та густина вина. Кластеризація формує чітко виражені групи з центроїдами, які відображають середні значення кожного кластеру.

Вплив параметра pH на процес кластеризації є більш значущим порівняно з густиною, оскільки основні групи формуються вздовж осі pH.

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

Результати цього дослідження демонструють можливість розподілу вибраних даних на п’ять кластерів. Зокрема, показано, що використання ChatGPT значно спрощує процес розробки програмного коду, роблячи його доступним навіть для початківців у цій галузі. ChatGPT надає докладні варіанти рішень, проте для ефективного використання необхідно правильно формулювати запити.

Однією з перспектив подальшого розвитку може бути застосування методів глибинного навчання. Використання штучних нейронних мереж (Artificial Neural Networks, ANN) або глибоких нейронних мереж (Deep Neural Networks, DNN) дозволить моделювати складні нелінійні взаємозв'язки у даних.

Подальші дослідження можуть включати як вдосконалення існуючих підходів, так і впровадження сучасних інструментів штучного інтелекту, таких як глибоке навчання, ансамблеві методи, нові техніки кластеризації, а також підвищення інтерпретованості моделей. Це відкриє нові можливості для аналізу даних та покращення точності прогнозів.

Література

1. Understanding Random Forests: From Theory to Practice. URL: https://www.researchgate.net/publication/264312332_Understanding_Random_Forests_From_Theory_to_Practice
2. Linear Regression in Python. URL: https://www.scribd.com/document/702480908/Linear-Regression-in-Python
3. Unsupervised K-Means Clustering Algorithm. URL: https://www.researchgate.net/publication/340813602_Unsupervised_K-Means_Clustering_Algorithm
4. Provost F., Fawcett T. Data Science for business. 2013. First edition. P. 44 – 71. URL: https://www.advisory21.com.mt/wp-content/uploads/2023/05/Data-Science-for-Business.pdf
5. Regression Analysis URL: https://corporatefinanceinstitute.com/resources/data-science/regression-analysis/

6. Red Wine Quality. URL: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009/data (дата звернення: 03.09.2024)

References

1. Understanding Random Forests: From Theory to Practice. URL: https://www.researchgate.net/publication/264312332_Understanding_Random_Forests_From_Theory_to_Practice
2. Linear Regression in Python. URL: https://www.scribd.com/document/702480908/Linear-Regression-in-Python
3. Unsupervised K-Means Clustering Algorithm. URL: https://www.researchgate.net/publication/340813602_Unsupervised_K-Means_Clustering_Algorithm
4. Provost F., Fawcett T. Data Science for business. 2013. First edition. P. 44 – 71. URL: https://www.advisory21.com.mt/wp-content/uploads/2023/05/Data-Science-for-Business.pdf
5. Regression Analysis URL: https://corporatefinanceinstitute.com/resources/data-science/regression-analysis/
6. Red Wine Quality. URL: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009/data