Enhancing Open Access to Data Science Education: Analyzing Skill Patterns Using LDA and K-Means Clustering in the Learning Path Index Dataset

Ibrahiem M. M. El Emary[1,*], Iwona Chomiak-Orsa[2], Elwira Gross-Gołacka[3]

[1] King Abdulaziz University, Kingdom of Saudi Arabia

[2] Wroclaw University of Economics and Business, Wroclaw, Poland

[3] University of Warsaw, Warsaw, Poland

ABSTRACT

This study examines the application of Latent Dirichlet Allocation (LDA) and K-Means clustering techniques to analyze the Learning Path Index Dataset, with the aim of identifying and categorizing data science education skills. By employing these machine learning models, the research reveals distinct skill patterns and clusters that characterize the dataset, highlighting prevalent skills and potential gaps in data science education accessible through open educational resources (OER). The findings demonstrate specific clusters of beginner to advanced data science topics, offering insights into the accessibility and distribution of educational content. These results can guide educators and platform developers in enhancing the structure and delivery of data science education, thereby improving learner outcomes and resource allocation. The study also discusses the broader implications for educational strategy and policy, emphasizing the role of targeted analytics in optimizing educational offerings in an increasingly digital landscape. Future research directions include expanding the dataset and applying similar analytical frameworks to other fields within open education to further validate and refine these findings.

Keywords Data Science Education, Topic Modeling, Latent Dirichlet Allocation (LDA), K-Means Clustering, Open Educational Resources (OER)

Introduction

The global demand for skills in data science, machine learning, and artificial intelligence (AI) has surged dramatically in recent years due to the rapid pace of technological advancement and AI's integration into diverse sectors. This growing demand is evidenced by an increasing number of job advertisements seeking individuals with expertise in these areas, demonstrating the necessity for a blend of both technical and business-oriented capabilities. Verma et al [1] emphasize the significance of having a strong grasp of business concepts for strategic decision-makers, alongside technical proficiencies in data analytics and visualization, to maximize the efficacy of AI solutions. Organizations across various industries are striving to leverage AI-driven innovations to enhance productivity and drive innovation, further fueling the demand for professionals proficient in these domains [2].

In response to this rising demand, the reliance on online resources for skill acquisition has grown notably, especially with the accelerated shift to online learning environments triggered by the COVID-19 pandemic. Educational institutions have adopted online platforms to maintain continuity and broaden

Submitted 1 February 2025
Accepted 29 March 2025
Published 3 June 2025

Corresponding author
Ibrahiem M. M. El Emary,
omary57@hotmail.com

Additional Information and Declarations can be found on page 110

DOI: 10.63913/ail.v1i2.45

How to cite this article: I. M. M. El Emary, I. Chomiak-Orsa, E. Gross-Gołacka, "Enhancing Open Access to Data Science Education: Analyzing Skill Patterns Using LDA and K-Means Clustering in the Learning Path Index Dataset," *Artif. Intell. Learn.*, vol. 1, no. 2, pp. 93-113, 2025.

access, showcasing the benefits of flexibility, accessibility, and global reach in digital education [3]. Massive open online courses (MOOCs) have emerged as pivotal tools for democratizing access to high-quality educational materials, enabling learners to progress at their own pace while customizing their educational experiences to align with individual goals [4]. However, as Zhao and Liu [5] note, existing online educational platforms often struggle to keep up with the surging demand for data access, highlighting the need for robust infrastructures to manage and disseminate educational content efficiently.

Navigating educational resources in open access environments presents a range of challenges for learners, despite the transformative potential of platforms like MOOCs and OER. While these tools have democratized access to learning and bridged geographical and socio-economic gaps, the overwhelming volume of available resources can lead to significant difficulties. Learners often experience information overload, which complicates their ability to synthesize knowledge from diverse sources. Evans et al [6] observe that open access platforms, while enhancing educational equity, sometimes overwhelm users with excessive information, making it challenging to integrate these resources into structured learning paths. The lack of a consistent peer review process among many open access materials further complicates the issue, as learners must assess the quality and relevance of available content on their own, leading to inconsistent educational experiences [7].

Language barriers further exacerbate these challenges, limiting access for non-English-speaking learners, who may find it difficult to engage fully with predominantly English-language resources. Montoya and Soledad [8] emphasize that this language disparity can hinder learners' ability to leverage high-quality educational materials, effectively marginalizing a segment of the global learning community. Thakran and Sharma [9] expand on this by highlighting how geographical and demographic barriers, compounded by a scarcity of qualified educators, can make access to quality open resources even more complex. Furthermore, differing cultural contexts and educational practices may not align with the formats and approaches used by open resources, reducing their efficacy for diverse learners.

The transition to online learning, especially during the COVID-19 pandemic, has shed further light on the challenges of open access education. While online platforms offer flexibility, Nsengimana et al [10] note that learners often struggle with the self-discipline and motivation required for effective learning in less structured environments. This lack of structured guidance can lead to isolation, particularly for those more accustomed to traditional, face-to-face learning settings. Kurelović [11] points out that a shift to open educational resource policies within institutions often faces resistance from educators and learners who are entrenched in conventional teaching methods and cultural norms.

The integration of technology into education also introduces barriers that must be addressed. Although OER and online learning can enhance educational access, their effectiveness often hinges on the technological proficiency of both educators and learners [12]. Unequal access to necessary technology and reliable internet connectivity remains a critical obstacle, creating a digital divide that hampers learners' ability to engage with these resources. Addressing these disparities is crucial for ensuring that the potential benefits of open educational resources reach a broader and more inclusive audience.

The Learning Path Index Dataset functions as a curated resource specifically designed to improve access to high-quality learning paths in the fields of data science, machine learning, and AI. It provides a structured repository that empowers both educators and learners to efficiently navigate the extensive range of available educational resources. This dataset addresses the increasing demand for personalized educational experiences, which allow learners to tailor their paths based on individual goals and learning preferences. As Joseph et al [13] emphasize, the ability to customize learning content enhances student engagement and success by aligning educational materials with personal interests and objectives. The structured nature of the dataset ensures that users can seamlessly find, evaluate, and engage with resources that are most relevant to their learning journeys, reducing the complexity often associated with navigating vast digital learning landscapes.

A key strength of the Learning Path Index Dataset lies in its systematic organization of resources according to specific criteria, such as subject area, complexity, and user preferences. This method streamlines the process of identifying appropriate learning paths, making it especially valuable in open-access environments where an overwhelming volume of information can hinder effective learning. As noted by Joseph et al [13], the challenge of sifting through extensive educational material is a common obstacle for learners. By offering a structured framework, the dataset mitigates information overload, enabling learners to concentrate on high-quality, targeted content that aligns with their needs. Such a focused approach not only maximizes learning efficiency but also ensures that learners have access to relevant, curated resources designed to support their educational progress.

Data mining serves a pivotal function in evaluating and optimizing educational content by enabling the analysis of large volumes of educational data to uncover patterns and derive actionable insights. Within educational environments, this process, often referred to as Educational Data Mining (EDM), facilitates the improvement of both skills coverage and accessibility. Castro et al [14] describe how EDM focuses on extracting meaningful patterns from data generated through learning interactions, thereby enhancing the quality of educational offerings. This application of data mining provides institutions with a framework to assess and refine the effectiveness of their educational content. Specifically, it allows for a detailed examination of how well specific skills are covered in educational materials, which can help identify gaps in content delivery. Brambila and González [15] illustrate how data mining techniques, such as association rule mining, can reveal relationships between learner characteristics and academic outcomes, thereby empowering educators to tailor content more effectively to meet diverse learner needs.

