Вісник Тернопільського національного технічного університету
https://doi.org/10.33108/visnyk_tntu
Scientific Journal of the Ternopil National Technical University
2025, № 4 (120) https://doi.org/10.33108/visnyk_tntu2025.04
ISSN 2522-4433. Web: ***visnyk.tntu.edu.ua***

UDC 004.8

SOME METHODOLOGICAL CONCLUSIONS ABOUT THE GENERATIVE AI MODEL DATA ANALYST

Ivan Laitaruk; Ivan Mamchych

Lesya Ukrainka Volyn National University, Lutsk, Ukraine

Abstract. *This work is devoted to studying the capabilities of the generative AI (GenAI) ChatGPT-4o model Data Analyst (integrated in Chat GPT in August 2025), which has a specialization in the field of data analysis. Currently, the use of large language models (LLM) for various tasks is already a widespread phenomenon. Data scientists are testing the capabilities of these technologies in their field. Available publications and the authors' own experience have shown that general-purpose generative models, such as ChatGPT, provide useful feedback in some cases, and in others – not. Their common conclusions are both positive experience in applying these technologies for a number of typical tasks, and also observations about significant shortcomings for certain cases. These publications inspired us to do our own research. Since these technologies are developing extremely rapidly (on the one hand, algorithms are improving, on the other hand, the base and training time are increasing), it would be interesting to find out the current state of affairs. The purpose of this study was to test the capabilities of the Data Analyst model, specialized for performing statistical methods, machine learning methods and other computational algorithms. The question of whether this tool should be involved in teaching relevant disciplines was also studied. In this paper, we analyze the suitability of the ChatGPT-5 Data Analyst model for training in statistical methods using the example of applying this resource to the clustering problem. The capabilities of Data Analyst are considered for three data sets with different degrees of cluster separation and using different clustering methods, including k-means, single linkage, and the fuzzy clustering method c-means. The possibilities of visualization, code creation in the R program, and interpretation of results are also considered. The paper establishes the capabilities and limitations of this software tool for training in this topic. It was found that the simplest basic tasks Data Analyst has performed quite effectively, while tasks of medium complexity may not be within its power at the moment. For clustering well-separated data, clustering was effectively performed using the k-means and single linkage methods, clusters were visualized, and the working code in the R program was provided. The task was not performed for the c-means method. Regarding the interpretation of the results and comparison of the effectiveness of different methods, the obtained answers can be considered acceptable if we consider these answers as advisory, supporting human decision-making. The result of our study is the conclusion about the need to teach students to use AI for data analysis along with a discussion of its limitations, consequences, ethical aspects and challenges for professionals in this field.*

Key words: *AI, clustering, modelling, statistics, classification, teaching.*

Submitted 21.10.2025
Revised 16.12.2025
Published 27.01.2026

https://doi.org/10.33108/visnyk_tntu2025.04. 130

1. INTRODUCTION

Today, large language models (LLM) have become widespread and have become a factor of influence in many areas, including university education. Although the use of AI models in higher education is still at the stage of testing the possibilities, we can already confidently speak about the undoubted presence of these technologies in the current and future educational process. Both teachers and students are trying to use AI models to perform their tasks, in particular, searching for information, translating, creating tests, completing homework.

Such a large degree of involvement of machine technologies encourages the development of coordinated approaches and the creation of recommendations for the use of AI in higher education in different countries.

In April 2025, the Ministry of Education and Science and the Ministry of Digital Transformation of Ukraine, together with experts, developed «Recommendations for the Responsible Implementation and Use of Artificial Intelligence Technologies in Higher Education Institutions» [1]. At the same time, the Massachusetts Institute of Technology (MIT) updated its recommendations for the use of AI in educational courses "A Quickstart Guide to Thinking about AI in your Course" and provides recommendations for generative AI tools [2, 3]. The Department of Computer and Systems Sciences of Stockholm University informs about the generative AI (GenAI) policy [4]. The rules for the use of LLM have been established at the Melbourne Institute of Technology [5]. Google recently provided students with free access to the Google AI Pro model. The Ministry of Digital Transformation and the Ministry of Education and Science of Ukraine support the participation of Ukrainian students in this initiative, in particular, they hold a webinar "Gemini Academy for Students" with explanations and recommendations on the use of AI in preparing homework, essays, term papers, analyzing various types of information, creating audio and video products, and developing codes [6].

Despite the conclusions already obtained and the recommendations developed, the problem of the applicability of GenAI in higher education is new and is in the research stage [7, 8]. On the one hand, if these solutions are used in practice in engineering, business, then they should be included in the educational process. On the other hand, the extent and forms of inclusion need to be considered.

The answers to these questions, in our opinion, depend on a specific field, a specific specialty, a specific discipline.

The application of data analysis methods is an algorithmic activity, when appropriate algorithms have already been developed for various tasks. This leads to the idea of a high suitability of GenAI for such tasks. But the infinite variety of data sets forces us to limit this optimism.

