Smart Learning Environments

RESEARCH

Open Access

Empowering students through active learning in educational big data analytics

Yun-Cheng Tsai[1*]

*Correspondence:
pecu@ntnu.edu.tw

[1] Department of Technology Application and Human Resource Development, National Taiwan Normal University, No. 162, He-Ping East Road Sec1, Taipei 10610, Taiwan

Abstract

Purpose: This paper explores how Educational Big Data Analytics can enhance student learning. It investigates the role of active learning in improving students' data analysis skills and critical thinking. By actively engaging students in data analysis assessments, the aim is to equip them with the skills to navigate the data-rich educational landscape.

Methods: The study uses a teaching strategy that combines structured and unstructured data analysis using Python tools and ChatGPT APIs. It presents five assignments, each highlighting data analysis skills and encouraging critical thinking.

Results: The paper offers insights into how the teaching strategy effectively enhances students' data analysis and critical thinking skills. It investigates the specific impact of active learning on students' engagement with educational data. The study reveals that all students can complete a comprehensive project, integrating the skills they have learned in the five assignments related to educational big data while incorporating the educational implications from their respective disciplines.

Conclusion: The key lies in instructors being able to design individual assignments that link practical experiences, enabling each teaching session's effectiveness to accumulate in students' personal experiences and practical skills, ultimately empowering them with the abilities necessary to work effectively with Educational Big Data Analytics. The findings of this study make a valuable contribution to the ongoing conversation about enhancing the educational experience for students in this data-rich era.

Keywords: Active learning, Inquiry-based learning, ChatGPT Python APIs, Exploratory Data Analysis (EDA), Text mining, Latent Dirichlet Allocation (LDA), Co-occurrence network, Minimum Viable Product (MVP), Principal Component Analysis (PCA)

Introduction

In today's rapidly evolving landscape of artificial intelligence (AI) and Big Data Analytics, education systems face the challenge of adapting to these technological advancements (Shi, 2022). This paper explores the integration of Educational Big Data Analytics in teaching environments, highlighting its potential to enhance student learning experiences and outcomes.

The core of our investigation revolves around active learning, a pedagogical approach gaining traction for its effectiveness in teaching complex subjects like data analysis and

critical thinking (Prince, 2004; Tsai, 2023). This study delves into how active learning can be utilized within Big Data Analytics in education, aiming to make learning experiences more engaging and relevant to students from diverse academic backgrounds, including those with limited IT knowledge (Donoghue et al., 2021; Romanow et al., 2020).

We recognize the varied expertise levels of students, especially those from non-IT fields, and address this by designing our course to be inclusive and accessible. Our teaching strategy, which includes a series of five well-structured assignments, is tailored to cater to a broad spectrum of students. These assignments are carefully crafted to balance the required expertise and ability, ensuring that students with different skill levels and disciplinary backgrounds can effectively engage and benefit from the course (Brewer & Cunningham, 2023).

In response to concerns about maintaining objectivity in grading and assessment, we have implemented a standardized rubric that addresses students' diverse thought processes and problem-solving approaches from different disciplines. This rubric evaluates students' understanding and application of data analytics concepts while considering their unique perspectives and backgrounds.

Moreover, we propose adjustments and enhancements to our course standards, including the possibility of extending the duration of the final project. This extension would provide students, especially those with less programming experience, ample time to fully engage with and understand the complexities of Big Data Analytics.

We also acknowledge the importance of evaluating the long-term impact of our teaching strategies. To this end, we plan to conduct a follow-up study that assesses the sustained effect of our approach on students' data analysis skills and critical thinking capabilities over time.

Lastly, this paper provides detailed insights into our study's Python tools and ChatGPT APIs. We elaborate on how these tools were integrated into our teaching strategy, enhancing learning and offering students hands-on experience in practical data analytics applications.

Through this study, we aim to contribute to the ongoing conversation about enhancing educational experiences in a data-rich era, focusing on student empowerment through active learning in Educational Big Data Analytics.

Key elements of active learning

In Educational Big Data Analytics, active learning includes several key components that enhance students' data analysis and critical thinking abilities. These crucial elements consist of:

1. Exploratory Data Analysis (EDA): Encouraging students to explore and analyze real educational datasets to gain hands-on experience (Milo & Somech, 2020).
2. Problem-Solving: Challenging students to identify and solve data-related problems, fostering critical thinking skills (Aslan, 2021).
3. Data Visualization: Teaching students how to present data through visualizations, enhancing their communication skills effectively (Waskom, 2021).
4. Group Activities: Promoting collaborative learning through group projects, allowing students to learn from their peers (Qureshi et al., 2023).

5. Inquiry-Based Learning: Encouraging students to actively ask questions and seek answers, promoting curiosity and self-directed learning (Husni, 2020).

These components form a comprehensive approach to active learning. They empower students to gain essential skills and foster a growth-oriented mindset, enabling them to navigate the data-rich educational environment more effectively and tackle the evolving challenges associated with Educational Big Data Analytics.

The impact on student empowerment

Active learning in Educational Big Data Analytics empowers students by:

1. It cultivates a profound comprehension of data analysis principles, ensuring students develop a solid foundation in this critical field.
2. By promoting critical thinking, problem-solving, and analytical skills, active learning equips students with the cognitive tools necessary to tackle complex data-related challenges effectively.
3. Active learning enhances students' proficiency in navigating and interpreting educational data, allowing them to derive meaningful insights from vast datasets.
4. Practical skills acquired through active learning become invaluable in real-world data analysis scenarios, preparing students for professional applications.
5. Importantly, it fosters a lifelong learning mindset, enabling students to stay adaptable and responsive to the continuous evolution of data technologies and educational data analytics. This holistic approach ensures that students are well-equipped to thrive in the dynamic landscape of Educational Big Data Analytics.

Active learning in the context of Educational Big Data Analytics is a powerful tool for empowering students. By actively engaging with data analysis, students acquire valuable skills and develop the confidence to tackle the challenges of the data-driven education landscape. Encouraging students to be active learners in this field is essential for their success and adaptability in an ever-changing educational environment.

