Factors that Impact Success in Higher Education: Analysis of Educational Data

Diellor Hoxhaj[1], Jan Hric[1] and Iveta Mrázová[1]

[1]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Abstract

Higher education shapes society both at the individual and country-wide levels. This study examines the character of US four-year colleges in the context of the achieved graduation rates. Based on the data provided by the British Open University, we further investigate the precursors indicating student success/failure. Finally, we explore the impact of a higher education background on the structure of the parliament in the UK. Data mining techniques like clustering and decision trees adopted to analyze relevant educational data confirm a substantial impact of demographic factors and study behavior on academic success. Social network-based methods assist in revealing alumni connections in the UK parliament.

Keywords

educational data mining, higher education, graduation rates, clustering, decision trees, social network analysis

1. Introduction

Higher education (HE) shapes society both at the individual and country-wide levels. In 2023, the US Bureau of Labor Statistics reported a median income of individuals with a bachelor's degree two times higher than that of those with a high school diploma or equivalent. In addition, the friendships and acquaintances built during the formative years at university may aid in achieving personal goals and overcoming professional challenges. Graduation rates reflect the percentage of students who complete their degree within a specified timeframe. At Charles University, e.g., the BSc. graduation rates stand at 55% [28]. The case is further exacerbated by lower graduation rates in Science, Technology, Engineering, and Mathematics (STEM) areas, e.g., 35,6% at the Faculty of Mathematics and Physics of Charles University, or when accompanied by high tuition fees and costly student loans.

To pinpoint demographic factors likely to raise the chance for a successful graduation, this study first examines the character of US four-year colleges in the context of the achieved graduation rates. Based on the data provided by the British Open University, we further inspect the e-learning-based precursors indicating student success/failure. To address the possible impact of HE on society, we finally explore the UK parliament's university education-based social network structure. The following section surveys the research on HE success. Section 3 discusses the chosen data mining techniques. Section 4 debates the data used and its acquisition. Section 5 analyzes the obtained results. The concluding section summarizes the factors uncovered that affect graduation success and the role of HE in society.

2. Related Work

HE is tied to improved career opportunities, life, and even health. Student success is fundamental to the sector because of the enormous costs associated with HE. Early detection of at-risk students could call for timely actions improving students' success rates [3], [23]. In HE, success exists at different levels [26]. At the personal level, it may be acceptance into a university or securing stellar grades for students and the perceived success can substantially impact an individual's wellbeing, retention, and future career prospects [19]. For institutions, success may be measured by graduation or retention

ITAT'25: Information technologies – Applications and Theory, Slovakia

diellorhoxhaj@gmail.com (D. Hoxhaj); jh@mff.cuni.cz (J. Hric); iveta.mrazova@mff.cuni.cz (I. Mrázová)

0000-0002-3765-1400 (I. Mrázová)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2025-11-13

CEUR-WS.org/Vol-4092/paper7.pdf

rates, or international rankings. For a country, success in HE may translate into economic development, and globally, advanced skills and knowledge contribute to a civilized society, wealth, and stability.

In general, student success is defined by academic achievement demonstrated through strong grades and ultimately graduation, followed by advanced career prospects. Further, HE students are expected to develop critical thinking skills, the ability to solve complex problems, and refine effective communication [3], [23], [26]. In addition, a study involving prestigious Swedish HE programs noticed a new phenomenon of effortless achievement perceived by students as an indicator of their ability to juggle studies and extracurricular activities or future professional life [19]. Anyway, frequent short-term employment may compromise longer-term academic outcomes. Beyond HE, employment rates track graduates' professional success, while alumni engagement reflects connection to the institution.

The top two aspects predicting academic success include prior academic achievement and student demographics like gender, age, race, and socioeconomic status (SES), as students who are not burdened by financial worries can better focus on their studies [3]. Further factors comprise students' environment, psychological attributes like motivation, study behavior, and integrated student e-learning activities. In general, feeling safe and secure together with a strong sense of belonging to the campus community boosts students' wellbeing and increases retention [25], [26]. On the other hand, a lack of diversity might negatively affect students' sense of belonging and inhibit their academic achievement. Surprisingly, a longitudinal study [17] found that Hungarian students do not benefit from higher-level twenty-first-century skills like critical thinking or inductive reasoning. Moreover, good problem-solvers had higher chances of dropping out than graduating. Emerging tools such as AI and ChatGPT were found to enhance learning performance [23].

3. Background

Educational databases enabled the emergence of the Educational Data Mining (EDM) research field. The iterative EDM process comprises six stages: data collection, initial data preparation, statistical analysis, data preprocessing, data mining implementation, and result evaluation [18]. Lately, machine learning techniques have been extensively used for predictive purposes. For a preliminary data analysis, clustering techniques provide interpretability without employing costly labeled data. Clustering methods partition the data into subsets of mutually similar data, dissimilar to data grouped in other clusters. Decision trees induce a hierarchical sequence of decisions organized in a tree-like model to facilitate explainable data classification.