Optimizing the accessibility of educational content is another critical dimension where data mining has a transformative impact. Analyzing demographic and engagement data offers institutions a window into the barriers that learners face in accessing educational resources. Wanjau et al [16] demonstrate the potential of predictive models to uncover enrollment trends, particularly in STEM fields, providing valuable insights into how courses can be made more accessible to underrepresented groups. This data-driven approach enables institutions to design targeted interventions that address access disparities and promote inclusivity, ensuring that learners from diverse backgrounds can acquire essential skills. Additionally, by examining learner engagement patterns,

educational content can be adapted and refined to enhance its relevance and accessibility, ultimately leading to more equitable educational outcomes.

The primary objective of this study is to identify patterns in skill coverage within open-access data science resources by employing LDA and K-Means clustering. This analytical approach allows for the extraction of latent topics and skill clusters, offering insights into the structure and focus areas of data science educational content. The use of LDA facilitates the identification of underlying themes and skill domains within the Learning Path Index Dataset, while K-Means clustering groups courses and learning materials based on their topical similarities.

Literature Review

Data Science Education Trends and Challenges

The literature on online education in data science and artificial intelligence (AI) reveals a complex and evolving landscape, shaped by issues of accessibility, inclusivity, and the effectiveness of instructional approaches. As the transition to online learning accelerated in response to the COVID-19 pandemic, educational institutions faced increased pressure to provide equitable and effective learning experiences. Accessibility remains a significant hurdle, particularly for students in developing regions. Aboagye et al [17] point out that inadequate internet connectivity and limited access to digital devices disproportionately impact learners from marginalized communities, exacerbating educational inequalities. Zhou [18] highlights how institutional infrastructure deficiencies, such as insufficient library resources to support remote learning, further impede students' ability to fully participate in data science and AI courses. These barriers underline the critical need for targeted interventions to bridge the digital divide and ensure all learners can engage meaningfully with online educational resources.

Inclusivity is also a critical aspect of online education, influenced by the design and delivery of learning materials. Ullah et al [19] emphasize the importance of motivational factors in fostering student engagement within digital learning environments. The absence of traditional face-to-face interactions often diminishes learners' sense of community and motivation, which are vital for successful learning. Addressing this challenge requires the creation of inclusive and adaptable online platforms tailored to diverse learning needs. Muslimin and Harintama [20] argue that flexibility in online learning can empower students to choose conducive learning environments, thereby enhancing their educational experiences. However, they stress that this flexibility must be paired with sufficient support systems to ensure that all students can thrive in online settings.

The effectiveness of online education in data science and AI depends on the quality of instructional design and the pedagogical strategies employed. Paudel [21] asserts that effective online courses should incorporate interactive elements, such as simulations and collaborative activities, to foster engagement and improve learning outcomes. The rapid shift to online learning during the pandemic has led educators to adopt innovative and interactive teaching methodologies, which often surpass traditional approaches in engaging students. Akbar et al [22] found that the quality of video content and the overall user experience significantly influence learner perceptions of online platforms. High-quality, adaptable content is essential for maintaining engagement and

ensuring the effective delivery of complex data science concepts.

Technology plays a pivotal role in enhancing the educational effectiveness of online learning. Adedoyin and Soykan [23] emphasize the necessity for educators to understand both the potential and limitations of online learning technologies to design impactful courses. This includes recognizing the value of synchronous and asynchronous learning modalities, which cater to different learning styles and preferences. Additionally, the integration of learning analytics offers valuable insights into student performance, enabling timely interventions and personalized support [24]. Such data-driven approaches are essential for continually refining online educational programs and ensuring their alignment with learner needs and industry demands.

Topic Modeling in Education

LDA is a widely used statistical model in educational content analysis for identifying latent topics and common themes within large collections of text data. LDA conceptualizes documents as mixtures of topics and topics as mixtures of words, which allows researchers to uncover hidden structures in unstructured text. This capability is particularly useful in educational contexts, where vast amounts of data, from course materials to student feedback, need to be analyzed for meaningful insights. Schwartz et al [25] demonstrated the power of LDA by extracting significant topics from extensive educational materials, such as chapter-length texts, thereby revealing thematic structures that can inform curriculum design. This process not only facilitates a deeper understanding of content focus areas but also helps educators align their teaching materials with the identified themes, ensuring that the curriculum remains relevant and comprehensive.

In addition, LDA plays a critical role in curriculum development through its ability to analyze student responses and feedback. Insights drawn from LDA can help educators discern which topics resonate most with learners and which areas may require further emphasis or additional resources. Inoue et al [26] illustrated this utility by applying LDA to free-text responses, revealing how the COVID-19 pandemic impacted nursing research. The uncovered themes informed adjustments to educational practices, underscoring LDA's potential to guide real-time curriculum changes based on learner feedback. This adaptability ensures that educational content evolves to meet the changing needs and expectations of learners.

Beyond identifying specific topics, LDA can reveal common themes that span across various educational materials, thereby enriching the learning experience. Fang et al [27] employed LDA to analyze library electronic references and identified evolving research topics and patterns, a methodology that can similarly be applied to educational content. Understanding these overarching themes enables educators to foster interdisciplinary connections, creating integrated learning experiences that bridge different courses or disciplines. Such insights can promote a more cohesive educational framework, enhancing student engagement and fostering a broader understanding of complex subject areas.

Insights from LDA analyses can also inform the design of educational interventions aimed at boosting student engagement and learning outcomes. Yin and Yuan [28] highlighted the potential of LDA to detect trends in blended learning environments, helping educators adapt their teaching strategies to

better align with students' evolving needs and interests. This adaptability is vital in an educational landscape that is continuously reshaped by technological advancements and shifting learner preferences. However, it is important to note that LDA's effectiveness relies heavily on the quality and relevance of input data. Guo [29] cautions that variations in dataset characteristics can impact LDA performance, while Huang et al. [30] emphasize the need for expert input to determine the optimal number of topics, ensuring that the identified themes accurately reflect the content.

LDA is a generative probabilistic model that has become an essential tool for topic modeling across various domains, including educational research. At the heart of LDA is the probabilistic relationship between words and topics, represented mathematically as P(w|z), where w denotes a word, and z represents a topic. This formula encapsulates the probability that a word belongs to a given topic, offering a means of identifying and understanding the thematic structure of large text corpora. LDA assumes that documents are composed of multiple topics and that each topic is characterized by a distribution of words. This approach enables researchers to model the hidden thematic layers in text data and infer patterns that are not immediately visible. Pérez-Encinas and Rodríguez-Pomeda [31] illustrate how this probabilistic framework allows for the extraction of meaningful themes, aiding in content analysis and revealing the latent structures within educational materials.

In the LDA process, each document d is associated with a distribution over topics, denoted as P(z|d), which is sampled from a Dirichlet distribution. For each word w in the document, a topic z is then chosen according to this distribution, and the word itself is generated based on the topic-specific distribution P(w|z). This generative process highlights LDA's ability to represent documents as mixtures of topics and to describe topics in terms of probabilistic distributions over words. Fang et al [27] explain that this methodology allows LDA to uncover latent themes, making it a powerful tool for analyzing complex educational content, such as course descriptions or student feedback. The probabilistic nature of LDA ensures that topics reflect the inherent variability and diversity present in textual data, offering a robust mechanism for thematic exploration.