Methodology

Our research leveraged the capabilities of specific Python libraries and the ChatGPT API from OpenAI, which played a crucial role in augmenting students' learning experience. These technologies were strategically selected to align with our educational objectives, supporting our pedagogical approach and empowering students to grasp and apply their acquired knowledge more intuitively. The Python libraries we employed, including Pandas for data manipulation, NumPy for numerical computations, and Matplotlib for data visualization, are celebrated for their powerful data processing abilities and versatility in visualizing complex datasets. These tools enabled students to delve into data analysis projects, offering them a platform to apply fundamental concepts and techniques in data science. Through practical exercises, learners could manipulate datasets, perform statistical analysis, and visualize results, thus cementing their understanding of data analysis principles.

Furthermore, our integration of the ChatGPT API played a pivotal role in enhancing the interactivity of data exploration and analysis. Using this API, students could perform data queries and analyses through natural language commands, making learning more engaging and aligned with real-world data science workflows. This integration allowed for a seamless blend of coding and natural language processing, enabling students to interact with data more intuitively and efficiently.

In terms of practical application, our methodology incorporated assignments that required students to use these Python tools in conjunction with the ChatGPT API to solve real-world data analysis problems. For instance, students were tasked with collecting data from various sources, cleaning and processing it using Pandas and NumPy, and then visualizing their findings with Matplotlib. Following this, they used the ChatGPT API to generate natural language summaries of their data analysis, effectively bridging the gap between technical data analysis and communicative reporting.

During classroom sessions, we embraced a live coding strategy, which allowed students to witness the immediate impact of their code, fostering an interactive and dynamic learning environment. Complemented by real-time interactions with the ChatGPT API, this approach greatly enhanced student engagement and facilitated a more collaborative and responsive learning atmosphere.

Integrating these Python tools and the ChatGPT API into our teaching strategy provided a comprehensive platform for students to engage in hands-on data analysis projects. This enhanced the effectiveness of our instructional methods and equipped students with the practical skills and knowledge necessary for tackling complex data-driven challenges in their future careers.

Participants

These students learned a lot about analyzing data. They just had an introductory class in programming in their first year. This helped them understand some parts of the code.

Looking at the information in Table 1, we can describe what the students are like in this way:

1. *Gender* Male participants accounted for 15 individuals, representing 36.59%. Female participants comprised the majority, with 26 individuals making up 63.41%.
2. *Education level* One Freshman participant is in the study, constituting 2.44% of the sample. Sophomore students numbered eight individuals, making up 19.51%. Junior students accounted for 14 individuals, representing 34.15%. Senior students comprised 16 individuals, constituting 39.02%. Two participants were pursuing master's degrees, representing 4.88%.
3. *Fields of study* Participants from the College of Education numbered 5, comprising 12.19%. The College of Liberal Arts had 2 participants, making up 4.88%. The College of Management was represented by 3 participants, accounting for 7.32%. The College of Science had 12 participants, representing 29.27%. The College of Technology and Engineering had the most significant representation, with 19 participants making up 46.34% of the sample (see Fig. 1).

Table 1 Participant demographic information

Participant demographics	Numbers	Percentage
Gender		
Male	15	36.59
Female	26	63.41
Total	41	100.00
Education level		
Freshman	1	2.44
Sophomore	8	19.51
Junior	14	34.15
Senior	16	39.02
Master	2	4.88
Total	41	100.00
Fields of study		
College of Education	5	12.19
College of Liberal Arts	2	4.88
College of Management	3	7.32
College of Science	12	29.27
College of Technology and Engineering	19	46.34
Total	41	100.00

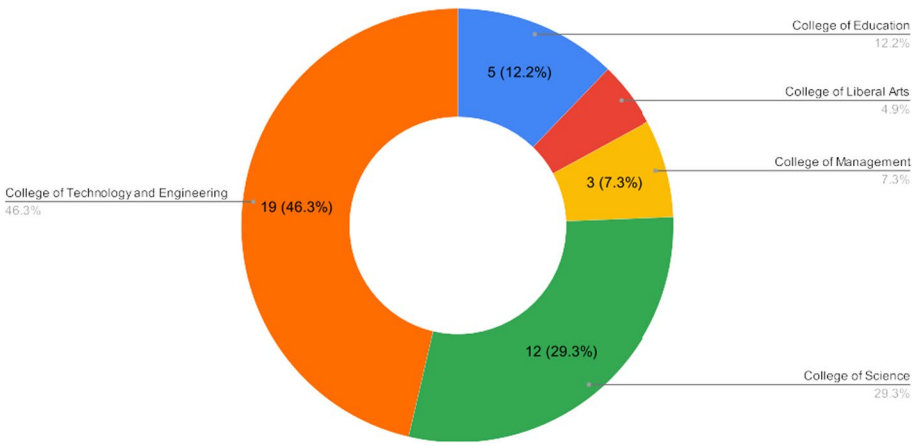

Fig. 1 Percentage of fields of study

This participant profile provides a breakdown of gender distribution, education levels, and fields of study, offering insights into the composition of the study's participants.

Teaching approach

Our course's essence, "Learning Analytics Tools Implementation Applications", revolves around an active learning strategy, moving away from the traditional lecture-focused teaching approach. In this 16-week course, we have designed five assignments, each spanning two weeks. During the first ten weeks, students acquire analytical skills through these assessments. In the eleventh week, they form teams for collaborative projects, receiving practical guidance and feedback on the feasibility of their proposals. Up to the fifteenth, the subsequent weeks are dedicated to integrating

these projects, culminating in their Minimum Viable Product (MVP) presentation in the sixteenth week. To address the varying levels of expertise and backgrounds of our students, especially those from non-IT fields, we have incorporated several critical adjustments in our teaching method:

1. *Modular learning design* Each assignment is structured modularly, with complex components. This allows students to engage with content at a level appropriate to their current understanding and skill set, ensuring inclusivity and accessibility.
2. *Additional learning resources* We provide supplementary materials such as tutorials, step-by-step guides, and practical examples of Python programming and data analysis basics. These resources benefit lower-grade students or those new to the field of information technology.
3. *Differential assessment difficulty* Assignments are designed with varying difficulty levels, offering entry-level assessments for beginners and more challenging assessments for advanced learners. For the high-difficulty evaluations, the course has developed a series of assignments requiring students to actively utilize Python APIs available on open-source platforms, such as the OpenAI API. These advanced assignments demand that students be capable of autonomously collecting big data related to custom topics through web scraping techniques. They must also choose suitable AI analysis modules based on these topics and data. The advanced assessments include the following steps:

    a. *Data collection* Students must learn how to use web scraping technologies automatically to collect data from various sources. This includes understanding how to avoid legal and ethical issues and efficiently process and store the collected data.
    b. *API application* Students will be guided to use open-source APIs, like the OpenAI API, for data processing and analysis. This requires students to be familiar with using these APIs and understand how to integrate these tools into their data analysis workflows.
    c. *Selection and application of AI analysis modules* Students must select appropriate AI analysis modules based on the collected data and specific topics. This may involve the application of machine learning, natural language processing, or other related technologies, requiring students to have an essential ability to train and evaluate models.
    d. *Data analysis process and practical code* Finally, students must propose a complete data analysis process, including every step from data collection, processing, and analysis to the result presentation. They must also provide executable code explaining how to implement this process and interpret the analysis results.