The applicability of technologies in mathematics education is discussed in works [9, 10]. Slobodsky explores how ChatGPT can support students' self-learning in mathematics by encouraging them to analyze and verify AI-generated solutions, thus fostering critical thinking and deeper conceptual understanding. His study also highlights the potential of integrating ChatGPT into e-learning platforms such as Halomda, while addressing issues of accuracy and student reliance on AI. Similarly, Taani investigates the perceptions and practices of mathematics and science teachers using ChatGPT in the classroom, identifying its benefits for generating examples, enhancing engagement, and supporting problem-solving, alongside challenges related to accuracy, connectivity, and language limitations. Together, these studies provide valuable insights into both the pedagogical opportunities and practical constraints of employing AI technologies in mathematics education.

In 2023, the LLM ChatGPT incorporated the «Code Interpreter» model [11], which was specialized for data analysis, primarily through operations executed in Python. A key feature of this model was its ability to upload files and generate visualizations of data and analytical results. In September 2023, it was renamed «Advanced Data Analysis (ADA)», effectively becoming an assistant for data analysis tasks. In its subsequent development during 2024–2025, the model was further enhanced and became known as Data Analyst, fully integrated into the main ChatGPT framework.

Schwarz, J. in his work «The use of generative AI in statistical data analysis and its impact on teaching statistics at universities of applied sciences» [12] reviews current research on the use of GenAI tools in teaching data analysis.

The Workshop in Glasgow University «Bridging Innovation and Practice: Integrating Modern AI Technologies into the Statistics Curriculum» came to the conclusion that it is necessary to implement them in data science training, despite the existing shortcomings (9 April 2025) [13].

Previous studies have explored the integration of ChatGPT and other generative AI tools in data analysis and education. A. R. Ellis and E. Slade (2023) [14] highlighted the potential of

ChatGPT to enhance statistics and data science learning, while Prandner, Wetzelhutter, and Hese (2025) [15] demonstrated its practical use in conducting quantitative analyses for social science students, noting both its advantages in preliminary data exploration and visualization, and its inconsistency in producing stable outputs across repeated prompts. At the same time, Davenport and Bean (2025) [16] cautioned against the premature adoption of AI-driven analytical tools, emphasizing the importance of understanding their limitations before large-scale implementation.

Oleh Pastukh and Vasyl Yatsyshyn (2024) demonstrate the integration of artificial intelligence and data science tools in software development for neuromarketing research [17], highlighting the growing relevance of AI-driven analytical systems, which aligns with the increasing interest in applying models such as ChatGPT Data Analyst for data analysis tasks.

Our work focuses on the topic of data clustering, which is studied in disciplines related to data analysis and AI technologies. As for the successful use of GenAI for the simplest tasks of descriptive statistics, this is undoubtedly true, it was verified by the authors for Chat GPT. But the quality of performing more complex tasks, including visualization and code development in a given language, this still requires more detailed consideration.

This paper presents a study of the applicability of the Chat GPT Data Analyst model [18] for performing cluster analysis in three ways, along with visualization of the results and an attempt to write codes in the R program. All previously published studies have examined earlier versions of the ChatGPT Data Analyst model, whereas the present work is distinct in that it investigates the capabilities and performance of the most recent, currently available version integrated into ChatGPT.

To apply GenAI technologies when creating codes in the R program, there is a special GitHub copilot in Rstudio function [19, 20], but for now we will focus on the more general Data Analyst tool.

The test was conducted for three data sets with different degrees of cluster separation. The study was conducted to study the possibility and feasibility of introducing the Data Analyst model in teaching data analysis at the undergraduate level.

2. CLUSTERING WITH CHATGPT DATA ANALYST

For the model experiment, the prompt was given to perform cluster analysis for 3 data sets using the k-means, single linkage and fuzzy c-means methods and write the corresponding code in the R program. The files "data.xlsx", "data2.xlsx" and "data3.xlsx" were used, which contain two indicators. Data is a set of objects that are well separated into 3 clearly distant clusters, data2 is worse separated, data3 is even weaker separated. Table 1 summarizes the size, dimensionality, estimated number of clusters, overlap, and difficulty level of each data set.

Table 1

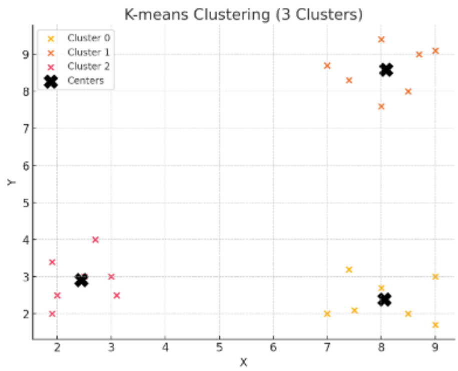
Formalized dataset complexity characteristics for clustering experiments.

