Unveiling Educational Archetypes of High Achievers: A K-Prototypes Clustering Analysis of Academic Pathways

Daniel Mashao[1], Ayorinde Olanipekun[2,*]

[1]Faculty of Engineering and the Built Environment, University of Johannesburg, South Africa

[2]Data Science Across Disciplines Research Group, Institute for the Future of Knowledge, Faculty of Engineering and the Built Environment, University of Johannesburg, Auckland Park, South Africa

ABSTRACT

While individual stories of successful people are often celebrated, a systematic, data-driven understanding of common educational pathways to high achievement remains underdeveloped. This is partly due to the analytical challenges of studying datasets with mixed numerical and categorical attributes. This study addresses this gap by identifying distinct educational archetypes from a diverse group of 108 high-achieving individuals using a computational approach. The research leverages the K-Prototypes clustering algorithm, a method specifically designed for mixed-attribute data, to analyze a dataset detailing the educational backgrounds of successful people. A comprehensive preprocessing pipeline was developed to clean, standardize, and transform features such as degree, field of study, university ranking, and GPA into a format suitable for clustering. The optimal number of clusters was determined using the Elbow method, balanced with a focus on the practical interpretability of the resulting groups. The analysis successfully identified four distinct and meaningful educational archetypes: (1) The Elite US STEM Achiever, characterized by advanced degrees from top-ranked American universities; (2) The Elite US Business & Law Professional, a similar high-prestige path focused on MBA and JD degrees; (3) The Global Entrepreneurial Path, a more internationally diverse route where institutional prestige and formal awards were less critical; and (4) The International STEM Scholar, defined by scholarship-funded education at a global range of institutions. The primary conclusion is that success is not predicated on a single educational model. The existence of these varied archetypes challenges monolithic definitions of prestigious education and provides a more nuanced understanding of the diverse foundations of high achievement. These findings have significant implications for the development of AI-driven educational guidance systems, which can be enhanced to provide more personalized and globally-aware recommendations.

Keywords AI in Education, Clustering, Educational Data Mining, K-Prototypes, Success Pathways

Submitted 7 July 2025
Accepted 3 August 2025
Published 1 September 2025

*Corresponding author
Ayorinde Olanipekun,
atolanipekun@uj.ac.za

Additional Information and Declarations can be found on page 269

DOI: 10.63913/ail.v1i3.35

Introduction

The increasing interest in the educational backgrounds of successful individuals has sparked a multidisciplinary exploration involving fields such as entrepreneurship, science, and the arts. Research indicates that understanding these trajectories could provide valuable insights into common educational paths that lead to success across diverse domains. Given the vast amounts of educational data available today, there is a significant opportunity for artificial intelligence (AI) to uncover hidden patterns within these datasets, allowing for a new level of analysis that can drive educational policies and practices [1] .

How to cite this article: D. Mashao and A. Olanipekun, "Unveiling Educational Archetypes of High Achievers: A K-Prototypes Clustering Analysis of Academic Pathways," *Artif. Intell. Learn.*, vol. 1, no. 3, pp. 258-270, 2025.

However, a notable gap exists in the systematic, data-driven understanding of educational trajectories across these fields. Current literature often lacks a comprehensive analysis that considers the mixed data types—both categorical and numerical—that typify educational records [2]. This complexity is compounded by challenges such as data quality and the biases that AI systems may inadvertently magnify, which risk perpetuating existing inequalities within educational landscapes [3], [4]. Thus, the examination of educational data through sophisticated AI methodologies warrants further investigation to adequately address these challenges and leverage the full potential of such technologies [5].

In particular, the application of AI in higher education holds promise for improving student learning experiences and academic planning [6]. For example, generative AI platforms can personalize learning and enhance engagement, ushering in novel pedagogical approaches and reshaping existing educational frameworks [7], [8]. Nevertheless, it is crucial that AI implementations are approached with caution, ensuring that ethical considerations guide these technologies' integration into primary educational systems [3], [4]. This call for responsible AI adoption is echoed by researchers who advocate for transparency and accountability within AI-driven educational practices, emphasizing the importance of stakeholder engagement in shaping effective policies [1], [3].

While the potential of AI to uncover patterns in educational trajectories is substantial, it is essential to systematically tackle the accompanying complexities and biases present in educational data. This examination can ultimately inform future educational practices and policies that are both equitable and effective in meeting the needs of diverse learner populations [5], [6].