4. *Collaborative learning opportunities* Group assignments and projects are integral to the course, encouraging peer-to-peer learning and mentorship. This fosters a collaborative learning environment and allows students from diverse backgrounds to bring unique perspectives and knowledge.

5. *Continuous feedback mechanism* We have established a system of regular feedback sessions where students can receive personalized instructor guidance. This approach allows for real-time adjustments in learning strategies and provides students with the support they need to successfully engage with the course material. The analytical capabilities of our instructors significantly enhance the effectiveness of this mechanism. They are proficient in analyzing the code related to each student's implementation of the analysis process. This proficiency enables instructors to perform live debugging sessions and incorporate additional, possibly undiscovered, packages on the fly. Such capabilities allow for deeper, more insightful discussions and hands-on problem-solving during these feedback sessions. This real-time, data-driven feedback system is not just about correcting errors; it is about enriching the learning experience. Instructors can identify patterns in student challenges and address them immediately, often introducing new tools and techniques that students might not have encountered otherwise. This dynamic interaction ensures that learning is not a one-way transmission of information but a collaborative process of discovery and improvement. By leveraging the instructors' ability to analyze and interact with students' code live, we address individual learning needs and foster a culture of continuous improvement and innovation. This approach ensures that students learn to solve present problems and are equipped to tackle new, unforeseen challenges with confidence and creativity.
6. *Use of intuitive analytical tools* The course integrates user-friendly and intuitive analytical tools, making complex data analysis more approachable for students with limited programming experience. These tools are chosen for their ease of use and ability to represent data analysis concepts visually.

By implementing these strategies, we aim to create a more inclusive learning environment that accommodates students from various academic backgrounds and expertise levels and ensures that each student can effectively grasp and apply the concepts of Educational Big Data Analytics. Through this approach, we strive to empower all students with the necessary skills and confidence to navigate the ever-evolving data analytics landscape. We propose a comprehensive teaching method that includes five different assignments, each focusing on various data analysis skills:

1. *Individual assignment 1* Exploratory Data Analysis (EDA) of Structured Data Sets—Students will learn to use Python Pandas for fundamental data analysis.
2. *Individual assignment 2* Exploratory Data Analysis (EDA) of Personal Data Sets—Students will apply data analysis to educational data related to their field.
3. *Individual assignment 3* Text Mining, Theme Analysis of Unstructured Data—Students will learn how to use ChatGPT API and Latent Dirichlet Allocation (LDA) for analyzing text themes.
4. *Individual assignment 4* Data Visualization and Interpretation using Latent Dirichlet Allocation (LDA)—Students will combine data visualization tools with structured and unstructured data for interpretive analysis.

5. *Individual assignment 5* Co-occurrence Network Analysis for Association and Reasoning—Students will apply data analysis skills to analyze associations and make reasoned conclusions using co-occurrence networks.

Here are the completion requirements for each assessment:

1. Individual assignment 1—Exploratory Data Analysis (EDA) of Structured Data Sets:

   a. Successfully import the dataset and perform fundamental data analysis using Python Pandas.
   b. Generate at least two data visualization charts to showcase essential data features.
   c. Provide a brief analysis report discussing data trends and critical insights.

2. Individual assignment 2—Exploratory Data Analysis (EDA) of Personal Data Sets:

   a. Apply data analysis skills to an education dataset relevant to your field.
   b. Use appropriate data visualization tools to present patterns and associations in the data.
   c. Provide a brief analysis report emphasizing the potential implications of the data on your field.

3. Individual assignment 3—Text Mining, Theme Analysis of Unstructured Data:

   a. Successfully analyze the themes of textual data using ChatGPT API and Latent Dirichlet Allocation (LDA).
   b. Present visual results of text themes and key associated keywords.
   c. Provide an explanatory report detailing the identified textual themes and related insights.

4. Individual assignment 4—Data Visualization and Interpretation using Latent Dirichlet Allocation (LDA):

   a. Successfully combine structured and unstructured data using LDA for data interpretation.
   b. Create at least three data visualization charts to explain patterns and trends in the data.
   c. Provide a detailed analysis report highlighting the role of LDA in data interpretation.

5. Individual assignment 5—Co-occurrence Network Analysis for Association and Reasoning:

   a. Successfully analyze data associations and make reasoned inferences using co-occurrence networks.
   b. Provide clear charts or visual results highlighting relationships in the data.

c. Deliver a structured report explaining the relationships between data elements and the associated inferences.

The effectiveness scores for each assignment are based on three indicators. Students can earn a maximum score of 3 if they meet all three hands for a given assessment. If they meet two of the three indicators, they will receive a score of 2. Meeting just one of the indicators earns them a score of 1 while failing to meet any of the indicators results in a score of 0 for that assignment.

Based on the provided criteria, here is an explanation of how to assign scores of 3, 2, and 1 for the Final Demo MVP:

1. Completeness of core functionality:

   a. If the MVP includes all core functionalities and they are fully implemented and operational, a score of 3 is awarded.
   b. If the MVP includes most core functionalities but some are partially implemented, a score of 2 is given.
   c. If the MVP lacks several core functionalities or is non-operational, a score of 1 is assigned.

2. Problem-solving ability:

   a. A score of 3 is granted if the student demonstrates how the MVP effectively addresses the core problem, providing a detailed solution.
   b. If the student provides some problem-solving approach but needs more clarity or detail in the solution, a score of 2 is appropriate.
   c. If the student fails to adequately explain the problem and solution, a score of 1 is given.