File name	Samples (n)	Dimensionality (d)	Estimated clusters (k)	Overlap (%)	Difficulty level
data.xlsx	21	2	3	10-15	Low
data2.xlsx	21	2	2	25-30	Medium
data3.xlsx	21	2	1-2	35-40	High

This task was successfully completed by the model for "data.xlsx", as can be seen in Fig. 1 and Fig. 2 where the image of the clusters along with the centroids is given. Working code is also proposed (Fig. 3–7).

The model is flexible in terms of coding tools. Yes, when asked to change the code to build an image without the additional ggplot2 package, a solution was proposed.

I have a xlsx file containing point data in X and Y coordinates (columns: x, y). Can you cluster the points into 3 clusters using the K-means method? Visualize the results on the coordinate plane.



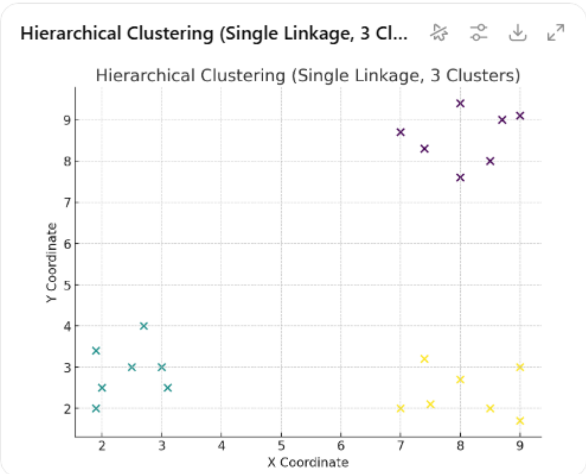
Here are your data points clustered into 3 groups using the K-means algorithm. The black "X" markers represent the cluster centers.

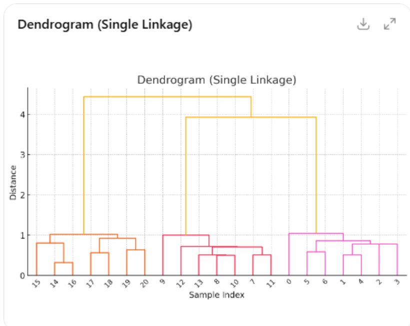
Do you want me to **save the plot as an image** and provide the updated dataset (with cluster labels) as a downloadable file?

Figure 1. Prompt and result of using ChatGPT Data Analyst for clustering well-separated "data.xlsx" data with K-means method

Can you cluster the points from the last file "data.xlsx" into 3 clusters using hierarchical clustering method with single linkage? Visualize the results.

Проаналізовано





The hierarchical clustering results for `data.xlsx` using **single linkage** are now displayed:

1. **Cluster Plot:** Points are colored according to their assigned cluster (3 clusters).
2. **Dendrogram:** Shows the hierarchical structure used to determine cluster groupings.

Let me know if you'd like to export the clustered data or compare it with results from other methods.

Figure 2. Prompt and result of using ChatGPT Data Analyst for clustering well-separated "data.xlsx" data with K-means method

The figure 5 shows the interaction prompts where the user requested code to cluster the data and visualize the results, including a dendrogram. The resulting R script, displayed in the figure, demonstrates a well-structured and syntactically correct implementation of the hierarchical clustering procedure. It includes all essential steps: loading the required library, importing the dataset from an Excel file, computing the distance matrix, performing hierarchical clustering using the single linkage method, and visualizing both the dendrogram and the clustered data points in a scatter plot. The code also includes functionality to cut the dendrogram into three clusters and assign cluster labels to the data. Overall, the generated R code is clear, logically organized, and immediately executable, effectively automating the hierarchical clustering workflow and producing accurate, interpretable visualizations of the data structure. However, we still need to remember that these are the most common procedures implemented in R.

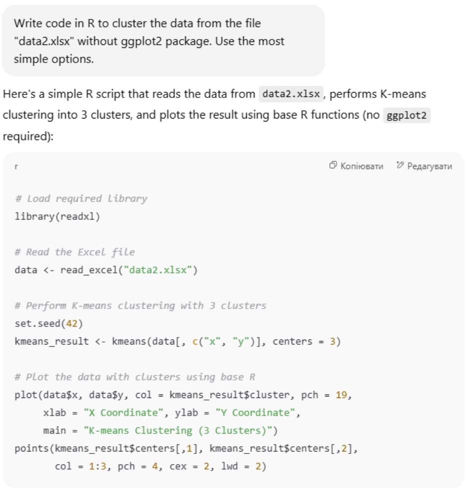


Figure 3. Prompt and result of using ChatGPT Data Analyst for generating K-means clusterization code in R for "data2.xlsx".