The educational backgrounds of highly successful individuals have long been a subject of public fascination and academic inquiry. From tech entrepreneurs and Nobel laureates to world-renowned artists and political leaders, their formative academic years are often scrutinized for clues that might explain their subsequent achievements. In an era of increasing data availability, the potential to move beyond individual anecdotes and uncover systematic patterns in these educational trajectories has grown significantly. By leveraging modern computational techniques, it is now possible to analyze the complex interplay of academic choices and institutional environments that shape the pathways to success, offering a more holistic understanding of the foundations of high achievement across diverse fields.

Despite this potential, a systematic, data-driven understanding of common educational archetypes among high achievers remains less developed. While individual success stories are widely told, they often obscure the broader patterns that may exist. This study addresses a key gap by seeking to identify these underlying models through a quantitative lens. The primary research question is: What distinct educational archetypes can be identified among a diverse group of high achievers using K-Prototypes clustering? A secondary question follows: What are the defining characteristics—such as degree types, fields of study, institution tiers, GPA levels, and scholarship prevalence—of these identified archetypes?

To answer these questions, this study sets forth three clear objectives. The first is to preprocess and prepare the "Educational Backgrounds of Successful

People" dataset, a unique collection of academic histories that contains a challenging mix of numerical and categorical data. The second objective is to apply the K-Prototypes clustering algorithm, a method specifically suited for such mixed-attribute datasets, to identify distinct groups based on their shared educational attributes. The final objective is to thoroughly profile and interpret these resulting clusters, conceptualizing them as distinct "educational archetypes" that represent common pathways to success.

The significance of this research is threefold. First, it contributes directly to the field of AI in Learning by demonstrating how data-driven archetypes can inform the development of more personalized and nuanced educational guidance systems, moving beyond simplistic, one-size-fits-all recommendations. Second, the findings offer potential benefits for educators, career counselors, and aspiring individuals by providing a clearer understanding of the diverse models of success. Finally, this work makes a methodological contribution by applying the K-Prototypes algorithm to a unique socio-educational dataset, showcasing its utility in uncovering meaningful patterns in complex human-centric data.

Literature Review

Education and Achievement

A wealth of research links educational attainment to career success across various fields, particularly highlighting differences in pathways among entrepreneurs, Nobel laureates, and CEOs. Generally, higher educational levels are associated with enhanced leadership abilities and entrepreneurial success. For instance, studies have demonstrated a positive correlation between CEO education and corporate environmental performance, suggesting that more educated CEOs lead organizations that prioritize sustainable and responsible practices [9], [10]. Furthermore, it has been observed that educational background significantly influences managerial styles, further affecting corporate decision-making processes and overall firm performance [11], [12].

Particularly noteworthy is the analysis of high achievers from diverse backgrounds, who often experience non-linear career trajectories influenced by their educational experiences. Such findings underscore the importance of educational pathways in shaping innovative leadership and success narratives across domains [13]. For instance, within entrepreneurial circles, it has been established that educational background often intersects with socio-economic factors, leading to variations in career attainment and perceptions of success based on educational quality and accessibility [14], [15].

Clustering in Educational Data Mining (EDM) and Learning Analytics

Clustering methodologies in Educational Data Mining (EDM) and Learning Analytics have provided significant insights into student profiles and learning patterns. One noteworthy application of clustering is the identification of distinct learner archetypes based on their academic performances and behavioral data, which facilitates personalized learning experiences [16]. For instance, clustering techniques have been embraced to analyze mixed attribute data, effectively revealing typologies that enable educators to tailor instructional strategies and interventions [16].

Among the various clustering methods, K-Prototypes has emerged as a prominent algorithm for dealing with mixed data types, allowing for the effective categorization of both categorical and numerical data within educational

contexts [16]. This methodological advancement is particularly critical as it enables more nuanced analyses of educational datasets, leading to a deeper understanding of how diverse educational journeys affect collective and individual learning outcomes [17]. Additionally, these methodologies can highlight intersections between various factors influencing educational trajectories and subsequent career success, constructing a comprehensive landscape of academic and professional development paths [18].