3. User feedback and improvement potential:

   a. A score of 3 is warranted when the student effectively presents how they considered user feedback and outlines a clear improvement plan.
   b. If the student provides some level of user feedback and an improvement plan, though lacking specificity, a score of 2 is suitable.
   c. If the student needs to adequately address user feedback and provide an improvement plan, a score of 1 is assigned.

This scoring approach allows for a quantitative evaluation of each MVP, differentiating between those that fully meet the requirements (scored 3), partially meet the requirements (scored 2), and those that do not meet the criteria (scored 1).

This scoring system ensures that students are evaluated on their ability to achieve specific learning outcomes and skills in each assignment, providing a clear and measurable assessment of their performance. It also emphasizes the importance of addressing all three indicators to attain the highest possible score, encouraging comprehensive and proficient completion of the evaluations.

These criteria will ensure that students achieve specific learning objectives and skills in each assessment, providing clear measurement standards for completion.

Teaching effectiveness assessment

In our teaching approach, we uphold a uniform standard for assessment across all student groups, irrespective of their academic backgrounds. This is feasible due to our robust learner-centered support system, which includes:

1. *Comprehensive rubrics* We utilize detailed rubrics that clearly outline the criteria for each assignment. These rubrics are designed to be discipline-agnostic, focusing on the core skills and competencies required in data analytics, thereby maintaining a consistent evaluation standard for all students.
2. *Adaptive learning resources* Our course has various learning materials catering to multiple skill levels and learning styles. This ensures that all students can access the resources necessary to meet the established standards regardless of their initial proficiency or background.
3. *Regular feedback mechanisms* Our teaching methodology incorporates continuous feedback loops. This allows us to identify and address any learning gaps among students from different backgrounds, ensuring they all have an equal opportunity to understand and meet the assessment criteria.
4. *Peer learning and collaboration* By promoting a collaborative learning environment, students from different disciplines can share their perspectives and knowledge, mitigating disciplinary biases in understanding the course material.
5. *Transparency in grading* We maintain transparency in our grading process, where students are informed about the assessment criteria and grading rubrics at the outset. This clarity helps in setting uniform expectations and reduces ambiguities in grading.

These measures ensure that our grading system is fair, unbiased, and consistent, providing all students an equal opportunity to demonstrate their skills and knowledge regardless of their background. This approach aligns with our commitment to fostering an inclusive and equitable learning environment. The teaching strategy emphasizes assessing students' real-world project outcomes rather than traditional pen-and-paper exams or legionnaires. The key to evaluating the effectiveness of this approach lies in several assessment methods. These methods involve closely observing the process of students' project work, collecting their feedback, and ensuring they produce tangible, integrated projects that address authentic problems. Here are the assessment methods that can be used:

1. *Project deliverables evaluation* Assess the quality and completeness of the final projects produced by students. This evaluation should focus on whether they effectively integrate data analysis techniques, programming code, and problem-solving skills to address real-world issues.
2. *Process observation* Observe and document the students' workflow throughout the project development. This includes tracking their problem-solving approaches, code

development, data analysis procedures, and collaborative teamwork during the project.

3. *Peer and self-assessment* Encourage students to assess their work and the work of their peers. Self-reflection and peer feedback can provide valuable insights into the students' understanding of the project's challenges and achievements.
4. *Presentation and defense* Have students present their projects to their peers, instructors, or an external panel. Require them to defend their project choices, methodologies, and outcomes. This can showcase their problem-solving skills and communication abilities.
5. *Feedback and reflection* Regularly collect feedback from students about their experiences with the assignments and the overall learning process. Please encourage them to reflect on their progress, challenges, and lessons learned.
6. *Client or industry feedback* If applicable, involve external clients or industry experts to evaluate the students' projects. This provides an authentic, real-world assessment of the projects' applicability and quality.

Our study emphasizes how combining structured and unstructured data analysis in teaching enhances students' data analysis and critical thinking skills. This approach helps students develop comprehensive analytical skills, enabling them to adapt to the constantly changing digital learning environment.

By combining these assessment methods, the student's learning outcomes and the effectiveness of the teaching strategy. The project outcomes demonstrate that focusing on real-world problem-solving, data integration, and programming skills is vital evidence of the strategy's success.

Results

Table 2 shows the scores of students from five different colleges in five assessments. For each assessment, students can get up to 3 points. They get 3 points if they do all three things needed for the evaluation, 2 points if they do two, and 1 point if they do one. If they do not do anything needed, they get 0 points.

The assessments are about looking at data in different ways. There are assessments for checking data, looking at text, making charts, and thinking about the data's meaning. The final score in the table is a total of how well the students did in all five assessments. Some students did very well and got the highest score of 3, but some did not do as well in some assessments and got lower scores.

This way of giving scores helps us see how well each student did in each assessment. It makes sure students learn what they are supposed to in each evaluation. The system also tells students to try their best in all parts of the assessment to get the highest score.

Principal Component Analysis (PCA) is a way to make complicated data more accessible and understandable (Ivosev et al., 2008). It keeps the essential parts of the data but makes it simpler. Here is an explanation of the PCA analysis process to visualize student performance by the college:

1. *Step 1* Data Gathering—We start with a dataset containing students' scores for five assignments and a final demo across different colleges or fields of study.

Table 2 The scores for all assignments of each enrolled student in five different colleges (https://docs.google.com/spreadsheets/d/1TZjqplNEM61RR8Hofk9KUWNn7u7Q4vTzST5crr4hrCk/edit?usp=sharing)