Social Network Analysis (SNA) studies interconnection patterns between individuals within a larger system, e.g., an education-based one. At ETH Zurich, SNA was used to analyze the factors explaining academic failure and success of engineering undergraduates [22]. In critical examination periods, functional studying relationships strongly impact students' success. Socially isolated students, on the other hand, tend to score remarkably worse and are more likely to drop out of university regardless of their SES and cognitive abilities.

3.1. Clustering

Let $k < p$ for the number of clusters k and the number of data patterns p. The patterns assigned to the same cluster are considered to be mutually similar, whereas the patterns from different clusters are regarded as mutually dissimilar [9]. In the case of numerical data, the goal of clustering sounds to find the best partition of a finite set of patterns $X \subset R^d$ into subsets $C_1, \ldots, C_k$ called clusters (and represented by the centroids $\vec{c}_1, \ldots, \vec{c}_k$) such that the value of the applied objective function, e.g.,

$$O = \sum_{j=1}^{k} \sum_{\vec{x}_i \in C_j} \|\vec{x}_i - \vec{c}_j\|^2 \; ; \; \vec{x}_i \in X, 1 \leq i \leq p \tag{1}$$

is optimized. The k-means clustering algorithm [15] and Kohonen Self-Organizing Feature Maps [13] belong to popular techniques used for this purpose.

3.1.1. k-means Clustering

Initially, we choose the desired number of clusters k, and the algorithm randomly assigns the data patterns to the clusters. Initial cluster centroids correspond to the mean of all data patterns from the same cluster. Afterwards, the k-means clustering algorithm [15] iteratively reassigns the patterns to clusters to minimize the objective function (1). As the centroids might not actually be present within the analyzed data, median values help to interpret the found cluster characteristics.

3.1.2. Kohonen Self-Organizing Feature Maps

Also, Kohonen Self-Organizing Feature Maps (SOMs) [13] can be used for preliminary analyses and visualizations of high-dimensional data. SOMs map high-dimensional data onto the (output) neurons arranged on a 2D topological grid that usually preserves the topography of the data in the original space. Given an input pattern $\vec{x}$, SOM finds the neuron with the closest weight vector $\vec{w}_i$. This neuron is called the *winner*. During training, the weights $\vec{w}_j$ of the winner and its neighbors on the grid are updated at time t according to:

$$\vec{w}_j(t+1) = \vec{w}_j(t) + \alpha(t) \cdot h_{ij}(t) \cdot (\vec{x} - \vec{w}_j(t)) \tag{2}$$

$\alpha(t) \in (0,1)$ denotes the learning rates decreasing in time and $h_{ij}(t)$ is the lateral interaction function value, e.g., of the Mexican hat form, at time t.

3.1.3. Silhouette Score

Silhouette score helps estimate the adequate number of clusters. The quality of the underlying clustering is assessed by comparing the similarity $d(i,j)$ of patterns i and j from the same clusters to the (dis)similarity of patterns from different clusters [21]. For each pattern i from cluster C_I, $|C_I| > 1$, we evaluate its average similarity to all other patterns $j \in C_I$:

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I,\, j \neq i} d(i,j). \tag{3}$$

Then, the minimum average (dis)similarity of i to all patterns $j \in C_J$; $J \neq I$ will be determined as:

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j). \tag{4}$$

For $C_I > 1$, the silhouette score $s(i)$ corresponds to:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}. \tag{5}$$

For $|C_I| = 1$, $s(i)$ is defined to be 0. The mean over all $s(i)$ determines the overall quality of the considered clustering. Its values range from -1 to 1, with higher values indicating a better clustering.

3.2. Decision Trees

Decision trees are built in a greedy manner, starting at the root and choosing the most informative attribute j at each node to split the data D in a more class-uniform fashion. The C4.5 [20] and CART [7] algorithms belong to the most common techniques. C4.5 uses the entropy-based information gain ratio gainRT(D, j) for data splitting:

$$\text{gainRT}(D,j) = \frac{\text{entropy}(D) - \text{entropy}_j(D)}{-\sum_{i=1}^{v_j} \left(\frac{|D_i|}{|D|} \cdot \log_2 \frac{|D_i|}{|D|} \right)} \tag{6}$$

with $\text{entropy}_j(D)$ for attribute j having v_j different values and $|D_i|$ standing for the number of patterns from D with value i of attribute j:

$$\text{entropy}_j(D) = -\sum_{i=1}^{v_j} \left(\frac{|D_i|}{|D|} \cdot \text{entropy}(D_i) \right). \tag{7}$$

For the number of patterns from class c in D, $|D^c|$:

$$\text{entropy}(D) = -\sum_{c=1}^{C} \frac{|D^c|}{|D|} \log_2 \frac{|D^c|}{|D|}. \tag{8}$$

3.2.1. CART algorithm

The CART algorithm [7] can handle both categorical and numerical attributes. During training, it raises a binary tree by iteratively splitting the data received at each node into two subsets based on the selected attribute and its value. CART stops when all the present data points belong to the same class.