Methodological Considerations for Mixed Attribute Data

Despite the advantages offered by clustering methodologies such as K-Prototypes, challenges remain in effectively analyzing mixed attribute data unique to educational contexts. Specifically, the complexity of reconciling categorical and numerical data often poses difficulties in obtaining coherent clusters that genuinely reflect underlying educational archetypes [19]. Furthermore, the interpretability of clusters remains a significant concern, as educators and researchers struggle to derive actionable insights from the resulting categorizations [20].

To address these challenges, it is essential to develop robust frameworks that prioritize the appropriate application of clustering algorithms tailored to specific educational datasets. The application of K-Prototypes, in this regard, stands out as a prospective solution, facilitating the extraction of holistic archetypes from high-achieving individuals' educational backgrounds—more importantly, highlighting the multifaceted influences of socio-economic factors and individual experiences on their educational trajectories [21], [22].

Identifying the Gap

While existing research has made strides in linking educational backgrounds to career outcomes, a distinctive gap exists in the application of advanced clustering techniques, such as K-Prototypes, to identify educational archetypes among high achievers encompassing diverse fields. Given that current methodologies often focus on homogeneous datasets, employing K-Prototypes on a unique dataset of high achievers can unravel complex intersections of educational experiences and career trajectories [15], [23]. Such analyses could ultimately contribute to a more profound understanding of the educational pathways that foster success, influencing academic policies aimed at nurturing diverse and innovative future leaders [24].

The inherent complexities of educational achievement necessitate a closer examination of educational paths utilizing robust methodologies. Advanced clustering techniques, particularly K-Prototypes, provide a pathway to exploring the rich tapestries of high achievers' educational backgrounds, ultimately illuminating the nuanced relationships between education, achievement, and career success.

Method

This study employed a quantitative approach using computational data analysis to identify distinct educational archetypes from a dataset of high-achieving individuals. The methodology was designed to be systematic and reproducible, encompassing several key stages: a detailed description and assessment of the source dataset, an extensive multi-step data preprocessing and feature

engineering pipeline, the application of the K-Prototypes clustering algorithm, a rigorous process for determining the optimal number of clusters, and finally, a systematic framework for cluster profiling and interpretation.

Dataset

The primary data for this research was sourced from the "Educational Backgrounds of Successful People" dataset, a publicly available collection detailing the academic histories of 108 notable individuals across a diverse range of professions, including technology, arts, science, and public service. The raw dataset contained multiple attributes for each individual, such as their profession, degree obtained, field of study, graduating institution, country of study, university global ranking, graduation year, GPA or equivalent academic honors, and information on scholarships or awards. An initial assessment revealed a heterogeneous data structure, characterized by a mix of numerical and categorical data types. This presented a significant analytical challenge, as standard clustering algorithms are often designed for homogenous data. Furthermore, the dataset exhibited common real-world data quality issues, including inconsistencies in formatting (e.g., varied representations for GPA and rankings) and a notable presence of missing values, particularly in the GPA and ranking columns. These characteristics necessitated a thorough and carefully designed preprocessing phase to ensure the data was clean, standardized, and suitable for the chosen clustering algorithm.

Data Preprocessing

To prepare the data for robust and meaningful cluster analysis, a multi-step preprocessing pipeline was executed. The initial step was feature selection, where a subset of the most relevant educational attributes was chosen to form the basis of the archetypes. The selected features included GPA (or Equivalent), Degree, Field, University Global Ranking, Country, and Scholarship/Award, as these were deemed most representative of an individual's core academic trajectory.

Subsequently, a comprehensive feature engineering and transformation process was undertaken to standardize these attributes and reduce noise. The Degree text field, which contained highly specific and varied entries, was consolidated into distinct, analytically useful categories such as "Bachelor's," "Masters," "PhD," "MBA," "MD," "JD," and "Dropout/Incomplete." This aggregation allowed for the grouping of similar qualifications to reveal broader patterns. Similarly, the Field of study was mapped into overarching disciplines like "STEM," "Business/Economics," "Humanities/Arts," "Social Sciences," "Law," and "Medicine/Health" to reduce dimensionality and facilitate higher-level comparisons. For the University Global Ranking, raw text entries (e.g., ranges like "10-20" or inequalities like "<50") were parsed to extract a single numerical rank, which was then categorized into meaningful ordinal tiers: "Top 50," "51-200," "201-500," ">500," and "Unknown/Not Ranked." This transformation from a noisy continuous variable to a structured ordinal one makes the feature more robust. The GPA (or Equivalent) feature underwent the most significant cleaning to harmonize its various formats; numerical values were standardized to a 4.0 scale where possible, and textual equivalents (e.g., "summa cum laude," "first class honours") were carefully mapped to corresponding numerical values based on common academic conventions. Finally, the Scholarship/Award feature was converted into a binary format (1 for yes, 0 for no) to simply indicate

the presence or absence of a recorded award, simplifying it for the clustering model.