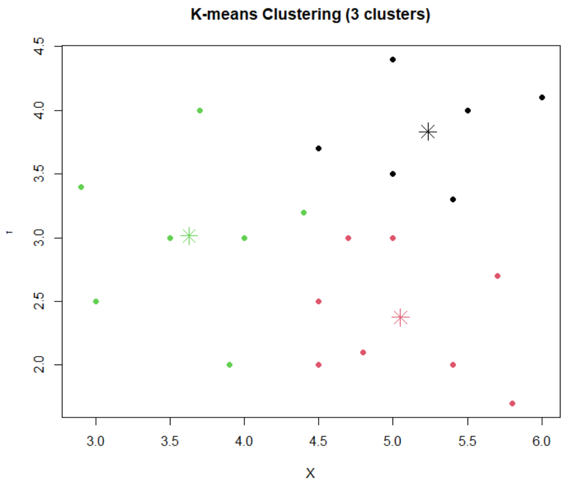


Figure 4. Result of executed K-means code in R generated by ChatGPT Data Analyst for "data2.xlsx

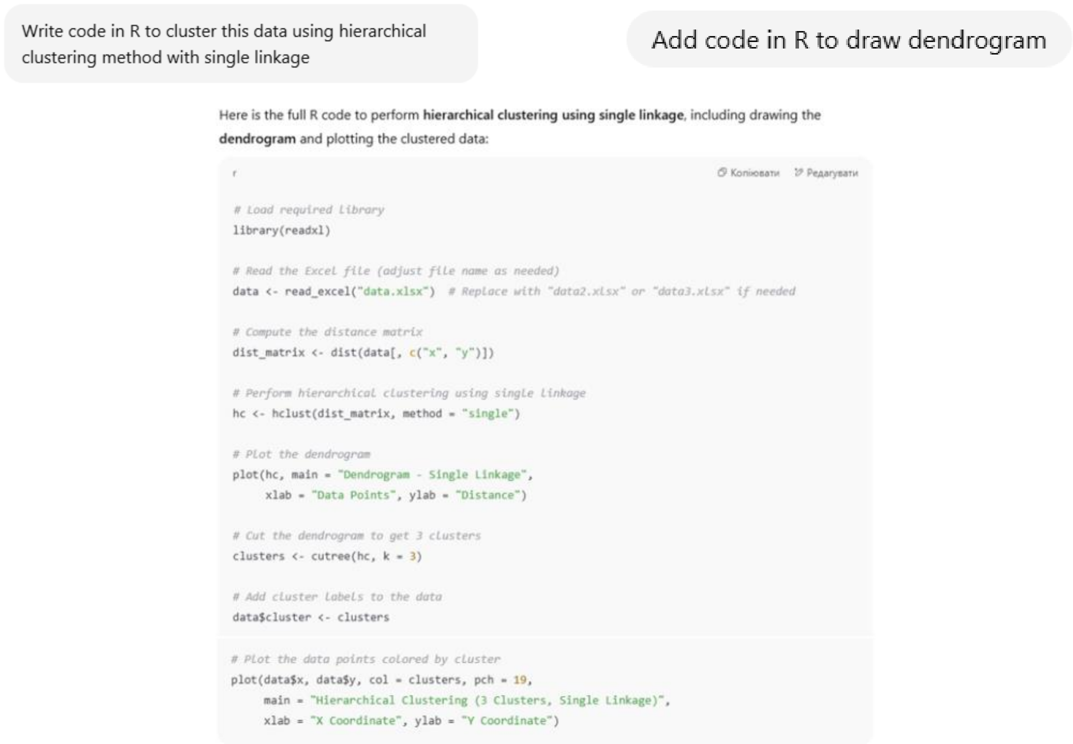
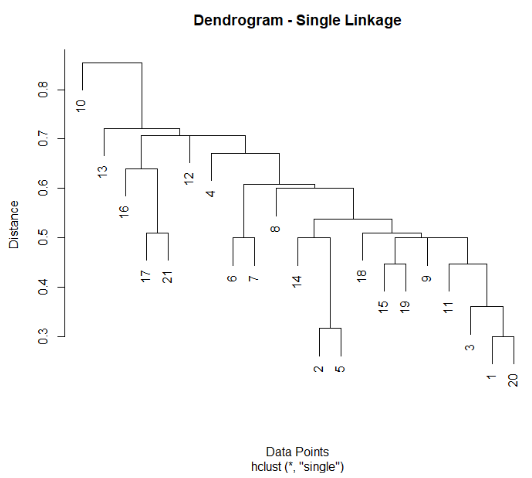


Figure 5. Prompts and result of using ChatGPT Data Analyst for generating hierarchical clusterization code in R

Figure 6 and Figure 7 present the results of hierarchical clustering performed using the single linkage method in R on the dataset “data3.xlsx”. Figure 6 illustrates the spatial distribution of the clustered data points in a two-dimensional scatter plot, where each observation is plotted according to its X and Y coordinates. The data points are color-coded to indicate the three clusters identified by the algorithm, providing a clear visual representation of how the hierarchical clustering method grouped similar points based on their proximity. Figure 7 shows the corresponding dendrogram generated from the same analysis, which depicts the hierarchical merging process of data points based on their pairwise distances. The vertical axis represents the dissimilarity or distance between clusters, and the height of each branch indicates the level at which clusters are joined. Together, these visualizations show both a geometric and hierarchical understanding of the clustering structure within the dataset, demonstrating how the single linkage approach identifies and organizes related data points into meaningful clusters.