No	Fields of study	Assessment 1	Assessment 2	Assessment 3	Assessment 4	Assessment 5	Final
1	College of Education	3	3	3	3	3	3
2	College of Education	3	3	1	2	2	2
3	College of Education	3	3	3	3	3	2
4	College of Education	3	3	2	2	2	2
5	College of Education	3	3	3	2	3	3
6	College of Liberal Arts	3	3	3	2	3	3
7	College of Liberal Arts	3	3	3	2	2	2
8	College of Management	3	3	0	1	1	2
9	College of Management	3	3	0	2	2	2
10	College of Management	3	3	3	3	3	3
11	College of Science	3	3	3	2	2	3
12	College of Science	3	3	3	2	3	2
13	College of Science	3	3	0	2	2	2
14	College of Science	3	3	3	2	3	2
15	College of Science	3	3	1	3	3	3
16	College of Science	3	3	3	3	3	3
17	College of Science	3	3	1	2	2	3
18	College of Science	3	3	3	1	0	2
19	College of Science	3	3	1	1	0	3
20	College of Science	3	0	1	2	2	2
21	College of Science	3	3	3	2	2	3
22	College of Science	3	3	1	2	2	2
23	College of Technology and Engineering	3	0	0	2	2	3
24	College of Technology and Engineering	3	2	2	3	2	2
25	College of Technology and Engineering	3	3	3	3	3	2
26	College of Technology and Engineering	3	3	3	3	3	3
27	College of Technology and Engineering	3	3	3	2	2	2
28	College of Technology and Engineering	3	3	3	3	3	3
29	College of Technology and Engineering	3	3	3	3	3	2
30	College of Technology and Engineering	3	3	1	2	2	2
31	College of Technology and Engineering	3	3	3	3	3	1
32	College of Technology and Engineering	3	3	3	3	3	3
33	College of Technology and Engineering	3	3	2	3	2	3
34	College of Technology and Engineering	3	3	3	3	3	2
35	College of Technology and Engineering	3	0	0	1	0	0
36	College of Technology and Engineering	3	3	3	2	2	2

Table 2 (continued)

No	Fields of study	Assessment 1	Assessment 2	Assessment 3	Assessment 4	Assessment 5	Final
37	College of Technology and Engineering	3	3	3	2	3	2
38	College of Technology and Engineering	3	3	3	2	3	2
39	College of Technology and Engineering	3	3	3	2	2	2
40	College of Technology and Engineering	0	3	3	2	0	2
41	College of Technology and Engineering	3	3	3	2	3	3

2. *Step 2* Data Transformation—Each student's performance is represented as a six-dimensional vector, with each dimension corresponding to one of the six assessments (Assignments 1–5 and Final Demo).
3. *Step 3* Dimensionality Reduction with PCA—To enhance our ability to gain insights and effectively visualize student performance, we utilize PCA. This technique reduces the dataset's dimensionality while retaining its essential information. In this instance, we reduce the data from six dimensions to two, which we will refer to as Principal Component 1 (PC1), Principal Component 2 (PC2), and Principal Component 3 (PC3).
4. *Step 4* Scatter Plot Visualization—The reduced data is used to create a scatter plot where each point on the plot represents an individual student. The position of each point on the scatter plot is determined by their PC1, PC2, and PC3 values derived from the PCA.
5. *Step 5* Labeling and Interpretation—To enhance the understanding of the scatter plot, we label each point with the corresponding college or field of study. Clusters or patterns in the scatter plot indicate groups of students performing similarly in the assignments and final demo. These clusters suggest that students from specific colleges have more cohesive learning outcomes.
6. *Step 6* Analysis—By analyzing the scatter plot, we can identify whether there are visible clusters of students based on their college or field of study. The proximity of points in the plot may indicate that students within the same college tend to have similar learning outcomes.
7. *Step 7* Conclusion—This analysis provides insights into how students from different colleges perform and whether there are any noticeable patterns or clusters in their learning outcomes. Educators and administrators can use this information to make informed decisions about curriculum improvements or tailored support for specific colleges based on student performance.

In the analysis presented in Fig. 2, Principal Component Analysis (PCA) is adeptly utilized to transform the academic grades detailed in Table 2 into a comprehensive three-dimensional visualization. This advanced visualization enables a deeper investigation into students' academic performance and learning patterns from various colleges. The key insights drawn from the PCA visualization underscore the current

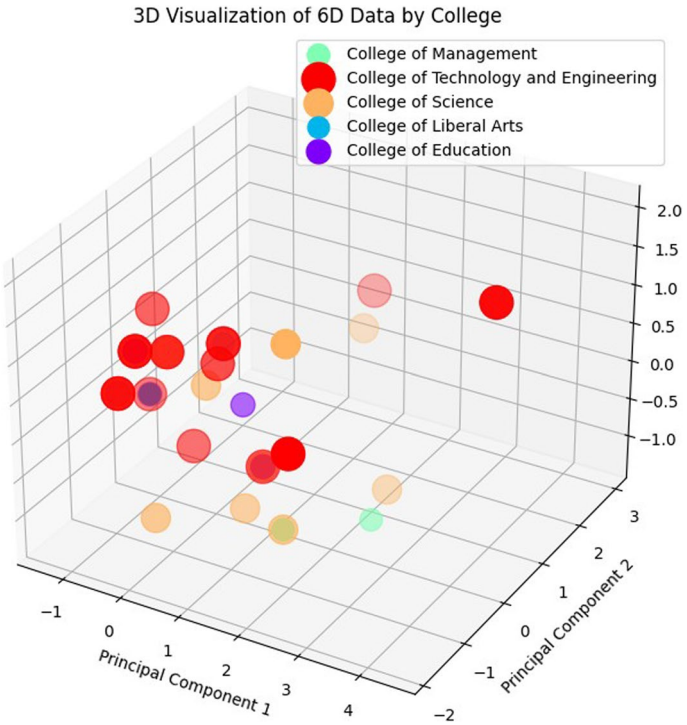

Fig. 2 3D Visualization of 6D data by college (https://colab.research.google.com/drive/1Xgl7uBdz7DqDZMh07RE8mnx7XZNC2JHt?usp=sharing)

state of student learning and highlight significant findings regarding their academic achievements.

The 3D scatter plot generated through PCA distinctly illustrates the distribution and clustering of students based on their performance metrics. This visualization is crucial for identifying and understanding the similarities and differences in learning outcomes among students from diverse academic backgrounds. Notably, the PCA results reveal that students from the "College of Education" and the "College of Technology and Engineering" exhibit comparable performance levels across assignments and final demonstrations. This observation suggests a noteworthy conclusion: the field of study may not be the primary determinant of a student's proficiency in acquiring data science and analysis skills.

Furthermore, the PCA visualization facilitates the identification of patterns or clusters among the student data, indicating groups of students with similar academic performances. Such clusters across different colleges suggest that irrespective of their specific disciplines. Students can achieve similar levels of success in data science and analysis competencies. This finding challenges traditional perceptions regarding the alignment of academic disciplines with particular skill sets and underscores the potential for interdisciplinary learning and achievement.