To address missing values, a critical step for many machine learning algorithms, a specific strategy was employed for the numerical GPA feature. The median of the cleaned GPA column was used to impute any remaining null entries. The median was chosen over the mean as it is less sensitive to outliers, providing a more robust measure of central tendency for a potentially skewed distribution of academic achievement. This final step created a complete, analysis-ready feature set consisting of one numerical feature (Cleaned_GPA_Imputed) and five categorical features (Degree_Category, Field_Category, Ranking_Tier, Scholarship_Binary, Country).

K-Prototypes Clustering Algorithm

The K-Prototypes algorithm was strategically selected for this study due to its specific design to handle datasets containing a mix of numerical and categorical data types, which was a core characteristic of our preprocessed data. This algorithm uniquely combines the principles of K-Means, used for numerical data, and K-Modes, used for categorical data. It partitions data by minimizing a composite cost function. For numerical features (in this case, GPA), it calculates the squared Euclidean distance to the cluster centroid, rewarding proximity to the numerical center. For categorical features, it uses a simple matching dissimilarity metric (equivalent to the Hamming distance), which counts the number of mismatches between a data point's attributes and the cluster's "prototype" (the mode of each categorical feature). A key hyperparameter, gamma (γ), can be used to weight the relative importance of the numerical and categorical components in the cost function, although for this study, a default weighting was used. This integrated approach allows for the identification of coherent clusters based on both the quantitative measure of academic performance (GPA) and the qualitative aspects of an individual's educational path (degree, field, institution tier, etc.).

Determining the Optimal Number of Clusters (K)

To identify the most appropriate number of clusters (K) for the dataset, the Elbow method was employed as a primary heuristic. The K-Prototypes algorithm was iteratively run with K values ranging from 2 to 7. For each value of K, the total within-cluster dissimilarity, or "cost," was calculated and plotted. This cost represents the sum of distances of all data points to their respective cluster centers. The resulting plot was then visually inspected to find the "elbow point"—the point on the graph where the rate of decrease in cost sharply declines, suggesting that adding more clusters beyond this point yields diminishing returns and risks overfitting the data. However, the final value of K was chosen not solely based on this mathematical heuristic. Strong consideration was also given to the practical interpretability and distinctiveness of the resulting clusters. The goal was to find a K that produced clusters where each represented a meaningful, unique, and easily describable educational archetype. After evaluating the trade-offs, K=4 was selected as it provided the best balance between model parsimony and the richness of the resulting archetypes.

Cluster Profiling and Archetype Definition

Following the successful partitioning of the data into four distinct clusters, a

detailed profiling analysis was conducted to interpret and define the emergent archetypes. This process moved from quantitative analysis to qualitative interpretation. For each cluster, the distribution of its constituent features was systematically examined. The central tendency (mean and median) and spread (standard deviation) of the numerical GPA feature were calculated to understand the academic performance level of the group. Concurrently, the frequency distributions and modes of the categorical features (Degree_Category, Field_Category, Ranking_Tier, Scholarship_Binary, Country) were analyzed to identify the most common attributes. The key to defining the archetypes was to identify the features that were significantly over-represented or under-represented within a cluster compared to the overall dataset average. This quantitative profiling allowed for the construction of a distinct narrative for each cluster, highlighting its dominant and differentiating educational characteristics. These profiles were then used to assign descriptive, thematic archetype names that encapsulate the primary educational pathway represented by each group.

Software and Tools

The entire data analysis pipeline, from initial data loading to final visualization, was implemented using the Python programming language (version 3.x), ensuring a transparent and reproducible workflow. Data manipulation, cleaning, and feature engineering were performed primarily using the pandas library, which provides powerful and flexible data structures. The core clustering analysis was conducted with the kmodes library, a specialized package that offers an efficient and well-maintained implementation of the K-Prototypes algorithm. All visualizations, including the Elbow plot for determining K and the various bar charts and boxplots used for cluster profiling, were generated using the matplotlib and seaborn libraries, which offer a high degree of control for producing publication-quality graphics.