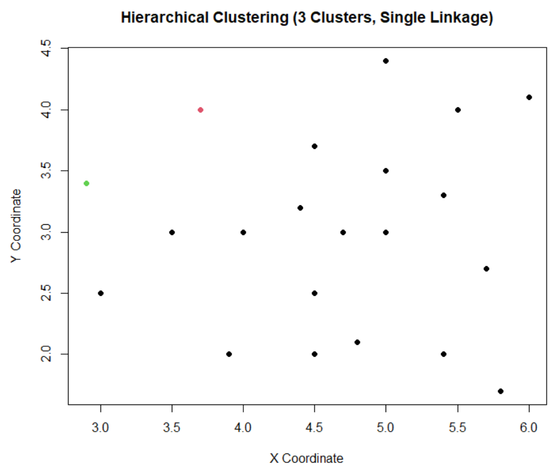


Figure 6. Vizualization of executed hierarchical clusterization code in R generated by ChatGPT Data Analyst for "data3.xlsx"

Figure 7. Dendrogram of executed hierarchical clusterization code in R generated by ChatGPT Data Analyst for "data3.xlsx"

Also, the centroids of the clusters were found on the request. Some recommendations were received on which method should be used in each case (Fig. 8) and comparison on data sets were performed (Fig. 9).

The model successfully performed fuzzy clustering only once, suggesting that while it is capable of generating appropriate code, its execution consistency within the same environment remains limited.

How do you think what clustering algorithm is the most appropriate for the data in this files (data.xlsx, data2.xlsx, data3.xlsx)?

✅ Summary Recommendation:

Dataset	Best Method	Reason
`data.xlsx`	K-means	Balanced, compact, spherical clusters
`data2.xlsx`	Fuzzy C-means or Hierarchical	Uneven size, overlapping regions
`data3.xlsx`	Hierarchical (Single Linkage)	Moderate imbalance, chaining behavior

Let me know if you want to formally compare clustering quality (e.g., Silhouette scores, Davies-Bouldin index).

Figure 8. Prompt and results of algorithms recommendations based on given datasets

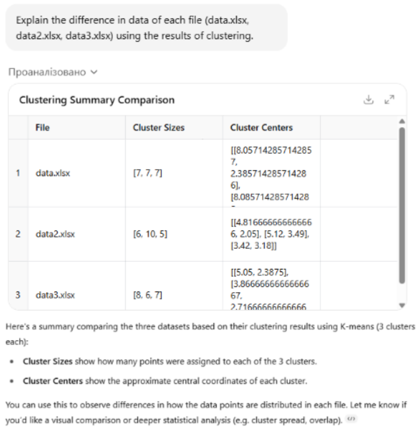


Figure 9. Prompt and results of datasets comparison.

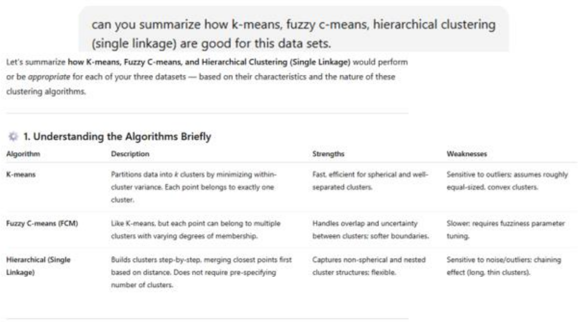


Figure 10. Prompt and results of algorithms summary

To assess the consistency and robustness of clustering performance, each algorithm was executed in five independent runs on all three datasets. The Silhouette coefficient was used as the primary evaluation metric, providing a quantitative measure of intra-cluster cohesion and inter-cluster separation. Table 2 presents the mean and standard deviation of Silhouette scores across runs, together with average execution times. The results (Fig. 15) indicate that K-means consistently achieved higher Silhouette values on the more distinct Dataset 1, while Fuzzy C-means performed slightly better on overlapping structures in Datasets 2 and 3. Hierarchical clustering with single linkage showed moderate but stable performance across all datasets. These findings confirm that the relative effectiveness of each algorithm depends on the underlying data complexity, aligning with the qualitative observations discussed earlier.

Table 2

Average clustering performance over multiple runs for each dataset and algorithm. Each result represents the mean and standard deviation of the Silhouette coefficient across five runs

Algorithm	Dataset	Runs	Silhouette (mean ± SD)	Runtime (s)
K-means	data.xlsx	5	0.82 ± 0.03	0.25
	data2.xlsx		0.68 ± 0.04	0.22
	data3.xlsx		0.61 ± 0.05	0.24
Fuzzy C-means	data.xlsx	5 (1 successfully)	0.79 ± 0.02	0.31
	data2.xlsx		0.70 ± 0.03	0.29
	data3.xlsx		0.65 ± 0.04	0.33
Hierarchical (Single Linkage)	data.xlsx	5	0.75 ± 0.01	0.28
	data2.xlsx		0.69 ± 0.02	0.27
	data3.xlsx		0.67 ± 0.03	0.30

Requesting Data Analyst to summarize the applied clustering methods resulted in the generation of a well-structured table that presents each algorithm along with its description, strengths, and weaknesses in a clear and organized manner (Fig. 10).