Let us consider a dataset D with p training patterns $\{(\vec{x}_i, y_i); 1 \leq i \leq p\}$. $\vec{x}_i$ is the vector of attribute values and y_i is its class label indicating one of C possible classes. The dataset D is split into D_l and D_r by selecting attribute j^* and its splitting value v_j^* to minimize

$$\text{Gini}(D, j, v) = \frac{|D_l|}{|D|} \text{Gini}(D_l) + \frac{|D_r|}{|D|} \text{Gini}(D_r). \tag{9}$$

Gini index for dataset D is then calculated as:

$$\text{Gini}(D) = 1 - \sum_{c=1}^{C} \left(\frac{|D^c|}{|D|} \right)^2 \tag{10}$$

where D^c comprises the patterns from D belonging to class c. To avoid deep and overfitted trees, maximum depth, number of nodes, or minimum number of data in a node can be constrained [27].

The so-called cost-complexity post-pruning might follow. This procedure systematically removes entire subtrees and replaces them with leaf nodes with previously found class labels. We will denote the error rates of tree T over the dataset D by $E(T, D)$. Further, let $\text{prune}(T, t)$ define the tree obtained by pruning subtree t from T. The subtree t^* that minimizes cost-complexity α,

$$\alpha = \frac{E(\text{prune}(T, t), D) - E(T, D)}{|\text{leaves}(T)| - |\text{leaves}(\text{prune}(T, t))|} \tag{11}$$

is then chosen for removal. The validation set is used to evaluate the error rates of the pruned trees.

3.2.2. Ensemble learning

Decision tree classifiers can be combined through ensemble learning to improve the overall performance for imbalanced data. Popular generalizations include adaptive boosting (AdaBoost) to lower the overall bias and Random Forests to reduce the variance of averaged predictions [1]. Alternatively, metrics like precision, recall, and F1-score can be used to reflect the actual model performance better.

AdaBoost [5], [10] assigns a weight to each pattern based on its difficulty for classification. At each iteration, an additional classifier is built with the weights updated according to the result of the previous classification. The final classification is determined as a weighted output of all previous classifiers, giving a higher weight to the more accurate ones. Random forests [8] inject more variety to the trees that may be used in parallel by randomly limiting attribute choices the trees can make at each node. Trained trees do not have to be pruned; the majority vote determines the final classification output.

3.3. Social Network Analysis

A *social network* is given by the graph (V, E). Its vertices represent actors and its edges correspond to the relationships between the actors. The importance of an actor u can be assessed by means of *centrality measures* [4].

- The *degree centrality* of actor u is defined as $C_D(u) = \mathrm{dg}(u)/(|V| - 1)$. High $C_D(u)$ values characterize influential actors with a direct relationship to others.
- The *closeness centrality* measures the efficiency of accessing other actors using the inverse of the average shortest path distance from u to all other actors v, $Dist(u, v)$:

$$C_C(u) = \sum_{v=1}^{|V|} Dist(u, v)/(|V| - 1), \tag{12}$$

- Let $\sigma_{u,w}$ be the total number of shortest paths between the actors u and w. Some of these paths go through actor v, let their number be $\sigma^v_{u,w}$. The *betweenness centrality* of v is the sum of the ratios $\sigma^v_{u,w}/\sigma_{u,w}$ for all available pairs of u and w normalized over $(|V| - 1)(|V| - 2)/2$:

$$C_B(v) = \frac{2}{(|V| - 1)(|V| - 2)} \sum_{u<w} \frac{\sigma^v_{u,w}}{\sigma_{u,w}}. \tag{13}$$

  Actors with a high $C_B(v)$ exhibit more control over the network by interconnecting its parts.

- The *eigenvector centrality* of actor v reflects the importance of all its neighbors. With the biggest eigenvalue of the adjacency matrix denoted as λ :

$$C_E(v) = \lambda^{-1} \sum_{u \in \mathrm{ne}(v)} C_E(u)\,. \tag{14}$$

3.3.1. Community Detection

Community detection identifies the subsets of actors that are more densely tied together than to the rest of the network. The so-called *network modularity* measures the concentration of edges within communities compared to the expected concentration of randomly distributed edges. High modularity values indicate optimum partitioning.