Result and Discussion

Dataset Overview and Preprocessing Outcome

The application of the K-Prototypes clustering algorithm to the preprocessed dataset of 108 high-achieving individuals successfully partitioned the data into four statistically distinct and qualitatively meaningful clusters. The selection of K=4 was guided by a heuristic analysis using the Elbow method, which indicated a point of diminishing returns in cost reduction, and was confirmed by a qualitative assessment of cluster interpretability, ensuring that each cluster represented a coherent and distinct narrative. The resulting clusters varied in size, with Cluster 0 being the largest (n=44), followed by Cluster 3 (n=24), Cluster 2 (n=21), and Cluster 1 (n=19). Detailed profiling of these clusters revealed four unique educational archetypes, each defined by a specific combination of academic achievement, degree level and type, field of study, institution prestige, and geographic context.

Cluster Profiles: The Educational Archetypes

Archetype 1 (Cluster 0): The Elite US STEM Achiever. This archetype, representing the largest cohort in the dataset, epitomizes a pathway of exceptional academic achievement within the most prestigious tier of US higher education. Individuals in this cluster exhibit the highest average GPA (mean 3.85), consistently performing at the top of their class. This academic excellence

is paired with an overwhelming tendency to have attended a "Top 50" global university (90.9%), institutions that function as epicenters of research and innovation. The educational focus is heavily concentrated in STEM fields (61.4%), suggesting a direct pipeline into technology, research, and engineering-driven industries. This path is not merely an undergraduate experience; a significant number pursue advanced degrees, including Master's (29.5%) and PhDs (13.6%), indicating a commitment to deep specialization. A defining characteristic is the high prevalence of scholarships or awards, with 90.9% of the group having received one, underscoring that their academic merit was formally recognized and financially supported. Geographically, this path is predominantly American, with 88.6% of individuals having studied in the USA, highlighting the central role of the US academic system in fostering this particular brand of elite talent.

Archetype 2 (Cluster 1): The Elite US Business & Law Professional. This archetype defines a highly specialized and equally prestigious route through elite professional programs, primarily in the United States, geared towards leadership in the corporate and legal sectors. The defining feature is the prevalence of advanced professional degrees, with a clear majority holding an MBA (57.9%) or a JD (10.5%). These degrees serve as powerful credentials and entry points into high-status professional networks. Correspondingly, the dominant fields of study are Business/Economics (68.4%) and Law (10.5%). Similar to the first archetype, this group is characterized by high academic achievement (mean GPA 3.67), attendance at "Top 50" institutions (78.9%), and a near-universal rate of receiving scholarships (94.7%). The slight difference in average GPA compared to the STEM group may reflect different grading curves or skill emphases in these programs. This pathway is also strongly US-centric (89.5%), reinforcing the notion of a distinct, American-style professional class forged in the nation's top business and law schools.

Archetype 3 (Cluster 2): The Global Entrepreneurial Path. This archetype represents a more internationally diverse and less institutionally-focused pathway, challenging traditional notions of elite education. The educational attainment is centered on foundational Bachelor's degrees (81.0%), predominantly in the practical field of Business/Economics (57.1%). A key distinguishing feature is the institutional background; a significant majority (66.7%) attended universities that were not highly ranked or whose rankings were unknown. This suggests that for this group, the specific brand of the institution was less important than the acquisition of core knowledge and skills. Furthermore, this group has the lowest rate of receiving scholarships, with 90.5% having no recorded award. This points towards a more self-reliant or self-funded educational journey, where success is not predicated on early academic accolades. This archetype is the most geographically diverse, with a wide distribution of individuals from the USA (38.1%), China (14.3%), India (9.5%), Germany (9.5%), and several other nations, reflecting the global nature of modern entrepreneurship.