The distribution P(w|z) plays a pivotal role in LDA's effectiveness as a topic modeling approach. It facilitates the identification of the most representative words associated with each topic, thereby offering insights into the primary themes and focus areas within a corpus. In educational contexts, this capability can be used to analyze course materials or student responses to reveal prominent themes that resonate with learners [26]. By examining words with high probabilities under specific topics, educators gain a better understanding of the key concepts emphasized in their teaching resources and can make informed decisions regarding content alignment and curriculum development.

Additionally, P(w|z) enables researchers to compare and contrast topics across different documents, providing a basis for thematic similarity and distinction analysis. This comparative approach proves especially valuable in curriculum development, where understanding topic overlap and divergence among courses can inform the creation of cohesive learning pathways [32]. LDA's probabilistic nature also accounts for the inherent uncertainty in topic assignments, offering a nuanced perspective on how themes evolve and interact over time. Fang et al [27] note that this capability is essential for capturing fine-

grained trends within academic literature, allowing educators and researchers to track the emergence of new topics and shifts in focus within their fields. The insights derived from such analyses can inform instructional strategies and drive the continuous improvement of educational content.

Clustering Techniques for Curriculum Analysis

K-Means clustering has proven to be a valuable technique in educational research, particularly for grouping similar courses and content based on keywords and levels of difficulty. As an unsupervised learning algorithm, K-Means partitions data into distinct clusters, thereby enabling researchers and educators to analyze and optimize educational resources and offerings. This method has been applied across numerous studies, demonstrating its utility in categorizing educational content, assessing student engagement, and tailoring curricula to meet diverse learner needs. For example, Ghifari and Putri [33] employed K-Means clustering to analyze courses based on student grades, effectively grouping courses that exhibited similar performance trends.

K-Means clustering has also been leveraged to analyze student engagement and performance data, as demonstrated by Davies et al [34]. In their study of online flipped classrooms, they employed longitudinal K-Means cluster analysis to investigate student learning behaviors, identifying patterns that were instrumental in refining instructional strategies and improving learning outcomes. By clustering students based on their engagement metrics, educators were able to tailor interventions and provide targeted support, ultimately enhancing student success. This application underscores the potential of K-Means in uncovering actionable insights from educational data, making it a powerful tool for curriculum analysis and the optimization of teaching practices.

K-Means clustering is equally effective for evaluating course difficulty and analyzing content characteristics. Research conducted by Biwer et al. [35] focused on university students' self-regulation behaviors during online learning amid the COVID-19 pandemic. By clustering students based on their adaptive learning strategies, the study revealed insights into how different courses varied in terms of required difficulty and support. Such findings enable educators to adjust their teaching approaches and provide the appropriate level of challenge and assistance to diverse student groups. Kwasi and Gyimah [36] similarly utilized K-Means clustering to explore learner typologies in project-based learning environments. Their analysis identified distinct student profiles and highlighted the varied difficulties students encountered, thereby informing more inclusive and accessible course designs.

Despite its numerous benefits, K-Means clustering does present some challenges that must be considered. The algorithm's sensitivity to the initial selection of cluster centroids can lead to variability in clustering outcomes, as noted by Maulana and Anugrah [37]. Determining the optimal number of clusters also involves subjectivity and often requires validation through methods such as the Silhouette Score or the Elbow method. Furthermore, researchers must ensure that the features selected for clustering, such as keywords and difficulty levels, accurately reflect the characteristics of the courses being analyzed. These considerations are critical for ensuring that the insights derived from clustering analyses are both meaningful and actionable.

The K-Means clustering algorithm is a fundamental method in data analysis

used to partition a dataset into k distinct clusters. Its objective function is mathematically represented as:

$$\min \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2 \quad (1)$$

In this expression, C_i represents the i-th cluster, μ_i denotes the centroid of that cluster, and x refers to individual data points within the cluster. The primary goal of the K-Means algorithm is to minimize the total within-cluster variance, quantified by the sum of squared distances between each data point and the centroid of its assigned cluster. This minimization results in the formation of compact and well-separated clusters, where intra-cluster data points exhibit a high degree of similarity.

The components of the objective function play distinct and essential roles in the clustering process. Clusters C_i represent subsets of the dataset, with each containing data points that are more similar to each other than to points in other clusters. The number of clusters, k, is typically predefined before executing the algorithm. The centroid μ_i serves as the representative point for each cluster, calculated as the mean of all data points within that cluster, and is iteratively updated during the algorithm's execution. The distance metric $(|x - \mu_i|^2)$ measures the squared Euclidean distance between a data point x and its cluster's centroid, serving as a critical factor in determining the quality of the clustering outcome.

The objective function plays a pivotal role in guiding the K-Means clustering process. One of its primary functions is to minimize variance within clusters by reducing the sum of squared distances between data points and their respective centroids. This optimization leads to clusters that are dense and well-defined, with minimal intra-cluster distance. The iterative nature of the K-Means algorithm involves repeatedly assigning data points to the nearest centroid and recalculating centroids based on the current cluster assignments. This process continues until convergence, which occurs when data point assignments remain unchanged or the centroids stabilize. The minimization of the objective function serves as a measure of clustering quality, with lower values indicating tighter and more coherent clusters.

In educational research, K-Means clustering has proven useful for grouping similar courses or content based on characteristics such as keywords and difficulty levels. Researchers can analyze course descriptions or performance data to form clusters of courses with shared features, providing valuable insights for curriculum design and resource allocation. This application underscores the utility of the K-Means objective function in creating meaningful groupings that enhance the understanding of educational content and inform targeted teaching strategies. The ability of K-Means to form well-defined clusters based on data-driven patterns makes it an indispensable tool for educational analysis and improvement.

Importance of Open Educational Resources (OER)

OER have emerged as pivotal tools in improving access to data science skills training, particularly in an era where data literacy and technological competence are becoming critical in many sectors. OER are defined as freely accessible,

openly licensed educational materials designed for teaching, learning, and research. Their ability to remove traditional barriers to education, including cost and geographical constraints, underscores their transformative impact on the democratization of learning. Guo et al [38] emphasize that the globalization of education has been significantly facilitated by the availability of OER, allowing learners from diverse backgrounds to engage with high-quality resources. In the context of data science, a field marked by rapidly evolving technologies and a growing demand for skilled professionals, access to comprehensive, up-to-date educational resources can substantially enhance learners' opportunities for career advancement.

OER's potential extends beyond mere access; they offer a flexible, customizable approach to learning that can cater to diverse educational needs. Malykhin et al [39] highlight the adaptability of OER in developing critical job skills, as they can be tailored to individual learners' career goals and preferred learning styles. This adaptability is particularly relevant in data science education, where new tools, methodologies, and techniques emerge frequently. By leveraging OER, educators can provide learners with customizable pathways to acquire and apply essential skills, helping them remain competitive in a fast-changing industry.

Despite their benefits, OER face challenges in terms of quality and relevance, which can impact their effectiveness as educational tools. This study aims to address these issues by implementing a systematic framework for evaluating OER, focusing on quality assessment, user engagement metrics, and alignment with industry standards in data science training. Collecting user feedback and analyzing how learners interact with these resources will help ensure that OER remain relevant and effective for skill development. Palkova et al [40] suggest that OER can promote collaborative and sustainable educational practices, underscoring the importance of fostering learning communities around these resources. Creating collaborative environments where learners share insights and experiences can deepen understanding and encourage the practical application of data science concepts.