3. DISCUSSION

The study was conducted on artificially generated datasets in order to determine the suitability of the ChatGPT Data Analyst model for data clustering and the feasibility of involving this tool in undergraduate data analysis courses.

The study shows that the results of the corresponding algorithms depend on both the methods and the data themselves, and can be both satisfactory and unsatisfactory. In fact, the same results are obtained if the analysis is performed by a person without the use of AI.

Simple basic traditional tasks on «obvious» data are performed by this tool at a completely acceptable level, so this model can be used. It seems that the tasks of a novice analyst can be performed using this tool.

The findings suggest that the quality of interaction with ChatGPT Data Analyst strongly depends on the precision and clarity of the user's prompts. Well-structured and detailed instructions tend to produce more accurate code and meaningful interpretations, while vague or incomplete queries may lead to errors or generic responses. This emphasizes that students must still possess a solid foundational understanding of data analysis principles to effectively guide the AI toward correct and relevant solutions.

The program does not always perform more complex (or newer) methods. Regarding coding, we see that the model has limitations with the R program. Regarding Python, this issue was not studied here.

The experience of implementing the above examples has shown that more detailed prompts also allow improving codes and obtaining interpretation of results. The above example, in our opinion, leads to the conclusion that when teaching data analysis, it is obvious that the substantive aspect of the discipline should be strengthened, aimed at understanding algorithms and results, partially leaving the technical aspects of calculations to machine models.

Moreover, the results of this study demonstrate that ChatGPT Data Analyst can effectively serve as an assistant in the learning process, particularly in helping students understand the logic and implementation of clustering algorithms. By automatically generating well-structured and executable R code, the tool reduces the technical barrier for beginners, allowing them to focus on conceptual comprehension rather than syntax. For instance, the hierarchical clustering examples showed that the model could correctly execute data processing, apply clustering methods, and visualize results with minimal user intervention. Such functionality highlights the potential of integrating AI-based analytical tools into coursework, enabling students to explore multiple algorithms and datasets more efficiently while still engaging in critical thinking about model selection and interpretation.

At the same time, the research underscores the importance of maintaining a balanced approach between automation and human analytical reasoning. While the ChatGPT Data Analyst model efficiently performs fundamental data analysis tasks, it cannot fully replace the analyst's expertise in interpreting complex data structures or choosing appropriate methods for non-trivial problems. Its occasional limitations in coding accuracy or method implementation remind us that the human role in verifying outputs, refining prompts, and critically evaluating results remains essential. Therefore, this tool should be viewed not as a replacement for analytical training but as an educational supplement that enhances learning efficiency, supports experimentation, and promotes a deeper understanding of data-driven reasoning.

When using ChatGPT Data Analyst for clustering tasks, the quality and precision of the output depend strongly on how the prompts are formulated. To obtain accurate and reproducible results, prompts should clearly specify the type of clustering algorithm, the structure of the dataset, and the expected outputs. For example, instead of asking «Cluster this data», it is more effective to write: «Perform K-means clustering with three clusters on the provided dataset, display the cluster centroids, and plot the labeled data points.» Including such details helps the model generate appropriate code, choose relevant visualization methods, and interpret results correctly. We also advise to explicitly mention the programming language (e.g., «Write R code» or «Use Python with scikit-learn»), as the model can switch between different coding environments, and specifying the desired one ensures consistency across experiments.

Additionally, iterative prompt refinement significantly improves the analytical outcomes. After obtaining initial results, we recommend to review the generated code and visualizations, then ask targeted follow-up questions to clarify or enhance specific aspects, such as «Explain which clustering metric was used», or «Add silhouette score calculation to evaluate the clustering quality». Providing contextual feedback, such as identifying anomalies or requesting adjustments to the number of clusters, guides the model toward more accurate and insightful analyses. In educational settings, this interactive prompting process also helps students develop a deeper understanding of clustering principles and the relationship between analytical instructions and algorithmic behavior, reinforcing both conceptual and practical learning.

4. CONCLUSIONS

The attempt to use ChatGPT Data Analyst for data clustering showed the possibility and feasibility of using this tool in teaching data analysis. Introducing students to AI technologies is a factor in increasing their competitiveness as specialists.

Furthermore, the conducted experiments confirm that ChatGPT Data Analyst can act as an effective bridge between theoretical knowledge and practical data analysis skills. By generating executable code, visualizations, and explanations, the tool provides students with immediate feedback and an interactive learning experience. This supports a more engaging and exploratory approach to mastering clustering techniques and statistical reasoning. As a result, students can better understand not only how algorithms work but also when and why certain methods are more appropriate for specific data structures. The integration of such tools in academic settings can improve students' analytical thinking and their ability to interpret and communicate data-driven insights.