Archetype 4 (Cluster 3): The International STEM Scholar. This final archetype captures a distinct and vital pathway for international scholars who achieve success primarily through deep specialization in STEM. While the most common degree is a Bachelor's (54.2%), this group also has a notable proportion of PhDs (12.5%), signaling a strong orientation towards research and academia. The field of study is overwhelmingly STEM (50.0%). Critically, this group is similar to

Archetype 3 in that most individuals (83.3%) attended institutions with unknown or lower global rankings. However, in stark contrast to the entrepreneurial path, this group has a very high rate of scholarship reception (83.3%). This juxtaposition is the core of the archetype's story: it represents talented individuals from around the world who leverage financial awards to access quality STEM education, regardless of the institution's global ranking. This archetype is defined by its international composition, with a large contingent from China (37.5%) and a broad mix of other countries, with a smaller US presence (20.8%). It exemplifies the global mobility of academic talent, driven by merit and enabled by financial support.

Discussion

The results of the clustering analysis successfully identified four distinct and interpretable educational archetypes, providing a data-driven narrative of the varied academic pathways pursued by high-achieving individuals. This section interprets these archetypes in the context of broader societal trends, discusses the critical theme of diverse success pathways, considers the tangible implications for the future of AI in education, and acknowledges the study's inherent limitations while proposing avenues for future research.

The four archetypes tell compellingly different stories about the journey to success, reflecting broader dynamics in the global economy and higher education. The Elite US STEM Achiever and The Elite US Business & Law Professional (Archetypes 1 and 2) represent traditional, high-prestige pathways that function within what can be described as a "gatekept ecosystem." These routes, characterized by exceptional academic performance at world-renowned institutions, align with the conventional narrative that success is forged in the crucible of elite higher education. These pathways appear to be predominantly accessible within the well-resourced US academic system, which acts as a powerful magnet for talent and a credentialing authority for entry into top-tier professions.

In stark contrast, The Global Entrepreneurial Path (Archetype 3) challenges this narrative of institutional gatekeeping. It suggests that for a significant cohort of successful individuals, the prestige of the institution and formal academic accolades are less critical than obtaining a foundational undergraduate education and then leveraging it through initiative and risk-taking. This pathway highlights a more self-reliant, internationally diverse route where success may be driven more by skills, networks, and opportunities seized outside the formal academic structure. It reflects the rise of a globalized startup culture where the value of an idea and its execution can outweigh the brand of one's alma mater.

Finally, The International STEM Scholar (Archetype 4) reveals another crucial narrative: the global pursuit of specialized knowledge. This archetype represents talented individuals who leverage scholarships as a key to unlock advanced education. Their journey underscores the role of financial support as a powerful engine for academic mobility and equity, suggesting that high-impact careers in STEM can be launched from a wide variety of institutions, provided the foundational training and support are present. This archetype speaks to the global competition for scientific talent and the use of scholarships as a tool for "brain circulation," benefiting both the individual and the host institution.

Diversity of Success Pathways

A central finding of this study is the empirical validation that there is no single, monolithic educational path to high achievement. The clear distinction between the four archetypes powerfully illustrates this diversity. The analysis reveals a spectrum of pathways, from the highly structured, resource-intensive routes centered in elite US universities to more varied, globally distributed routes where institutional prestige is a less dominant factor. This finding is significant because it challenges a one-size-fits-all definition of a "good education." It suggests that the currency of success varies by path: for some, it is the prestige of the degree; for others, it is the applicability of the skills; and for others still, it is the depth of the specialized knowledge acquired. This nuanced, evidence-based understanding is critical for educators, counselors, and aspiring individuals, as it moves beyond simplistic rankings and embraces a wider, more inclusive range of successful educational models.

Implications for AI in Learning and Education

The identified archetypes hold significant potential for enhancing AI-driven educational technologies, moving them beyond their current limitations. Many personalized guidance systems rely on simplistic metrics, often defaulting to recommending top-ranked universities or linear degree progressions. The archetypes from this study could inform the development of more sophisticated AI tutors and recommendation engines that operate with a narrative, contextual intelligence. For instance, an AI tool could present a student interested in entrepreneurship with "The Global Entrepreneurial Path," showing them a statistical profile of successful individuals who prioritized a business bachelor's degree over institutional rank. It could also illustrate "The International STEM Scholar" path to a student in a developing nation, highlighting the strategic importance of seeking scholarships. By incorporating these data-driven narratives, AI in learning can offer more holistic, personalized, and globally-aware guidance, presenting students with a portfolio of possibilities rather than a single, narrow ladder.