Definition 1 (Network Modularity). *Let $G = (V, E)$ be a graph and $c : V \to \mathbb{N}$ be a community assigning function. The network modularity Q^G_c of G with the communities given by c is defined as*

$$Q^G_c = \frac{1}{2|E|} \sum_{u,v \in V} \left(\mathbf{A}_{u,v} - \frac{\mathrm{dg}(u) \cdot \mathrm{dg}(v)}{2|E|} \right) \delta_{c(u),c(v)}\,. \tag{15}$$

δ denotes Kronecker delta ($\delta_{x,y} = 1$ if $x = y$, otherwise $\delta_{x,y} = 0$), $\mathbf{A}_{u,v}$ is the (u, v) entry in the adjacency matrix, i.e., the number of edges between u and v.

Various methods exist for community detection, e.g., the *Girvan-Newman algorithm* [11], the *Kernighan-Lin algorithm* [12], or the *Leiden algorithm* [24]. Further, we will use the *Louvain method* [6] capable of quickly detecting communities of varying sizes.

4. Data Acquisition

4.1. Data on US Four-Year Colleges

The dataset on graduation data of US four-year colleges was acquired from the National Center for Educational Statistics https://nces.ed.gov/. We gained the required university IDs from https:

//secondnature.org/wp-content/uploads/Workbookv7-Sheet1.pdf. Along with the numbers of enrolled students and graduation rates on 7 different student ethnic groups ('Asian', 'Black/African American', 'Hispanic/Latino', 'Race and Ethnicity Unknown', 'Two or More Races', 'US-Nonresidents', and 'White'), the total number of men, women, and tuition fees was obtained there.

We enhanced the dataset with the county hosting the respective university's median household income (MHI). We scraped this information from https://www.countyhealthrankings.org. To avoid single-year fluctuations, the data was collected over three years, 2020-2022, resulting in 3,975 data patterns, each comprising 11 numeric attributes.

4.2. UK Data on Student Success/Failure

The statistics we will use to analyze study behavior linked to students' success/failure stems from the UK Open University (OU). OU courses are taught mostly online (off-campus) [14]. The Open University Learning Analytics Dataset (OULAD) contains evidence about students' behavior while studying and is available at https://analyse.kmi.open.ac.uk/open_dataset. The file includes information on seven courses involving 32,593 students from 2013 and 2014. The collected data combines students' personal details along with their exam scores, and records how they interacted with the Virtual Learning Environment (VLE), given by the summaries of their clicks. After preprocessing, the available data comprised 35 attributes over 21663 patterns.

4.3. Data on the UK Parliament Structure

The UK Parliament consists of the House of Lords (787 appointed members as of April 2024) and the House of Commons with 650 Members of Parliament (MPs). MPs are elected in general elections, and the party or coalition with the majority in the House of Commons forms the government. Usually, the leader of that party or coalition becomes the Prime Minister. In this study, we will analyze the House of Commons structure based on its members' previous education.

For the data analysis, the information on the MPs, such as Name, Alma Mater, and Party membership, was scraped from the site: https://www.parallelparliament.co.uk/MPs in April 2024. We used the MPs' Wikipedia pages to find information on the universities they attended. We enhanced the data by adding additional attributes ('Party Membership', 'Age', and 'Gender'. Afterwards, we cleaned the data to avoid possible inconsistencies. In the formed social network, a vertex represents each MP. An edge will connect any two MPs who graduated from the same university.

5. Supporting Experiments

5.1. US Graduation Rates Analysis

This experiment aimed to understand the character of four-year HE institutions from the perspective of the achieved graduation rates. For the initial data analysis, a SOM with a hexagonal mesh of 35 x 35 neurons was trained for 500000 iterations using the scraped data on US four-year colleges. Afterwards, the SOM's weight vectors were clustered by the k-means algorithm. Silhouette score indicated two viable choices for k, namely $k = 2$ and $k = 6$, with the scores for both slightly below 0.36. We grouped the universities into six distinctive clusters to opt for higher variability.