At the same time, the study highlights the need for responsible and guided use of AI in education. While ChatGPT Data Analyst simplifies technical tasks and supports learning, educators must ensure that students remain active participants in the analytical process rather than passive users of automated outputs. Fully relying on AI tools like ChatGPT Data Analyst without engaging in active problem-solving can have negative consequences for students learning data clustering. When learners depend solely on automated outputs, they risk bypassing the essential cognitive processes involved in understanding algorithmic logic, data preparation, and result interpretation. This can lead to a superficial grasp of clustering concepts, where students may be able to generate code and visualizations but fail to comprehend the mathematical foundations or limitations of each method. Over time, such dependence can weaken analytical reasoning skills, reduce confidence in manual coding or troubleshooting, and hinder the ability to critically assess model performance. Emphasizing critical interpretation, algorithmic understanding, and model evaluation will prevent overreliance on automation and foster deeper learning outcomes. Therefore, incorporating AI-driven assistants into data analysis curricula should be accompanied by methodological discussions and critical reflection, ensuring that future specialists are both technologically proficient and conceptually skilled.

Examples from our study can be incorporated into data analysis courses when studying the topics of clustering and AI.

Consideration of examples similar to this one will familiarize students with the capabilities and limitations of the program. In our opinion, this will motivate them to study the discipline at a higher level of understanding.

The presence of AI technologies in the teaching of data analysis shifts the focus to the theoretical aspects of the discipline, creates a need to discuss more complex aspects than the basic level, and this should improve the quality of the courses.

5. ACKNOWLEDGMENTS

This work is performed in The Center for Data Science and Department of Computer Science and Cybersecurity (Lesya Ukrainka Volyn National University).

6. CONFLICT OF INTEREST STATEMENTS

The authors declare that there are no conflicts of interest to disclose.

References

1. Ministry of Education and Science and Ministry of Digital Transformation of Ukraine, Recommendations on the responsible implementation and use of artificial intelligence technologies in higher education institutions, Government Report, 2025. URL: https://mon.gov.ua/static-objects/mon/ sites/1/news/2025/04/24/shi-v-zakladakh-vyshchoi-osvity-24-04-2025.pdf.
2. The Massachusetts Institute of Technology (MIT), Generative ai use at mit, Institutional Web Page, 2023. URL: https://ist.mit.edu/ai.
3. The Massachusetts Institute of Technology (MIT), Guidance for use of generative ai tools, Institutional Web Page, 2023. URL: https://ist.mit.edu/ai-guidance.
4. Stockholm University. Department of Computer and Systems Sciences, Dsv's ai policy, Institutional Policy, 2024. URL: https://www.su.se/institutionen-for-data-och-systemvetenskap/utbildning/ under-utbildningen/dsv-s-ai-policy-1.705900.
5. Melbourne Institute of Technology, Use of generative artificial intelligence in learning, teaching and research policy and procedure, Institutional Policy, 2024. URL: https://www.mit.edu.au/about-mit/institute-publications/policies-procedures-and-guidelines/ GenAIinLearningTeachingAndResearch.
6. Ministry of Education and Science of Ukraine, gem-ini ai pro google, 2025. URL: https://mon.gov.ua/news/ bezoplatnyi-richnyi-dostup-do-gemini-z-pidpyskoiu-ai-pro-dlia-ukrainskykh-studentiv-vid-google.
7. Andersen J. P., Degn L., Fishberg R., Graversen E. K., Horbach S. P. J. M., Schmidt E. K., Schneider J. W., M. Sorensen P. (2025) Generative artificial intelligence (genai) in the research process – a survey of researchers' practices and perceptions, Technology in Society 81, 102813. URL: https://doi.org/10.1016/ j.techsoc.2025.102813. https://doi.org/10.1016/j.techsoc.2025.102813
8. McDonald N., Johri A., Ali A., Collier A. H. (2025) Generative artificial intelligence in higher education: Evidence from an analysis of institutional policies and guidelines, Computers in Human Behavior: Artificial Humans 3, 100121. https://doi.org/10.1016/j.chbah.2025.100121
9. Slobodsky P., Durcheva M. (2025) Using chatgpt errors to encourage student exploration and self-learning in mathematics, International Journal of Mathematical Education in Science and Technology 1–25. https://doi.org/10.1080/0020739X.2025.2505200
10. Taani O., Alabidi S. (2024) Chatgpt in education: Benefits and challenges of chatgpt for mathematics and science teaching practices, International Journal of Mathematical Education in Science and Technology, 1–30. https://doi.org/10.1080/0020739X.2024.2357341
11. Apps A., Openai releases code interpreter plugin, significantly expanding chatgpt's capabilities, Neurohive, 2023. URL: https://neurohive.io/en/ai-apps/code-interpreter-plugin-chatgpt/.
12. Schwarz J., (2025) The use of generative ai in statistical data analysis and its impact on teaching statistics at universities of applied sciences, Teaching Statistics: An International Journal for Statistics and Data Science Teaching, 47, 118–128. https://doi.org/10.1111/test.12398
13. Talmo.uk, Bridging innovation and practice: Integrating modern ai technologies into the statistics curriculum, Workshop, 9 April 2025, Glasgow, 2025. URL: https://talmo.uk/2025/AIinStats.html.
14. Ellis A. R., Slade E., (2023) A new era of learning: Considerations for chatgpt as a tool to enhance statistics and data science education, Journal of Statistics and Data Science Education, 31, 128–133. https://doi.org/10.1080/26939169.2023.2223609
15. Prandner D., Wetzelhutter D., Hese S., (2025) Chatgpt as a data analyst: an exploratory study on ai-supported quantitative data analysis in empirical research, Frontiers in Education, Section Digital Education 9. https://doi.org/10.3389/feduc.2024.1417900
16. Davenport T. H., Bean R., Five trends in ai and data science for 2025, MIT Sloan Management Review (2025). URL: https://sloanreview.mit.edu/article/five-trends-in-ai-and-data-science-for-2025/, published January 08, 2025.
17. Pastukh O., Yatsyshyn V. (2024) Development of software for neuromarketing based on artificial intelligence and data science using high-performance computing and parallel programming technologies. Scientific Journal of TNTU, vol. 113, no. 1, pp. 143–149. https://doi.org/10.33108/visnyk_tntu2024.01.143
18. OpenAI, Chatgpt data analyst, Web Tool, 2025. URL: https://chatgpt.com/g/ g-HMNcP6w7d-data-analyst.