Strengths and Limitations

The primary strength of this study lies in its novel application of K-Prototypes clustering to a unique dataset, yielding clear and interpretable educational archetypes from complex, mixed-type data. However, several limitations must be acknowledged. The dataset size of 108 individuals is small and cannot be considered fully representative of all successful people globally. The results, particularly the prominence of US-centric archetypes, may reflect a sampling bias in the source data, which likely over-represents publicly visible figures from Western media. Furthermore, the preprocessing steps, such as categorizing fields of study, involve a degree of subjectivity; the boundaries between disciplines are fluid, and different categorization schemes could yield different results. A further limitation is the static nature of the data, which captures a final outcome but not the dynamic, often non-linear journey of an individual's education and career. Finally, it is crucial to remember that this analysis reveals correlations, not causation. These archetypes represent common patterns, not prescriptive formulas for success.

Directions for Future Research

This study opens several promising avenues for future research. The most immediate step would be to replicate this analysis on a larger, more comprehensive, and globally representative dataset—perhaps by leveraging

public APIs from professional networks like LinkedIn or academic databases like Google Scholar—to validate and potentially expand upon the identified archetypes. Future work could also incorporate a richer set of features, such as data on extracurricular activities, mentorship, socio-economic background, and early-career milestones, to build more detailed and predictive profiles. Finally, and perhaps most importantly, this quantitative analysis should be complemented with qualitative research. Conducting in-depth interviews with individuals who fit these archetypes could provide invaluable context, adding a layer of lived experience to the statistical profiles and helping to understand the motivations, challenges, and decisions that define these pathways.

Conclusion

This study successfully employed K-Prototypes clustering to analyze the educational backgrounds of high achievers, unveiling four distinct archetypes: the Elite US STEM Achiever, the Elite US Business & Law Professional, the Global Entrepreneurial Path, and the International STEM Scholar. The key finding is the clear, data-driven evidence of multiple, diverse pathways to success, challenging the notion of a single, elite-centric model of education. The analysis reveals a fascinating tension between traditional, institutionally-gatekept routes to success and more globalized, democratized pathways where individual merit and initiative play a more central role. These findings empirically validate the multifaceted nature of educational journeys and provide a structured framework for understanding them. Ultimately, this research contributes a novel, nuanced perspective on the relationship between education and high achievement. The identified archetypes offer a valuable framework for developing more sophisticated and context-aware AI-powered guidance systems that can present learners with a portfolio of possibilities rather than a single, narrow ladder. By moving beyond simplistic metrics and embracing the complexity of real-world trajectories, we can better support the next generation of innovators, leaders, and scholars. This study underscores the power of computational methods to uncover hidden patterns in socio-educational data and broadens our perspective on the many roads to a successful and impactful life.

Declarations

Author Contributions

Conceptualization: D.M.; Methodology: D.M.; Software: A.O.; Validation: A.O.; Formal Analysis: A.O.; Investigation: D.M.; Resources: A.O.; Data Curation: A.O.; Writing Original Draft Preparation: D.M.; Writing Review and Editing: A.O.; Visualization: D.M.; 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] E. Hannan and S. Liu, "AI: New Source of Competitiveness in Higher Education," *Compet. Rev. Int. Bus. J. Inc. J. Glob. Compet.*, 2021, doi: 10.1108/cr-03-2021-0045.

[2] M. Jia, H.-C. Xu, and S. Zhang, "Comparison Analysis of ARIMA and Machine Learning Methods for Predicting Trend of US Semiconductor Stocks," pp. 1607–1614, 2022, doi: 10.2991/978-94-6463-052-7_178.

[3] B. George and O. S. Wooden, "Managing the Strategic Transformation of Higher Education Through Artificial Intelligence," *Adm. Sci.*, 2023, doi: 10.3390/admsci13090196.

[4] N. Humble and P. Mozelius, "The Threat, Hype, and Promise of Artificial Intelligence in Education," *Discov. Artif. Intell.*, 2022, doi: 10.1007/s44163-022-00039-z.

[5] G. Yun, K. M. Lee, and H. H. Choi, "Empowering Student Learning Through Artificial Intelligence: A Bibliometric Analysis," *J. Educ. Comput. Res.*, vol. 62, no. 8, pp. 2042–2075, 2024, doi: 10.1177/07356331241278636.

[6] J. Guan, J. Zhang, and X. Zhang, "The Influence of Generative Artificial Intelligence on High School Students Academic Planning ChatGPT," *Lect. Notes Educ. Psychol. Public Media*, 2023, doi: 10.54254/2753-7048/24/20230705.