Additionally, leveraging technology to enhance the accessibility and usability of OER is critical. Incorporating interactive elements, such as real-world projects, simulations, and quizzes, can increase learner engagement and improve knowledge retention. Tang [41] points out that the shift to digital education has heightened the need for effective OER integration to maintain high educational standards. This study focuses on developing innovative, technology-driven OER solutions to enhance the learning experience for data science students. By combining a rigorous evaluation framework, collaborative learning models, and interactive technological elements, this research seeks to maximize the potential of OER in equipping learners with essential data science skills.

Method

The research method employed in this study involves a series of meticulously designed steps to guarantee a thorough and precise analysis. Figure 1 presents a detailed flowchart that outlines the comprehensive steps of the research methodology.

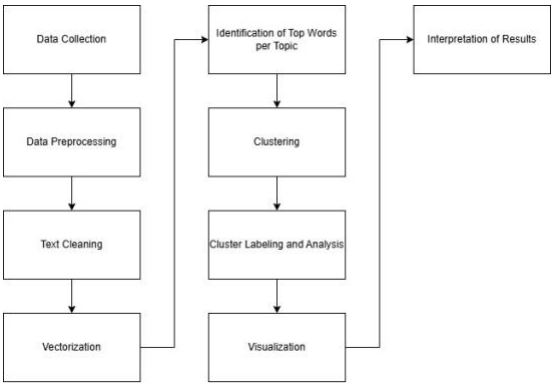

Figure 1 Research Method Flowchart

Data Overview

The Learning Path Index Dataset serves as the foundation for this research, providing a comprehensive collection of data science, machine learning, and courses. Key columns used in the analysis include Keywords_Tags_Skills_Interests_Categories, Course_Level, and Type_Free_Paid. The Keywords_Tags_Skills_Interests_Categories column contains relevant tags associated with course content, highlighting skill areas, topics, and learning objectives. Course_Level indicates the intended proficiency level of the course, ranging from beginner to advanced, while Type_Free_Paid specifies whether the course is freely available or requires payment.

An initial exploration of the dataset revealed a structured but diverse set of educational offerings. Summary statistics provided insights into the distribution of course types and levels, showcasing the variety of resources available to learners. For instance, a significant proportion of courses are categorized as beginner-level, reflecting the dataset's focus on foundational skills. The selected columns provided a basis for analyzing patterns in course offerings, keyword relevance, and accessibility, which are critical to achieving the study's objectives.

Exploratory Data Analysis (EDA)

To gain a deeper understanding of the dataset, EDA was conducted, focusing on key attributes such as Course_Level, Type_Free_Paid, and Keywords_Tags_Skills_Interests_Categories. Visualizations played a crucial role in summarizing and interpreting the data. A bar chart of the Course_Level distribution (figure 2) revealed that beginner-level courses dominated the dataset, with fewer offerings at intermediate and advanced levels. This pattern underscores the emphasis on introductory materials in the dataset, catering to learners new to data science.

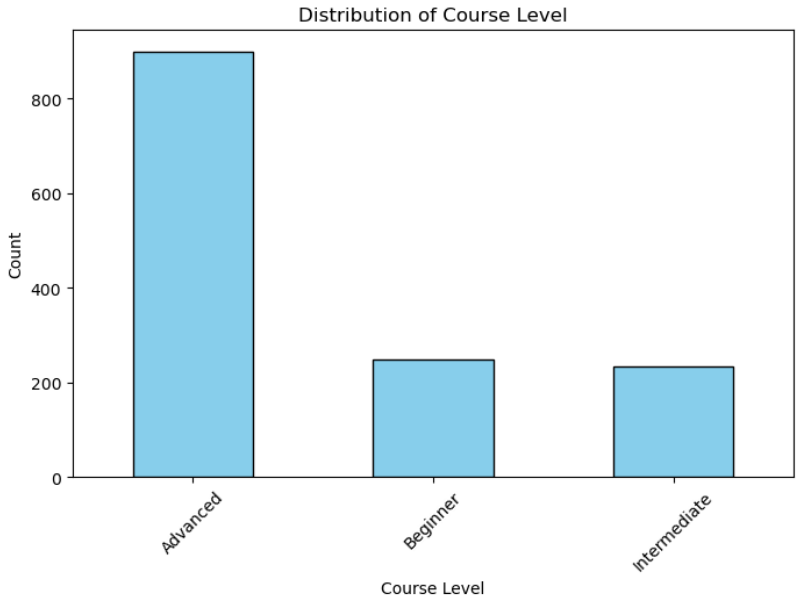

Figure 2 Distribution of Course Level

Another bar chart depicted the distribution of free versus paid courses (figure 3). Free courses represented a substantial majority, indicating the dataset's alignment with open educational resource principles. To analyze the keyword data, a function extracted and aggregated individual keywords from the Keywords_Tags_Skills_Interests_Categories column.

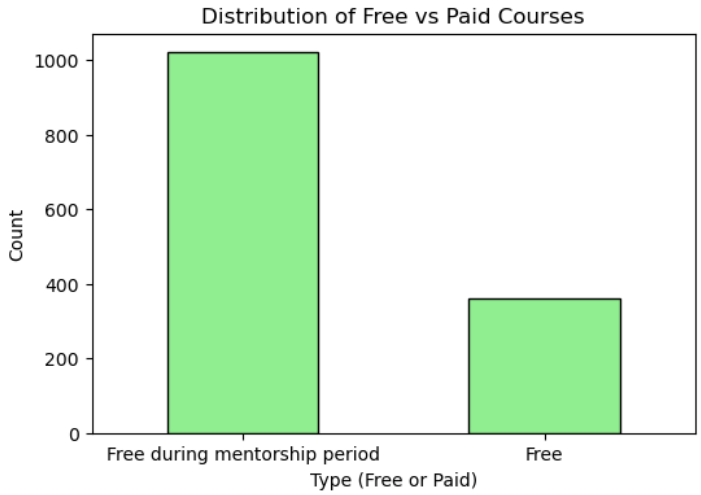

Figure 3 Distribution of Free vs Paid Courses

The top 10 most frequently occurring keywords were identified and visualized using a horizontal bar chart (figure 4). This analysis highlighted core concepts such as “machine learning,” “data analysis,” and “artificial intelligence,” which align closely with the study's focus on skill development in data science.

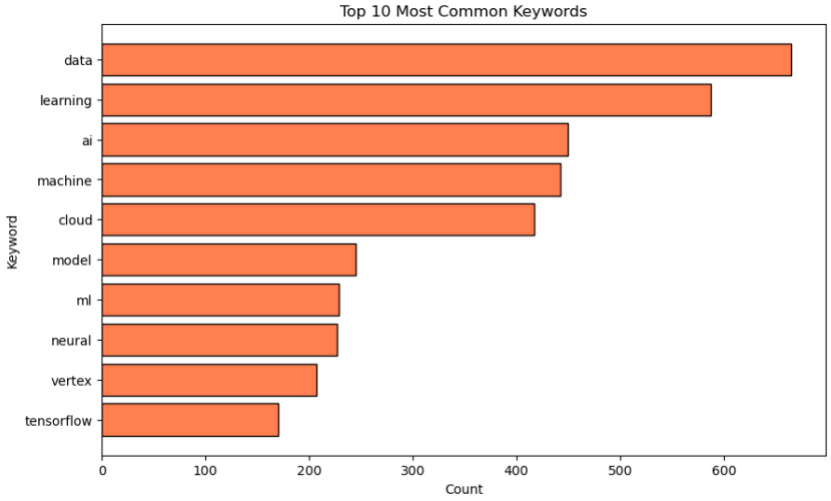

Figure 4 Top 10 Most Common Keywords

Additionally, a word cloud was generated (figure 5) to provide a visual representation of keyword diversity, offering a comprehensive overview of the dataset's thematic scope. These exploratory steps provided critical insights into the dataset's structure and content, informing subsequent analyses.