The analysis of the 3D PCA scatter plot not only provides a novel perspective on student performance but also serves as a compelling tool for educators and researchers. Educators can gain invaluable insights into their students' learning dynamics and achievement levels by examining the plot's proximity, distribution, and clustering

of data points (students). This, in turn, can inform targeted interventions, curriculum design, and teaching strategies to foster enhanced learning experiences across disciplines.

In essence, the PCA visualization results highlight the current learning situation among students, revealing that learning outcomes and academic success in data science and analysis are not confined to students' academic majors. This pivotal finding emphasizes integrating advanced visualization techniques in educational research to uncover more profound insights into student learning patterns and achievements.

Based on the provided data (see Table 3, 4, 5, 6, 7, 8), we can analyze the learning outcomes of different colleges as follows:

1. *Assignment 1 scores* Students from the "College of Education", "College of Liberal Arts", and "College of Management" achieved high scores in Assignment 1 with an average of 3.00 and no standard deviation (SD). Students from the "College of Science" and "College of Technology and Engineering" achieved lower average scores,

Table 3 The raw scores are used to calculate the average and standard deviation of Assignment 1 scores for each college across all assignments

Fields of study	The mean of Assignment 1	SD of Assignment 1	Max	Min
College of Education	3.00	0.00	3	3
College of Liberal Arts	3.00	0.00	3	3
College of Management	3.00	0.00	3	3
College of Science	3.00	0.00	3	3
College of Technology and Engineering	2.84	0.69	3	0

Table 4 The raw scores are used to calculate the average and standard deviation of Assignment 2 scores for each college across all assignments

Fields of study	The mean of Assignment 2	SD of Assignment 2	Max	Min
College of Education	3.00	0.00	3	3
College of Liberal Arts	3.00	0.00	3	3
College of Management	3.00	0.00	3	3
College of Science	2.75	0.87	3	0
College of Technology and Engineering	2.63	0.96	3	0

Table 5 The raw scores are used to calculate the average and standard deviation of Assignment 3 scores for each college across all assignments

Fields of study	The mean of Assignment 3	SD of Assignment 3	Max	Min
College of Education	2.70	0.67	3	1
College of Liberal Arts	3.00	0.00	3	3
College of Management	1.00	1.73	3	0
College of Science	1.92	1.16	3	0
College of Technology and Engineering	2.47	1.02	3	0

Table 6 The raw scores are used to calculate the average and standard deviation of Assignment 4 scores for each college across all assignments

Fields of study	The mean of Assignment 4	SD of Assignment 4	Max	Min
College of Education	2.40	0.55	3	2
College of Liberal Arts	2.00	0.00	2	2
College of Management	2.00	1.00	3	1
College of Science	2.00	0.60	3	1
College of Technology and Engineering	2.42	0.61	3	1

Table 7 The raw scores are used to calculate the average and standard deviation of Assignment 5 scores for each college across all assignments

Fields of study	The mean of Assignment 5	SD of Assignment 5	Max	Min
College of Education	2.60	0.55	3	2
College of Liberal Arts	2.50	0.71	3	2
College of Management	2.00	1.00	3	1
College of Science	2.00	1.04	3	0
College of Technology and Engineering	2.32	0.95	3	0

Table 8 The raw scores are used to calculate each college's average and standard deviation of Final Demo scores across all assignments

Fields of study	Mean of final demo	SD of final demo	Max	Min
College of Education	2.40	0.55	3	2
College of Liberal Arts	2.50	0.71	3	2
College of Management	2.33	0.58	3	2
College of Science	2.50	0.52	3	2
College of Technology and Engineering	2.16	0.76	3	0

with "College of Science" showing a lower SD and "College of Technology and Engineering" having a slightly higher SD (see Table 3).

2. *Assignment 2 scores* Most college students showed uniform performance in Assignment 2, with average scores of 3.00 and no SD. "College of Science" students achieved lower average scores and exhibited higher variability with a higher SD. "College of Technology and Engineering" students also achieved lower average scores with a moderate SD (see Table 4).
3. *Assignment 3 scores* "College of Liberal Arts" students excelled in Assignment 3 with an average score of 3.00 and no SD. "College of Education" students performed well, with an average score of 2.70 and a moderate SD. "College of Technology and Engineering" students achieved slightly lower average scores with a higher SD (see Table 5).
4. *Assignment 4 scores* "College of Education" and "College of Liberal Arts" students performed well in Assignment 4 with average scores of 2.40 and 2.00, respectively,

and moderate SD. "College of Management" students had relatively lower performance, with an average score of 2.00 and a higher SD. "College of Science" and "College of Technology and Engineering" students achieved lower average scores, with "College of Science" having a lower SD and "College of Technology and Engineering" having a slightly higher SD (see Table 6).

5. *Assignment 5 scores* "College of Education", "College of Liberal Arts", and "College of Management" students performed well in Assignment 5 with average scores of 2.60, 2.50, and 2.00, respectively. "College of Science" students had relatively lower performance, with an average score of 2.00 and a higher SD. "College of Technology and Engineering" students achieved lower average scores with a moderate SD (see Table 7).
6. *Final demo scores* Students from different colleges consistently performed in final demo scores, with average scores ranging from 2.16 to 2.50. "College of Management" students had a somewhat lower performance, with the lowest average score of 2.33 and a moderate SD (see Table 8).

The radar chart visualization in Fig. 3 distinctly illustrates the learning performance of students from different colleges across various assignments, highlighting significant differences in their academic accomplishments. This visualization enables a comprehensive analysis of how students from each college fare in specific areas, offering a detailed insight into the learning dynamics across disciplines. Key observations from the radar chart include the following:

1. *Broad performance spectrum* The "College of Education" and the "College of Liberal Arts" demonstrate comprehensive coverage across the radar chart's axes, indicating strong and consistent performance across multiple assignments. This suggests that students from these colleges have developed a versatile skill set, enabling them to excel in diverse academic challenges.
2. *Identified learning gaps* The radar chart shows that the "College of Management" has a more constrained presence, with specific axes revealing closer proximity to the center. This pattern highlights areas where students from this college exhibit lower performance, signaling potential gaps in their understanding or application of the course material.
3. *Varied achievements* Students from the "College of Science" and the "College of Technology and Engineering" display varied distances from the center across different axes of the radar chart. Such variability signifies that while these students may excel in specific assignments, they encounter challenges in others. This mixed performance points to a heterogeneity in their learning experiences and outcomes.