[7] S. Grassini, "Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings," *Educ. Sci.*, 2023, doi: 10.3390/educsci13070692.

[8] M. L. How and W. L. D. Hung, "Educing AI-thinking in science, technology, engineering, arts, and mathematics (STEAM) education," *Educ. Sci.*, vol. 9, no. 3, Sept. 2019, doi: 10.3390/educsci9030184.

[9] N. M. Tran and B.-N. T. Pham, "The Influence of CEO Characteristics on Corporate Environmental Performance of SMEs: Evidence From Vietnamese SMEs," *Manag. Sci. Lett.*, 2020, doi: 10.5267/j.msl.2020.1.013.

[10] M. D. Amore, M. Bennedsen, B. Larsen, and P. Rosenbaum, "CEO Education and Corporate Environmental Footprint," *J. Environ. Econ. Manag.*, 2019, doi: 10.1016/j.jeem.2019.02.001.

[11] S. Saidu, "CEO Characteristics and Firm Performance: Focus on Origin, Education and Ownership," *J. Glob. Entrep. Res.*, 2019, doi: 10.1186/s40497-019-0153-7.

[12] S. Silvina, R. Robin, and W. Yuwono, "The Impact on Firm Performance: Evidence From CEO Education," *Inovasi*, 2022, doi: 10.30872/jinv.v18i1.10477.

[13] V. Chotiyaputta and Y. J. Yoon, "Firm Performance by Thai CEOs in the SET100: Foreign or Locally Educated?," *Gatr J. Manag. Mark. Rev.*, 2016, doi: 10.35609/jmmr.2016.1.1(2).

[14] B. L. Perry, E. Martinez, E. W. Morris, T. Link, and C. G. Leukefeld, "Misalignment of Career and Educational Aspirations in Middle School: Differences Across Race, Ethnicity, and Socioeconomic Status," *Soc. Sci.*, 2016, doi: 10.3390/socsci5030035.

[15] M. Nieuwenhuis, A. S. R. Manstead, and M. J. Easterbrook, "Accounting for Unequal Access to Higher Education: The Role of Social Identity Factors," *Group Process. Intergroup Relat.*, 2019, doi: 10.1177/1368430219829824.

[16] M. Moore and J. Thaller, “Career Readiness: Preparing Social Work Students for Entry Into the Workforce,” *Front. Educ.*, 2023, doi: 10.3389/feduc.2023.1280581.
[17] O. Arek-Bawa, “Transitioning Doctoral Students to University Teachers a Case of an Online Teaching Development Programme,” 2023, doi: 10.29086/978-0-9869937-3-2/2023/aasbs14/8.
[18] O. Paixão and V. Gamboa, “Motivational Profiles and Career Decision Making of High School Students,” *Career Dev. Q.*, 2017, doi: 10.1002/cdq.12093.
[19] D. Spurk, A. Hirschi, and N. Dries, “Antecedents and Outcomes of Objective Versus Subjective Career Success: Competing Perspectives and Future Directions,” *J. Manag.*, 2018, doi: 10.1177/0149206318786563.
[20] B. Mintz, “Neoliberalism and the Crisis in Higher Education: The Cost of Ideology,” *Am. J. Econ. Sociol.*, 2021, doi: 10.1111/ajes.12370.
[21] S. Monteiro and L. S. Almeida, “Adaptation and Initial Validation of the Career Resources Questionnaire for Portuguese – HE Students Form,” *Análise Psicológica*, 2021, doi: 10.14417/ap.1841.
[22] A. Kosvyra, D. Filos, T. Cusack, and I. Chouvarda, “Identifying the PH.D. Students’ Needs for Career Enhancement Skills,” 2022, doi: 10.36315/2022v2end024.
[23] H. Han and J. W. Rojewski, “Economic Attainment Patterns of College-Educated Women in Mid-Career: An Objective Indicator of Career Success,” *J. Res. Tech. Careers*, 2017, doi: 10.9741/2578-2118.1019.
[24] W. Ghardallou, H. Borgi, and H. ALKHALIFAH, “CEO Characteristics and Firm Performance: A Study of Saudi Arabia Listed Firms,” *J. Asian Finance Econ. Bus.*, 2020, doi: 10.13106/jafeb.2020.vol7.no11.291.