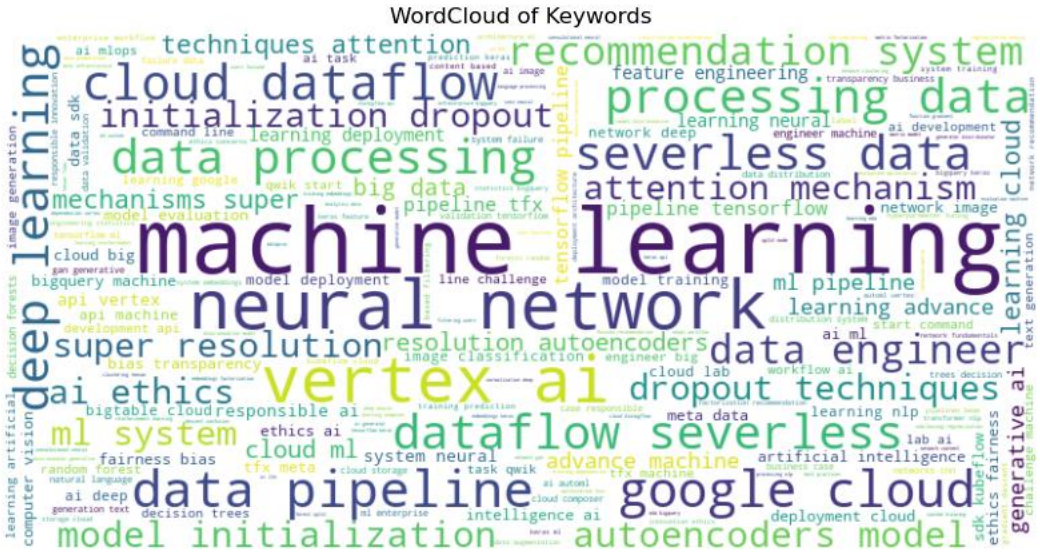

Figure 5 WordCloud of Keywords

Preprocessing

Effective text preprocessing is a crucial step in preparing the data for meaningful analysis. The Keywords_Tags_Skills_Interests_Categories column in the dataset underwent a series of preprocessing steps to transform raw text into a structured format suitable for analysis. The first step involved tokenization, where the text was split into individual words using a custom simple tokenizer. This tokenizer employed a regular expression to extract words while ensuring that punctuation and non-alphanumeric characters were excluded. Tokenization enabled the representation of each keyword as discrete units, forming the foundation for subsequent text processing.

After tokenization, stop-word removal was performed to eliminate common words that do not contribute significant meaning to the analysis. A predefined set of English stop words was used to filter out terms such as “and,” “the,” and “is,” which occur frequently but do not provide useful insights into the dataset's content. Removing these words reduced noise and improved the clarity of the processed text. Each tokenized and filtered string was then reassembled into a single space-separated text, which became the input for vectorization.

Vectorization Using TF-IDF

To numerically represent the processed text, the TF-IDF vectorization method was applied. TF-IDF assigns a weight to each word based on its frequency in a specific document relative to its frequency across all documents in the dataset. This method highlights terms that are particularly unique or relevant within individual records while minimizing the impact of commonly occurring words. The processed Keywords_Tags_Skills_Interests_Categories column served as the input for the TF-IDF vectorizer, resulting in a sparse matrix where rows corresponded to records and columns represented unique keywords.

The TF-IDF matrix provided a quantitative basis for analyzing patterns within the dataset. Each entry in the matrix reflected the importance of a specific word in a given record, enabling downstream analyses such as clustering and topic modeling. A sample of the generated matrix demonstrated the successful transformation of textual data into a numerical format. Additionally, the feature names extracted from the TF-IDF process revealed the most significant keywords in the dataset, highlighting terms central to data science, machine learning, and AI education. This preprocessing pipeline ensured that the text data was clean, structured, and ready for advanced analytical techniques.

Result and Discussion

Topic Modeling Results

The application of LDA to the dataset revealed five distinct topics, each characterized by unique sets of keywords. Topic 1 focused on terms related to data processing and cloud infrastructure, with keywords such as "data," "pipelines," "dataflow," and "cloud," reflecting its alignment with scalable machine learning systems. Topic 2 encompassed advanced neural network techniques, featuring keywords such as "neural," "recommendation," "autoencoders," and "attention," indicating a focus on deep learning and specialized models. Topic 3 emphasized machine learning pipelines and feature engineering, with terms such as "pipeline," "feature," "tfx," and "tensorflow" dominating the topic. Topic 4 was centered on machine learning applications in cloud environments, including "bigquery," "engineer," "nlp," and "tensorflow." Finally, Topic 5 addressed ethical considerations in AI, with keywords like "ethics," "mlops," "responsible," and "generative," highlighting its focus on responsible AI practices.

A table summarizing the most frequent keywords and example course titles for each topic provided a clearer understanding of the dataset's thematic structure. For instance, courses in Topic 1 related to data pipelines were strongly associated with terms like "processing" and "cloud," while Topic 2 emphasized advanced neural architectures with a focus on models like autoencoders and attention mechanisms. This analysis offered insights into the alignment between courses and key skill areas, enabling educators and learners to identify the most

relevant resources.

Clustering Results

The K-Means clustering algorithm grouped the courses into five distinct clusters based on their topic distributions. Cluster 0 primarily included foundational topics in machine learning, with entries such as "supervised learning" and "machine learning." Cluster 1 reflected a mix of introductory and conceptual topics, as evident from phrases like "problem statement" and "reducing loss." Cluster 2 focused on optimization and feature engineering, with keywords such as "learning rate" and "feature crosses." Cluster 3 highlighted fairness and bias in AI, demonstrated by entries like "fairness" and "identifying bias." Lastly, Cluster 4 contained topics on representation and toolkit development, with terms such as "tensorflow toolkit" and "representation."

A scatterplot of the clusters, based on the first two components of the LDA topic distribution, illustrated the separation between clusters. Observable trends suggested that beginner-friendly clusters, such as Cluster 0, were distinct from advanced technical clusters, such as Cluster 3. The balanced distribution of cluster sizes indicated a broad representation of skill levels and focus areas, with Cluster 1 being the largest and Cluster 4 the smallest. This clustering provided a nuanced view of course organization, aiding in curriculum design and resource recommendations tailored to specific learner needs.

The remaining topics, while less frequent, show relatively balanced distributions, with Topics 1, 2, 3, and 4 covering advanced neural network techniques, pipeline and engineering tools, cloud-specific applications, and AI ethics, respectively. The variation in topic frequency reflects the diverse nature of the dataset and aligns with the need to cater to learners with different expertise levels and interests.