Cluster ID 0 comprises high-performing institutions with reasonably many students diversified across various ethnicities (median student population size of cca 720). Their tuition fees are the highest (median of cca \$55,000 a year). The median household income of the counties hosting these universities is also high (cca \$74,000). Institutions from this cluster include Harvard, Princeton, and Yale universities, which consistently exhibit high median graduation rates across all ethnic groups (over 80%). The institutions from cluster ID 4 are notably more affordable (with median tuition fees of cca \$24,000), administer significantly more students (median student population size of cca 3400), yet still attain above-average

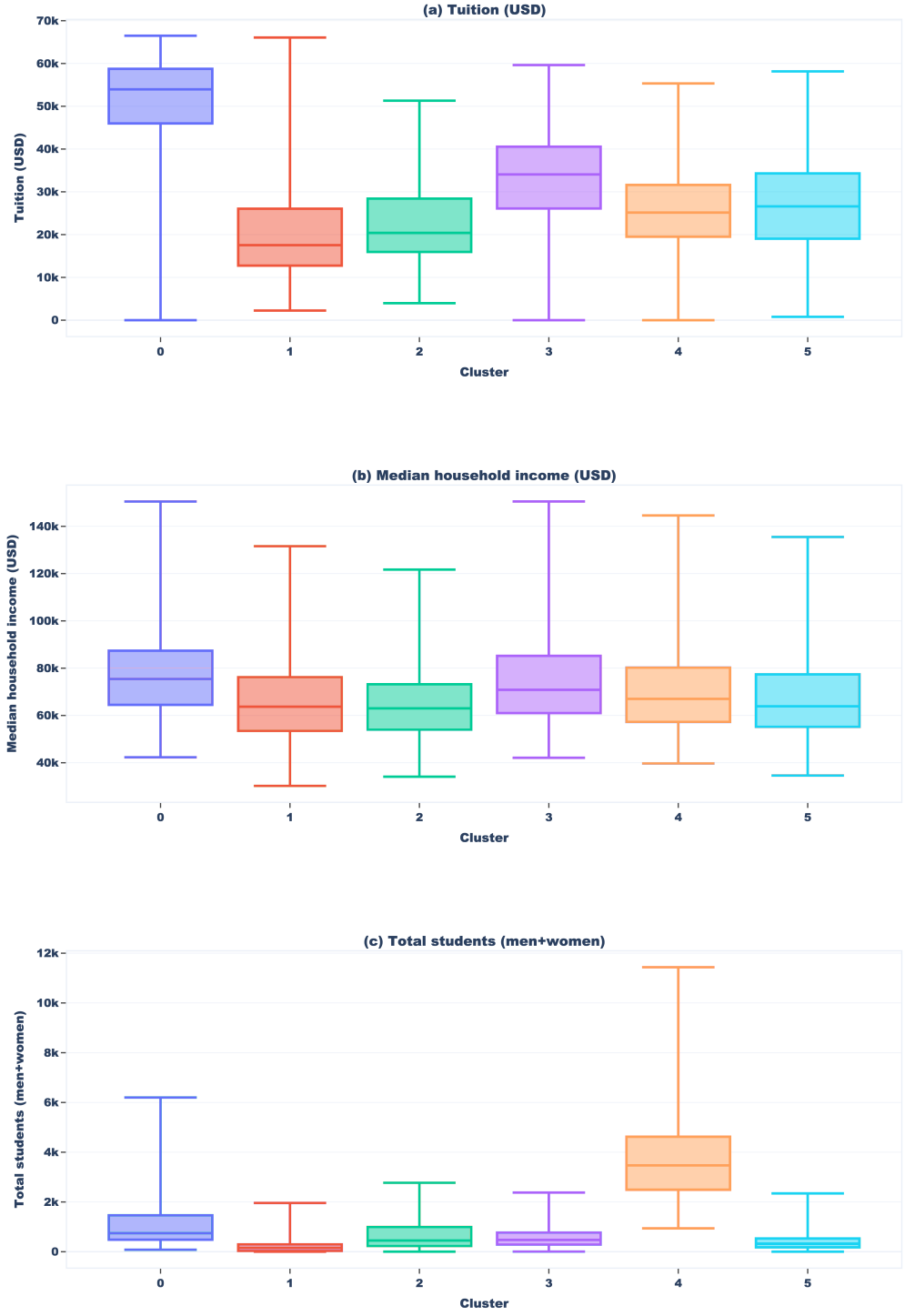


Figure 1: Distribution of tuition fees (a), median household income values (b), and student population size (c) across the found clusters.

graduation rates (between 55% and 70%) across the respective ethnicity groups. However, the hosting counties' median household income is lower (cca $65,000) when compared to cluster ID 0.

Colleges from cluster ID 1 do not demand high tuition fees (median at roughly $17,000) but achieve only low graduation rates (between 15% and 40% across the ethnicities). While they serve a relatively small student body (median of around 150), 'Asian', 'US-Nonresidents', and 'Two or More Races' ethnicity representatives do not seem to prefer enrolling in these universities. The median household