These insights drawn from the radar chart are pivotal for understanding the nuanced learning statuses of students across different colleges. They reveal areas of academic strength and pinpoint specific challenges students face, guiding educators in tailoring their instructional strategies to meet diverse learning needs. By leveraging this detailed analysis, educators can implement targeted interventions to support

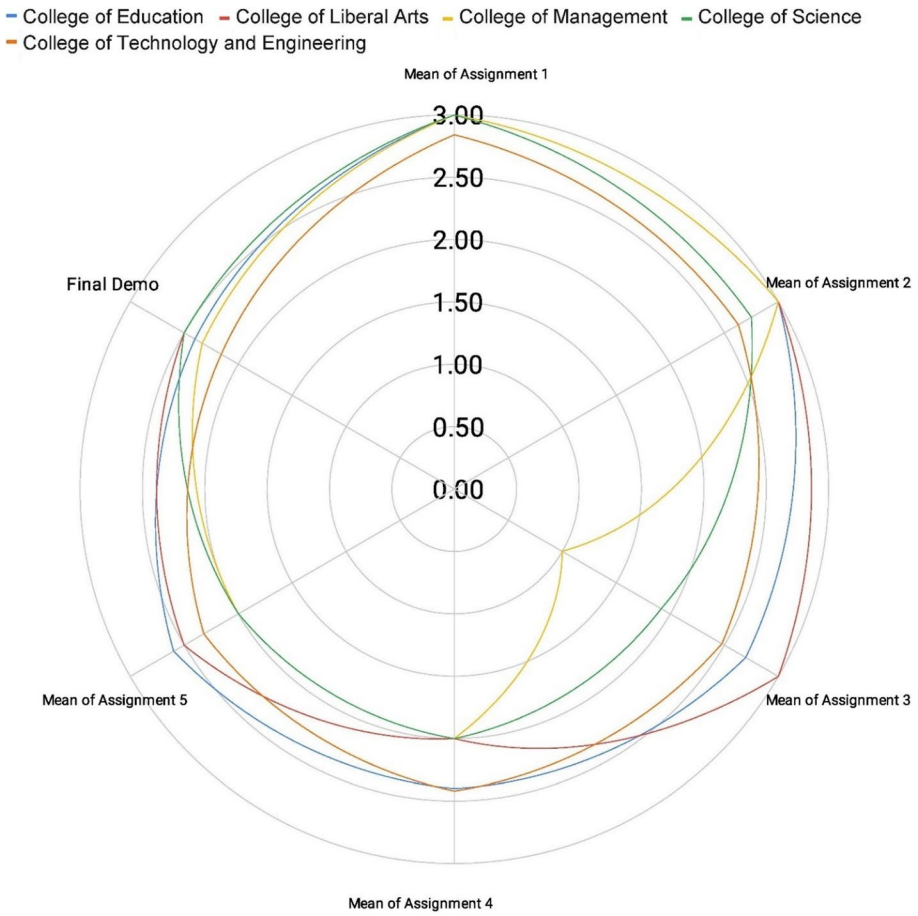

Fig. 3 Comparative performance of different colleges in multiple assignments

students in areas of difficulty, thereby enhancing the overall educational experience and outcomes for students across all colleges.

Furthermore, the positive feedback and successful project showcase shared by students, as highlighted in the provided YouTube playlist, validate the effectiveness of the teaching methodologies employed in the course. This external validation underscores the importance of a data-informed approach to understanding and improving student learning experiences, ensuring that teaching strategies are responsive and impactful.

The final project results are available on YouTube, and following the peer review, the feedback from students offers valuable insights into their learning experiences in the course (https://youtube.com/playlist?list=PLH3VeiMX0ckgYsdOMhixapnV8TfzrHJkp&si=BvRCT5L6wDWsxJqz). The course has proven highly beneficial, as students express gratitude for the rich content and guidance the instructors and teaching assistants provide. The final projects have stood out, showcasing the students' collective proficiency in applying the knowledge they acquired throughout the course to real-world applications.

This quantitative analysis reveals a consistent pattern where students from the "College of Education" and "College of Liberal Arts" generally performed better across assignments, suggesting that their learning strategies might effectively align with the assessment techniques. In contrast, the "College of Science" and "College of Technology and

Engineering" showed variability and generally lower performance, indicating potential mismatches between student learning styles or preparatory backgrounds and the assessments' demands. The "College of Management" showed significant variability, particularly in Assignment 3, which could reflect specific challenges or a need for curriculum adjustments to support these students better.

Discussion

The discussion is based on the scores from five assessments conducted throughout the course, with students from five colleges participating.

1. *Assessment scores distribution* The scoring system for each assessment was designed to objectively measure students' understanding and application of data analysis concepts. Students could earn up to 3 points for each assessment, with the final score in the table representing their cumulative performance across all five assessments. Our analysis revealed a diverse range of scores, reflecting the varying levels of students' engagement and comprehension of the subject matter.
2. *Performance across different disciplines* Students from different fields demonstrated varying strengths and challenges. For instance, students from the College of Education and Technology and Engineering generally performed well. In contrast, students from the College of Liberal Arts excelled in assignments requiring critical thinking and text analysis.
3. *Objective grading and standardized assessment* A standardized rubric was employed across all assessments to ensure fairness and objectivity in grading. This rubric focused on critical skills such as data analysis accuracy, clarity of data visualization, and effectiveness in problem-solving. It applied to all students, regardless of their academic background, ensuring that the grading system was unbiased and comprehensive.
4. *Feedback integration and learning progression* Students received regular feedback throughout the course, which was instrumental in their learning progression. This feedback, coupled with the collaborative learning environment and the modular structure of the assignments, allowed students to steadily build their skills, as evidenced by improved scores from the initial to the final assessments.
5. *Performance insights and clusters* Utilizing Principal Component Analysis (PCA), we visualized the performance data to identify patterns or clusters among the students from different colleges. The PCA plot (see Fig. 2) indicated that while there were clusters of students with similar performance levels, students from diverse academic backgrounds demonstrated comparable abilities in grasping data science and analysis skills.

Our course suggests that active learning and inclusive teaching methodology have effectively enhanced students' data analysis and critical thinking skills. These findings are instrumental in understanding the impact of our teaching strategy on students from diverse academic backgrounds and skill levels.