Clusters are distributed across the two components, showing distinct groupings. Cluster 0 includes foundational topics and appears densely packed, reflecting similarity in courses focusing on basic machine learning concepts. Cluster 3, focusing on fairness and bias in AI, is more spread out, highlighting diversity in the types of content in this cluster. Clusters such as Cluster 2 and Cluster 4 display intermediate separation, likely representing specialized topics like feature engineering and advanced neural networks.

The figure showcases the ability of K-Means to partition courses effectively based on their topical similarity, providing an organized structure for understanding content groupings. This separation can aid in creating tailored learning paths or targeted recommendations for users with specific educational needs.

Skill Coverage and Educational Gaps

The topic modeling and clustering results reveal important insights into the distribution of skills across the Learning Path Index Dataset, highlighting both gaps and redundancies in the resources provided. Topics extracted through LDA show a strong focus on foundational data science and machine learning concepts, such as "data pipelines," "cloud processing," and "machine learning training" (Topic 1 and Topic 4). Advanced topics like neural network architectures and recommendation systems (Topic 2) and ethical considerations in AI (Topic 5) are also represented but are less prevalent. This imbalance indicates that while the dataset provides significant coverage for foundational

skills, more advanced and niche topics, such as ethical AI and optimization strategies, may be underrepresented. Additionally, topics like "fairness" and "bias" in Cluster 3 show limited scope, suggesting potential gaps in coverage for equity and inclusivity in AI training.

Another observed pattern is the accessibility of beginner resources compared to advanced ones. Clusters such as Cluster 0, which focuses on introductory machine learning concepts, and Cluster 1, with basic terminologies and loss functions, are significantly larger in size, as evidenced by their distributions (328 and 330 entries, respectively). These clusters dominate the dataset, indicating that beginner-level content is readily available and more accessible to learners. However, advanced resources, such as those focusing on representation learning and feature engineering in Cluster 4, are smaller in size (200 entries). This suggests a disparity in the depth and breadth of content available for advanced learners, potentially hindering their ability to access specialized resources.

Cluster Analysis and Resource Distribution

The K-Means clustering results further highlight key trends in resource characteristics, such as average course duration, difficulty levels, and free vs. paid content distribution. Cluster 0 predominantly contains short-duration beginner courses that are primarily free, emphasizing accessibility and foundational knowledge. Conversely, Cluster 4, which focuses on advanced concepts like "representation" and "feature engineering," tends to include longer-duration courses with a mix of free and paid options. This variation underscores the need for balanced resource availability to ensure that learners at different skill levels have equitable access to quality materials.

A summary of the clusters also reveals redundancies in foundational topics. Clusters 0 and 1 both focus on introductory machine learning concepts, creating overlap that might result in inefficient use of resources. Conversely, gaps are evident in areas like ethical AI (Cluster 3), which has fewer resources despite its growing importance in the field. Furthermore, the distribution of courses in Cluster 3 suggests a lack of diversity in content, as most entries focus narrowly on topics like "fairness" and "bias" without expanding into related areas such as algorithmic transparency or responsible deployment.

Implications and Recommendations

These findings suggest that while the dataset provides a solid foundation for learners at the beginner level, advanced learners may encounter challenges in accessing high-quality, specialized resources. The observed redundancies in introductory topics could be addressed by consolidating overlapping content and reallocating efforts toward underrepresented areas, such as ethical AI and advanced optimization techniques. Additionally, increasing the availability of free, high-quality advanced courses could bridge the accessibility gap and cater to learners seeking to deepen their expertise in specialized topics.

In conclusion, the skill coverage and clustering analysis highlight both strengths and weaknesses in the dataset's educational offerings. Addressing the identified gaps and redundancies through strategic content development and improved curation practices could enhance the dataset's overall effectiveness, ensuring that learners at all levels can access resources tailored to their needs. These improvements would contribute to the broader goal of fostering equitable access

to data science education.

Educational Implications

The findings from the topic modeling and clustering analysis offer valuable insights into improving the organization and accessibility of data science resources in open education. The identification of distinct topics, such as "data pipelines" and "neural networks" (Topic 1 and Topic 2), suggests a need to streamline resources around these themes to avoid redundancy and better align them with learners' needs. The clustering analysis revealed a dominance of beginner-level resources in Clusters 0 and 1, which primarily focus on foundational topics such as "machine learning basics" and "terminologies." While this is beneficial for newcomers, the lack of balance in advanced topics, particularly those in Cluster 4 (e.g., "representation learning" and "feature engineering"), highlights the need for expanded content to cater to more experienced learners seeking specialized skills.

To address these disparities, educators and platform developers can adopt a targeted approach by enhancing the depth and diversity of advanced-level resources. For example, creating comprehensive learning paths that build upon beginner courses and transition seamlessly into advanced topics such as "MLOps" (Topic 5) or "ethical AI" (Cluster 3) could bridge the existing gaps. Additionally, integrating practical projects and case studies into these courses can reinforce complex concepts, ensuring learners not only acquire theoretical knowledge but also gain real-world applicability. This structured progression would help maintain learner engagement across skill levels and foster a more holistic understanding of data science.

From an accessibility standpoint, the clustering results underline the importance of making advanced resources as widely available as their beginner counterparts. Clusters focusing on specialized topics, such as "TensorFlow engineering" and "responsible AI," are often underrepresented and, in many cases, locked behind paid access. This disparity limits opportunities for underprivileged learners to advance their expertise. To counter this issue, open educational platforms could consider offering a mix of free and paid advanced courses, with scholarship options or modular access to key topics. Furthermore, integrating interactive elements such as quizzes, simulations, and peer-based learning communities into these resources can enhance accessibility by promoting collaborative and self-paced learning.

Lastly, the implications for educators extend to curriculum design. The insights from clustering analysis can inform the development of modular and adaptive learning pathways, allowing instructors to guide students through a logical progression of topics. For instance, courses identified in Cluster 2, which focus on "optimization techniques" and "feature engineering," could be recommended as prerequisites for more advanced content in Cluster 4. Similarly, content addressing ethical considerations in AI (Cluster 3) should be incorporated early in the learning journey to instill critical thinking about responsible practices in technology. These adjustments can create a more inclusive and effective learning environment, supporting diverse learner goals and fostering broader participation in data science education.

Conclusion

This study explored skill patterns in open-access data science resources using

topic modeling with LDA and clustering with K-Means. The analysis identified five distinct topics, including foundational themes like "data pipelines" and "machine learning basics" as well as advanced areas such as "MLOps," "ethical AI," and "representation learning." While beginner-level resources were abundant, advanced topics were comparatively underrepresented, particularly in clusters addressing specialized domains. Additionally, resources focusing on emerging skills, such as "generative AI" and "feature engineering," were largely confined to paid offerings, highlighting accessibility disparities. These findings reveal significant educational gaps in the availability and distribution of advanced and accessible resources. The clustering results also emphasized overlaps and redundancies in some foundational skills while showcasing limited depth in advanced skills. Courses grouped under beginner-level clusters often focused on similar topics, potentially leading to learner saturation in basic concepts. In contrast, advanced learners faced challenges in finding comprehensive, freely accessible content for their specific needs. This imbalance underscores the importance of refining educational resources to ensure balanced skill coverage and equitable access.