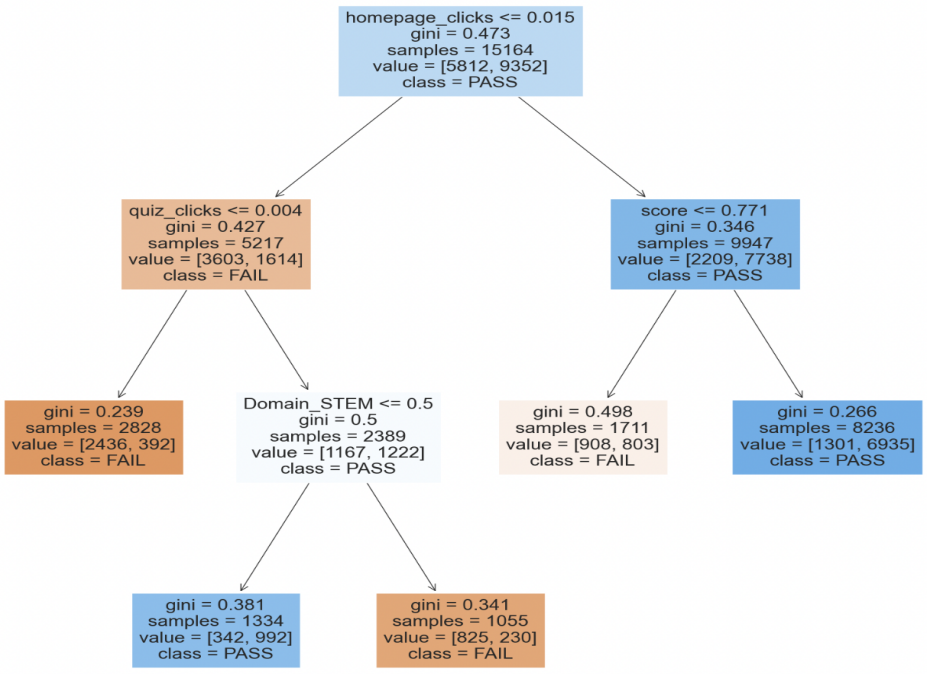


Figure 2: A pruned CART decision tree for the OULAD data.

income of the hosting counties is comparable to cluster ID 4 (cca $64,000). Clusters ID 2, 3, and 5 comprise HE institutions with a more petite student body and varying graduation rates across ethnicity groups (moderate to low-performance for cluster ID 2; on par with cluster ID 4 for cluster ID 3; and above average graduation rates for 'Asian' and 'White' for cluster ID 5, with 'US-Nonresidents' rarely attending). Median county household incomes and tuition fees also differ.

5.2. Prediction of student success/failure in a virtual learning environment

While many papers analyzed the OULAD dataset, the produced models still lack actionable interpretability. Therefore, this experiment focused on finding an accurate decision tree-based model for students' success/failure assessment. Random Forests performed the best with an accuracy 0.86 and an F1-score of 0.86. A recent paper [2], reports comparable classification results on the same dataset (accuracy of 0.83 for the CART algorithm vs. 0.86 for Random Forests) - see Table 1.

To facilitate explainability, we also provide a pruned CART decision tree 2 elucidating the attributes that most contribute to the student's success. Based on our findings, essential attributes to predict student success comprise the clicks recorded on the homepage of VLE, the number of times students interacted with the quizzes in VLE, and the domain of the course - STEM or social sciences, where students tend to succeed more, and their scores on assessments throughout the semester.

We also tested the option to predict students' success or failure using only demographic features comprising the deprivation index (IMD), age, disability, studied credits, highest education, and final result (PASS or FAIL). However, the formed trees performed much worse - see Table 1. The accuracy of the best C4.5 algorithm reached just 63 % indicating a deficiency of sole demographic features for accurate performance predictions. Out of the demographic features, the highest level of students' education best indicated their success or failure.

5.3. UK Parliament Analysis

In this experiment, our objective was to identify the most influential members in the parliament and to detect communities of MPs based on their educational background. Each MP could have attended

Table 1
Test performance of the considered decision tree classifiers induced over all 35 attributes. The values obtained for decision trees induced over six demographic attributes are in italics.

Method	F1-Score	Precision	Recall	Accuracy
Random Forest	**0.86** / *0.58*	**0.86** / *0.59*	**0.86** / *0.58*	**0.86** / *0.58*
AdaBoost	0.84 / ***0.59***	0.84 / *0.59*	0.84 / *0.58*	0.84 / *0.58*
C4.5	0.83 / ***0.59***	0.83 / ***0.61***	0.83 / ***0.63***	0.83 / ***0.63***
CART (No Pruning)	0.80 / *0.58*	0.80 / *0.59*	0.80 / *0.58*	0.80 / *0.58*
CART (Post-Pruning)	0.79 / *0.58*	0.78 / *0.59*	0.79 / *0.58*	0.79 / *0.58*

Table 2
Centralities computed for selected UK MPs and average centrality values over all MPs (before the 2024 election).