After reflecting on the diverse programming experiences of the students, we recognize the potential benefit of refining our course standards. Adjusting the curriculum to cater

to the varying skill levels can ensure a more inclusive and effective learning environment. Moreover, extending the duration of the final project could provide additional support for those with less programming experience, allowing them more time to grasp complex concepts and apply them effectively. This extension could also offer more advanced students the opportunity to delve deeper into sophisticated aspects of data analysis, thereby enriching their learning experience.

Conclusion

This study explored the transformative role of active learning in Educational Big Data Analytics. Our findings highlight the significant impact of a hands-on, student-centered approach in developing critical data analysis and problem-solving skills. Implementing a structured curriculum integrating Python tools and ChatGPT APIs showed marked improvements in students' ability to effectively engage with and analyze educational data.

Our study contributes to the broader discourse on educational methodologies in the big data and AI age. Using exploratory data analysis, problem-solving tasks, and collaborative projects facilitated a deeper, more practical understanding of data analytics concepts among students. These methodologies enriched the students' learning experiences and equipped them with essential skills for navigating the increasingly data-driven educational landscape.

We acknowledge the challenges posed by the diverse programming experience levels of the students. Our tailored teaching approach, including modular learning designs and differential assessment difficulties, effectively catered to this diversity. The positive outcomes observed across various academic backgrounds and skill levels underscore the efficacy of our inclusive educational strategies.

To assess the enduring impact of our teaching strategy on students' data analysis skills and critical thinking abilities, we propose conducting a follow-up study. Such a study would allow us to track the long-term effects of our approach and provide insights into how these skills evolve in a real-world setting. This longitudinal research could also offer valuable feedback for continuous improvements in our teaching methodologies, ensuring they remain practical and relevant in an ever-evolving educational landscape.

In conclusion, this study underscores the importance of active learning strategies in Educational Big Data Analytics. By empowering students through practical, interactive learning experiences, educators can foster a generation of learners adept at harnessing the power of data. As we navigate this data-rich era, educational practices must evolve to prepare students for the challenges and opportunities that lie ahead.

Abbreviations

EDA	Exploratory Data Analysis
LDA	Latent Dirichlet Allocation
MVP	Minimum Viable Product
PCA	Principal Component Analysis
SD	Standard deviation

Acknowledgements
Not applicable.

Author contributions
Y-CT conceptualized and designed the study, conducted the research, and wrote the manuscript. All authors have read and approved the final manuscript.

Funding
This research was supported by the Taiwan Ministry of Education's Education Big Data Micro-Program Project.

Availability of data and materials
The datasets generated and analysed during the current study are available in the (i) Data: https://docs.google.com/spreadsheets/d/1TZjqpINEM61RR8Hofk9KUWNn7u7Q4vTzST5crr4hrCk/edit?usp=sharing. (ii) Code: https://colab.research.google.com/drive/1Xgl7uBdz7DqDZMh07RE8mnx7XZNC2JHt?usp=sharing. (iii) Projects: https://youtube.com/playlist?list=PLH3VeiMX0ckgYsdOMhixapnV8TfzrHJkp&si=BvRCT5L6wDWsxJqz.

Declarations

Ethics approval and consent to participate
This study was conducted in full compliance with ethical standards, and verbal consent was obtained from all participants. The rationale for using verbal consent is grounded in the context of the educational environment. As the class instructor, I secured verbal consent from students to use their learning artifacts on a public GitHub repository to document and showcase their learning journey.

Consent for publication
Written consent for identifying information was not required in this instance, as all data utilized in this study comes from information that participants voluntarily made public on their GitHub profiles. The decision to use publicly available data was predicated on the understanding that participants were aware of the general nature of their contributions and consented to their visibility by sharing their information in a public GitHub repository. This approach aligns with ethical guidelines for using publicly available data, ensuring that all participant data included in this study was shared under the participants' consent through their public GitHub profiles. Proof of the general nature of this data can be provided upon request.

Competing interests
The authors declare that they have no competing interests.

Received: 2 November 2023 Accepted: 25 March 2024
Published online: 01 April 2024

References

Aslan, A. (2021). Problem-based learning in live online classes: Learning achievement, problem-solving skills, communication skills, and interaction. *Computers & Education, 171*, 104237.
Brewer, E., & Cunningham, K. (2023). *Integrating study abroad into the curriculum: Theory and practice across the disciplines*. Taylor & Francis.
Donoghue, T., Voytek, B., & Ellis, S. E. (2021). Teaching creative and practical data science at scale. *Journal of Statistics and Data Science Education, 29*(1), S27–S39.
Husni, H. (2020). The effect of inquiry-based learning on religious subjects learning activities: An experimental study in high schools. *Jurnal Penelitian Pendidikan Islam, 8*(1), 43–54.
Ivosev, G., Burton, L., & Bonner, R. (2008). Dimensionality reduction and visualization in principal component analysis. *Analytical Chemistry, 80*(13), 4933–4944.
Milo, T., & Somech, A. (2020). Automating exploratory data analysis via machine learning: An overview. In *Proceedings of the 2020 ACM SIGMOD international conference on management of data* (pp. 2617–2622). https://doi.org/10.1145/3318464.3383126
Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education, 93*(3), 223–231.
Qureshi, M. A., Khaskheli, A., Qureshi, J. A., Raza, S. A., & Yousufi, S. Q. (2023). Factors affecting students' learning performance through collaborative learning and engagement. *Interactive Learning Environments, 31*(4), 2371–2391. https://doi.org/10.1080/10494820.2021.1884886
Romanow, D., Napier, N. P., & Cline, M. K. (2020). Using active learning, group formation, and discussion to increase student learning: A business intelligence skills analysis. *Journal of Information Systems Education, 31*, 218–231.
Shi, Y. (2022). Advances in big data analytics. Adv Big Data Anal.
Tsai, Y.-C. (2023). Empowering learner-centered instruction: Integrating ChatGPT Python API and tinker learning for enhanced creativity and problem-solving skills. In *Innovative Technologies and Learning: Proceedings of the 6th International Conference, ICITL 2023* (pp. 531–541). https://doi.org/10.1007/978-3-031-40113-8_52
Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software, 6*(60), 3021.

Publisher's Note