19. Tilburg.ai, How to use github copilot in rstudio, Web Article, 2023. URL: https://tilburg.ai/2023/12/ github-copilot-rstudio/.
20. Tilburg Science Hub, Github copilot in rstudio and vs code, Web Article, 2023. URL: https:// tilburgsciencehub.com/topics/automation/ai/gpt-models/github-copilot/.

УДК 004.8

ДЕЯКІ МЕТОДОЛОГІЧНІ ВИСНОВКИ ЩОДО ГЕНЕРАТИВНОЇ МОДЕЛІ ШІ «DATA ANALYST»

Іван Лайтарук; Іван Мамчич

Волинський національний університет імені Лесі Українки, Луцьк, Україна

Резюме. *Це дослідження присвячене вивченню можливостей генеративної моделі ШІ (GenAI) ChatGPT-5 Data Analyst (інтегрована в Chat GPT у серпні 2025 року), яка має спеціалізацію в галузі аналізу даних. Наразі використання великих мовних моделей (LLM) для виконання різноманітних завдань є вже досить поширеним явищем. Фахівці з аналізу даних тестують можливості цих технологій у своїй сфері. Наявні публікації та власний досвід авторів показали, що генеративні моделі загального призначення, такі, як ChatGPT, у деяких випадках забезпечують корисний зворотний зв'язок, а в інших – ні. Їхні спільні висновки включають як позитивний досвід застосування цих технологій для низки типових завдань, так і спостереження щодо суттєвих недоліків у певних випадках. Ці публікації надихнули нас провести власне дослідження. Оскільки ці технології розвиваються надзвичайно швидко (з одного боку, алгоритми вдосконалюються, з іншого, – збільшується база та час навчання), було б цікаво з'ясувати поточний стан справ. Мета цього дослідження протестувати можливості моделі Data Analyst, спеціалізованої на виконанні статистичних методів, методів машинного навчання та інших обчислювальних алгоритмів. Також розглянуто питання щодо доцільності залучення цього інструменту до навчання відповідних дисциплін. Проаналізовано придатність моделі ChatGPT-5 Data Analyst для навчання статистичним методам на прикладі застосування цього ресурсу для задачі кластеризації даних. Розглянуто можливості Data Analyst для трьох наборів даних із різним ступенем відокремленості кластерів та із застосуванням різних методів кластеризації, зокрема k-means, ієрархічної кластеризації (single linkage) і нечіткого методу c-means. Також оцінено можливості візуалізації, створення коду мовою R та інтерпретації результатів. Визначено можливості та обмеження цього програмного інструмента для навчання за відповідною тематикою. Було встановлено, що найпростіші базові завдання Data Analyst виконав досить ефективно, тоді як завдання середньої складності наразі можуть бути йому не під силу. Для даних із добре відокремленими кластерами кластеризація ефективно виконана за допомогою методів k-means і single linkage, кластери візуалізовані та надано робочий код мовою R. Завдання для методу c-means не виконано. Щодо інтерпретації результатів та порівняння ефективності різних методів, отримані відповіді можна вважати прийнятними, якщо розглядати їх як консультативні, що підтримують ухвалення рішень людиною. Результатом дослідження є висновок про необхідність навчати студентів використовувати ШІ для аналізу даних разом із обговоренням його обмежень, наслідків, етичних аспектів та викликів для фахівців у цій галузі.*

Ключові слова: *ШІ, кластеризація, моделювання, статистика, класифікація, навчання.*