	Values of the respective centralities			
Name of the MP	Degree	Betweenness	Closeness	Eigenvector
John Glen	**0.30**	0.036	**0.43**	**0.11**
Matt Hancock	**0.30**	0.036	**0.43**	**0.11**
Tanmanjeet Dhesi	0.29	0.029	0.42	**0.11**
Ed Davey	0.23	**0.066**	**0.43**	0.10
Keir Starmer	0.20	**0.065**	0.41	0.10
Average (all MPs)	0.06	0.003	0.27	0.02

multiple universities. Each vertex in the graph thus represents an MP and the vertices are interconnected by an edge if both MPs participated at the same university – see Figure 3.

5.3.1. UK Parliament - the Graph Structure

Table 2 lists the centralities of 5 MPs selected for their consistently high scores. MPs with high centrality values typically graduated from two or more universities shared by other MPs, are well-connected to other influential MPs, and play a central role in the inspected network.

John Glen interlinks the graduates of Oxford and Cambridge. He also studied at King's College London and was named Paymaster General and Minister for the Cabinet Office in 2023. Matt Hancock and Tanmanjeet Singh Dhesi also interconnect the graduates from Oxford with those from Cambridge, see Figure 3. Matt Hancock served as the Secretary of State for Health and Social Care from 2018 to 2021. Tanmanjeet Singh Dhesi further studied at University College London. Ed Davey and Keir Starmer exhibit high betweenness centrality. Keir Starmer graduated from the University of Leeds and from Oxford, thus serving as a bridge in the MP social network. Keir Starmer became the leader of the Labour Party in 2020 and won last year's election. In 2024, he became Prime Minister of the UK.

5.3.2. UK Parliament - Community Detection

We used the Louvain algorithm for community detection in the UK MPs' graph, which was investigated. Twenty-nine communities were found, yet for further analysis, we considered only those communities with more than 15 MPs. Figure 4 lists those 7 communities comprising a total 548 MPs (based on the data from April 2024). The MPs most frequently attend Oxford and Cambridge universities. Explicitly, communities 1 and 2 emerged around these universities. They are predominantly male (around 70%) and approximately 70% of their MPs belong to the Conservative Party.

MPs from communities 4 and 7 are younger than those from other communities. Communities 6 and 7 appear more gender-equal, and communities 3, 4, 5, and 6 include more members from the Labour Party than the first two communities (significantly more than 30%). Scottish National Party members dominate community 7; they mostly attended the University of Glasgow and the University of Stirling,

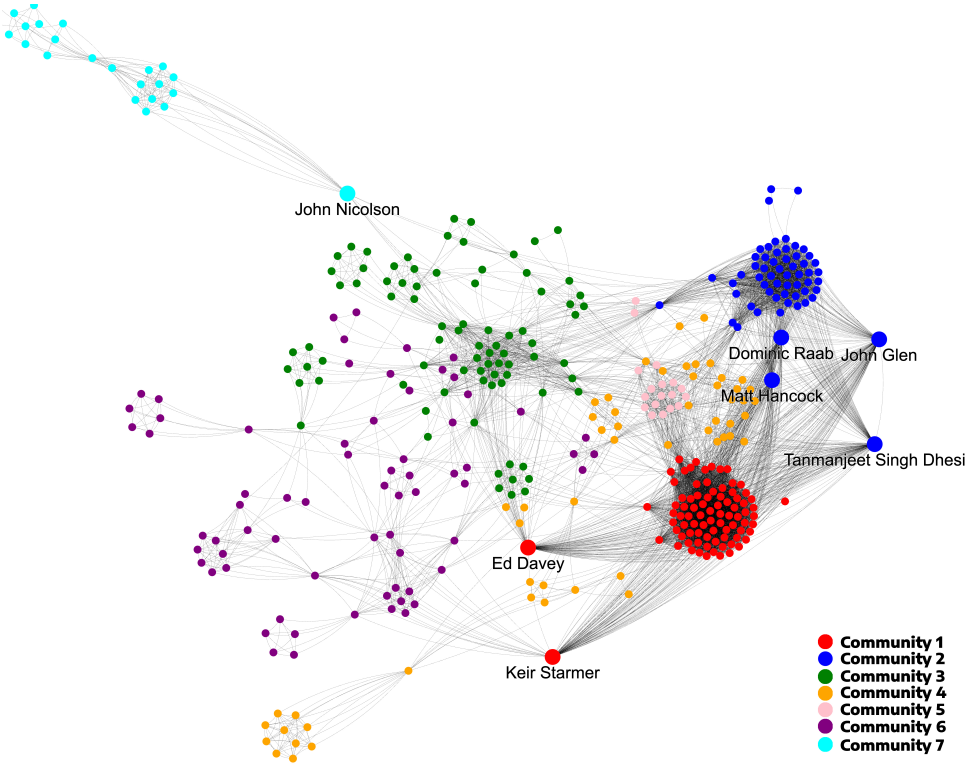


Figure 3: Community structure of the UK Parliament (in April 2024) based on the educational background of its members.

both in Scotland.