The insights from this study have significant implications for promoting accessible, skill-aligned resources in data science education. Open educational platforms and course providers can leverage these findings to design more comprehensive learning paths that address both beginner and advanced learner needs. Providing modular access to advanced content, such as "MLOps" and "neural networks," alongside foundational courses can enhance skill progression. Additionally, blending free and paid content models, especially for advanced topics, could improve accessibility for learners from diverse socioeconomic backgrounds. Furthermore, resource providers can optimize their content by addressing redundancy in foundational skills and tailoring advanced resources to industry demands. This could involve integrating real-world projects and interactive elements, such as quizzes and collaborative activities, to enhance learner engagement. Focusing on underrepresented skills, such as "AI ethics" and "data representation," can align resources with critical industry trends, thereby equipping learners with comprehensive and future-ready competencies.

Future research could extend this analysis to validate the observed skill patterns across larger, more diverse datasets. Expanding the scope to include other fields, such as healthcare analytics or environmental data science, could uncover discipline-specific gaps and inform resource development in those areas. Additionally, incorporating learner feedback into topic modeling and clustering could provide a more nuanced understanding of how learners perceive and engage with these resources. Further investigation is also warranted to explore the impact of cultural and linguistic factors on the accessibility of educational resources. Analyzing global datasets could highlight region-specific challenges and guide the creation of localized resources that cater to learners' unique needs. Moreover, integrating advanced machine learning techniques, such as neural topic models, could improve the granularity of topic identification and provide deeper insights into skill patterns.

The study faced several limitations, including constraints in dataset size and scope, which may affect the generalizability of the findings. The analysis was limited to a predefined number of topics and clusters, potentially overlooking subtler themes or patterns within the dataset. Additionally, the reliance on text

data from course descriptions and keywords may not fully capture the instructional quality or depth of these resources. Another limitation was the focus on only one domain, data science, within open education. As open educational resources span numerous fields, findings from this study may not translate directly to other disciplines. Addressing these limitations in future research, such as by incorporating multimodal datasets or expanding to cross-disciplinary analyses, could enhance the robustness and applicability of the results.

Declarations

Author Contributions

Conceptualization: I.C.; Methodology: I.M.M.E.; Software: E.G.; Validation: I.M.M.E.; Formal Analysis: I.C.; Investigation: I.M.M.E.; Resources: I.C.; Data Curation: I.M.M.E.; Writing Original Draft Preparation: E.G.; Writing Review and Editing: I.C.; Visualization: I.M.M.E.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Verma, K. Lamsal, and P. Verma, “An Investigation of Skill Requirements in Artificial Intelligence and Machine Learning Job Advertisements,” *Ind. High. Educ.*, vol. 36, no. 1, pp. 63–73, 2021, doi: 10.1177/0950422221990990.

[2] S. Morandini, F. Fraboni, M. D. Angelis, G. Puzzo, D. Giusino, and L. Pietrantoni, “The Impact of Artificial Intelligence on Workers’ Skills: Upskilling and Reskilling in Organisations,” *Informing Sci. Int. J. Emerg. Transdiscipl.*, vol. 26, no. 2, pp. 039–068, 2023, doi: 10.28945/5078.

[3] X. Xie, K. Siau, and F. F. Nah, “COVID-19 Pandemic – Online Education in the New Normal and the Next Normal,” *J. Inf. Technol. Case Appl. Res.*, vol. 22, no. 3, pp. 175–187, 2020, doi: 10.1080/15228053.2020.1824884.

[4] J. Reich and J. A. Ruipérez-Valiente, “The MOOC Pivot,” *Science*, vol. 363, no. 6423, pp. 130–131, 2019, doi: 10.1126/science.aav7958.

[5] Y. Zhao and H. Liu, “Cloud Curriculum Resource Management Platform Based on Hadoop,” *Meas. Control*, vol. 53, no. 9–10, pp. 1782–1790, 2020, doi: 10.1177/0020294020948088.

[6] F. M. Evans *et al.*, “Evaluation of Open Access Websites for Anesthesia

Education," *Anesth. Analg.*, 2022, doi: 10.1213/ane.0000000000006183.
[7] J. Khadpe, "Emergency Medicine Procedures: A Review of Approved Instructional Resources From the World of Free Open Access Medical Education," *Cureus*, vol. 15, no. 9, pp. 1-8, 2023, doi: 10.7759/cureus.45986.
[8] R. Montoya and M. Soledad, "Challenges for Open Education With Educational Innovation: A Systematic Literature Review," *Sustainability*, vol. 12, no. 17, p. 7053, 2020, doi: 10.3390/su12177053.
[9] A. Thakran and R. C. Sharma, "Meeting the Challenges of Higher Education in India Through Open Educational Resources: Policies, Practices, and Implications," *Educ. Policy Anal. Arch.*, vol. 24, no. 3, p. 37, 2016, doi: 10.14507/epaa.24.1816.
[10] T. Nsengimana *et al.*, "Online Learning During COVID-19 Pandemic in Rwanda: Experience of Postgraduate Students on Language of Instruction, Mathematics and Science Education," *Contemp. Math. Sci. Educ.*, vol. 2, no. 1, p. ep21009, 2021, doi: 10.30935/conmaths/10788.
[11] E. K. Kurelović, "Open Access Culture and Acceptance of Open Educational Resources In Croatian Public Universities," *Zb. Veleuč. U Rijeci*, vol. 6, no. 1, pp. 39–50, 2018, doi: 10.31784/zvr.6.1.3.
[12] G. Ghosh, "Utilisation of Open Educational Resources by Academics and Students Among Nursing College in Siliguri Subdivision," *Rev. Rev. Index J. Multidiscip.*, vol. 3, no. 4, pp. 28–35, 2023, doi: 10.31305/rrijm2023.v03.n04.005.
[13] L. Joseph, S. Abraham, B. P. Mani, and N. Rajesh, "Exploring the Effectiveness of Learning Path Recommendation Based on Felder-Silverman Learning Style Model: A Learning Analytics Intervention Approach," *J. Educ. Comput. Res.*, vol. 60, no. 6, pp. 1464–1489, 2022, doi: 10.1177/07356331211057816.
[14] A. Castro, L. Garcia, D. N. Prata, M. Lisboa, and M. Prata, "An Exploratory Study on Data Mining in Education: Practiced Algorithms and Methods," *Int. J. Inf. Educ. Technol.*, vol. 7, no. 5, pp. 319–323, 2017, doi: 10.18178/ijiet.2017.7.5.888.
[15] S. B. G. Brambila and J. F. González, "Discovering Relationships Among Personal and Academic Factors With Academic Performance Using Association Rules," *Res. Comput. Sci.*, vol. 118, no. 1, pp. 9–17, 2016, doi: 10.13053/rcs-118-1-1.
[16] S. K. Wanjau, G. Okeyo, and R. Rimiru, "Data Mining Model for Predicting Student Enrolment in STEM Courses in Higher Education Institutions," *Int. J. Comput. Appl. Technol. Res.*, vol. 5, no. 11, pp. 698–704, 2016, doi: 10.7753/ijcatr0511.1004.
[17] E. Aboagye, J. A. Yawson, and K. N. Appiah, "COVID-19 and E-Learning: The Challenges of Students in Tertiary Institutions," *Soc. Educ. Res.*, vol. 2, no. 1, pp. 109–115, 2020, doi: 10.37256/ser.122020422.
[18] J. Zhou, "The Role of Libraries in Distance Learning During COVID-19," *Inf. Dev.*, vol. 38, no. 2, pp. 227–238, 2021, doi: 10.1177/02666669211001502.
[19] A. Ullah, M. Ashraf, S. Ashraf, and A. Shehta, "Challenges of Online Learning During the COVID-19 Pandemic Encountered by Students in Pakistan," *J. Pedagog. Sociol. Psychol.*, vol. 3, no. 1, pp. 36–44, 2021, doi: 10.33902/jpsp.2021167264.
[20] A. I. Muslimin and F. Harintama, "Online Learning During Pandemic: Students' Motivation, Challenges, and Alternatives," *Loquen Engl. Stud. J.*, vol. 13, no. 2, p. 60, 2020, doi: 10.32678/loquen.v13i2.3558.
[21] P. Paudel, "Online Education: Benefits, Challenges and Strategies During and After COVID-19 in Higher Education," *Int. J. Stud. Educ.*, vol. 3, no. 2, pp. 70–85, 2020, doi: 10.46328/ijonse.32.
[22] A. F. Akbar, H. B. Santoso, P. O. H. Putra, and S. B. Yudhoatmojo, "User Perception Analysis of Online Learning Platform 'Zenius' During the Coronavirus Pandemic Using Text Mining Techniques," *J. Sist. Inf.*, vol. 17, no. 2, pp. 33–47, 2021, doi: 10.21609/jsi.v17i2.1065.
[23] O. B. Adedoyin and E. Soykan, "Covid-19 Pandemic and Online Learning: The Challenges and Opportunities," *Interact. Learn. Environ.*, vol. 31, no. 2, pp. 863–875, 2020, doi: 10.1080/10494820.2020.1813180.

[24] M. Limniou, T. Varga-Atkins, C. Hands, and M. Elshamaa, “Learning, Student Digital Capabilities and Academic Performance Over the COVID-19 Pandemic,” *Educ. Sci.*, vol. 11, no. 7, p. 361, 2021, doi: 10.3390/educsci11070361.

[25] C. E. Schwartz, R. B. Stark, E. Biletch, and R. B. Stuart, “Comparing Human Coding to Two Natural Language Processing Algorithms in Aspirations of People Affected by Duchenne Muscular Dystrophy,” *J. Methods Meas. Soc. Sci.*, vol. 13, no. 1, pp. 15-40, 2022, doi: 10.2458/jmmss.5397.

[26] M. Inoue, H. Fukahori, M. Matsubara, N. Yoshinaga, and H. Tohira, “Latent Dirichlet Allocation Topic Modeling of Free-text Responses Exploring the Negative Impact of the Early COVID-19 Pandemic on Research in Nursing,” *Jpn. J. Nurs. Sci.*, vol. 20, no. 2, 2022, pp. 1-13, doi: 10.1111/jjns.12520.

[27] D. Fang, Y. Hong-yuan, B. Gao, and X. Li, “Discovering Research Topics From Library Electronic References Using Latent Dirichlet Allocation,” *Libr. Hi Tech*, vol. 36, no. 3, pp. 400–410, 2018, doi: 10.1108/lht-06-2017-0132.

[28] B. Yin and C.-H. Yuan, “Detecting Latent Topics and Trends in Blended Learning Using LDA Topic Modeling,” *Educ. Inf. Technol.*, vol. 27, no. 9, pp. 12689–12712, 2022, doi: 10.1007/s10639-022-11118-0.

[29] J.-W. Guo, “Using Twitter to Identify Patient Education Topics for Using Medical Marijuana for Cancer Pain Management,” *Coj Nurs. Healthc.*, vol. 5, no. 2, 2019, doi: 10.31031/cojnh.2019.05.000610.

[30] M. Huang, O. ElTayeby, M. Zolnoori, and L. Yao, “Public Opinions Toward Diseases: Infodemiological Study on News Media Data,” *J. Med. Internet Res.*, vol. 20, no. 5, p. e10047, 2018, doi: 10.2196/10047.

[31] A. Pérez-Encinas and J. Rodríguez-Pomeda, “International Students’ Perceptions of Their Needs When Going Abroad: Services on Demand,” *J. Stud. Int. Educ.*, vol. 22, no. 1, pp. 20–36, 2017, doi: 10.1177/1028315317724556.

[32] G. Krishnan, “Trends and Trajectories: Mapping the Evolution of Consumer Switching Intentions Through the Push-Pull Mooring Framework,” *Qubahan Acad. J.*, vol. 3, no. 4, pp. 457–468, 2023, doi: 10.58429/qaj.v3n4a230.

[33] M. Ghifari and W. T. H. Putri, “Clustering Courses Based on Student Grades Using K-Means Algorithm With Elbow Method for Centroid Determination,” *Inf. J. Ilm. Bid. Teknol. Inf. Dan Komun.*, vol. 8, no. 1, pp. 42–46, 2023, doi: 10.25139/inform.v8i1.4519.

[34] R. Davies, G. N. Allen, C. Albrecht, N. Bakir, and N. Ball, “Using Educational Data Mining to Identify and Analyze Student Learning Strategies in an Online Flipped Classroom,” *Educ. Sci.*, vol. 11, no. 11, p. 668, 2021, doi: 10.3390/educsci11110668.

[35] F. Biwer *et al.*, “Changes and Adaptations: How University Students Self-Regulate Their Online Learning During the COVID-19 Pandemic,” *Front. Psychol.*, vol. 12, 2021, doi: 10.3389/fpsyg.2021.642593.

[36] D. Kwasi and E. Gyimah, “Using K-Means to Determine Learner Typologies for Project-Based Learning: A Case Study of the University of Education, Winneba,” *Int. J. Comput. Appl.*, vol. 178, no. 43, pp. 29–34, 2019, doi: 10.5120/ijca2019919320.

[37] A. D. Maulana and I. G. Anugrah, “Implementation of the Simple Additive Weighting Method in Determining Centroids in the Process of Clustering the Poor in Kakatpenjalin Village, Lamongan Regency,” *J. Dev. Res.*, vol. 5, no. 2, pp. 85–93, 2021, doi: 10.28926/jdr.v5i2.147.

[38] Y. Guo, M. Zhang, C. J. Bonk, and Y. Li, “Chinese Faculty Members’ Open Educational Resources (OER) Usage Status and the Barriers to OER Development and Usage,” *Int. J. Emerg. Technol. Learn. Ijet*, vol. 10, no. 5, p. 59, 2015, doi: 10.3991/ijet.v10i5.4819.

[39] O. Malykhin, N. Aristova, N. Dichek, and N. Dyka, “Formation of Top Job Skills of Tomorrow Among Computer Engineering and Information Technologies Undergraduate Students in the Process of Learning English,” *Environ. Technol. Resour. Proc. Int. Sci. Pract. Conf.*, vol. 2, no. 6, pp. 249–254, 2021, doi: 10.17770/etr2021vol2.6642.

[40] K. Palkova, O. Agapova, and A. Zīle, “OER as a Tool for Sustainable Development: The Ukrainian - Latvian Experience of Forensic Science Experts,” *Eur. J. Sustain. Dev.*, vol. 10, no. 3, p. 15, 2021, doi: 10.14207/ejsd.2021.v10n3p15.

[41] H. Tang, “Implementing Open Educational Resources in Digital Education,” *Educ. Technol. Res. Dev.*, vol. 69, no. 1, pp. 389–392, 2020, doi: 10.1007/s11423-020-09879-x.