6. Conclusions

This work aimed to address a burning issue in recent HE, namely, worldwide low graduation rates. We approached this task by clustering, decision trees, and social network analysis methods. The analysis of four-year US colleges indicates that universities consistently achieving high graduation rates are characterized by a student body that is of a reasonable size and ethnically diverse. These schools are located in counties with a higher median household income, but also demand high tuition fees.

Examining the OULAD student data highlights the gravity of a sustained learning effort and frequent use of studying materials to succeed in HE. The education-based UK Parliament structure analysis was elaborated using the 2019 election results. In July 2024, the structure of the House of Commons changed significantly. Labour Party won 404 seats (202 in 2019). The Conservative Party secured 121 seats (365 in 2019). Liberal Democrats raised the number of their seats to 72 (11 in 2019). The Scottish National Party obtained nine seats (48 in 2019) [16]. Still, past parliament centrality measures clearly identified influential MPs for the newly elected parliament. Keir Starmer became Prime Minister, and Ed Davey, the Leader of the Liberal Democrats, won 61 more seats compared to 2019.

Overall, the obtained results insinuate the significance of a supportive environment (both financially and socially) that encourages building of functional relationships during students' formative years and promotes mutual collaboration and frequent exchange of ideas. Talent and diligence are equally essential to succeed in HE. Further research should also contemplate other factors that might account for students' academic success, like the criteria applied during the admission process, the candidates' personality traits, talents, and motivations, or the character of offered classes and educational experience. When analyzing professional alliances, we plan to utilize large language models (LLMs) and the recently introduced graph neural network models.

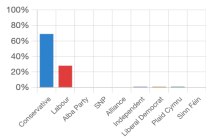
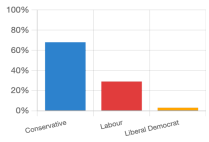
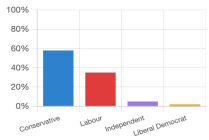
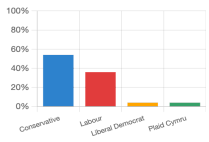
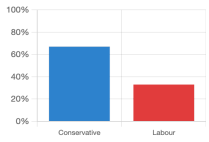
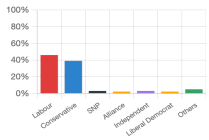
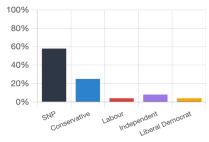
Communities	Description	Common Universities	Age of Members	Party Membership (%)
Community 1 (84 members, 59 Male, 25 Female)	Approximately 70% of the members in this community belong to the Conservative Party. 83 out of 84 members in this community attended Oxford while some of them also attended London School of Economics.	Oxford (83), London School of Economics (5)	The average age in Community 1 is **53**.	
Community 2 (59 members, 40 Male, 19 Female)	Approximately 70% of the members in this community belong to the Conservative Party, and around 30% are from Labour Party.	Cambridge (56), Oxford (4)	The average age in Community 2 is **55**.	
Community 3 (82 members, 53 Male, 29 Female)	Most of the members in this community belong to either Conservative or Labour Party.	University of London (29), Durham University (9), University of Sussex (10), University of Birmingham (10)	The average age in Community 3 is **53**.	
Community 4 (52 members, 33 Male, 19 Female)	In this community, approximately 36% of the members are from the Labour Party, and around 50% from the Conservative Party. We can observe that more members are from the Labour Party compared to the first and second community.	King's College London (13), University of Exeter (11), University College London (9), Aberystwyth University (9)	The average age in Community 4 is **51**.	
Community 5 (18 members, 11 Male, 7 Female)	This community has only members of the Conservative Party and the Labour Party; the community is more balanced in terms of gender.	London School of Economics (16), Brunel University (3)	The average age in Community 5 is **56**.	
Community 6 (71 members, 39 Male, 32 Female)	This community has more members of the Labour Party (around 45%).	University of Leeds (14), University of Hull (11), University of Salford (7), Open University (7)	The average age in Community 6 is **53**.	
Community 7 (24 members, 13 Male, 11 Female)	This community has a small number of the Labour Party and independent members, but it is dominated (approximately 60%) by members from the Scottish National Party. Most of the members attended the University of Glasgow, which is also located in Scotland.	University of Glasgow (13), University of Stirling (6)	The average age in Community 7 is **50**.	

Figure 4: Characteristics of the communities found in the UK Parliament based on the educational background of its members.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] C.C. Aggarwal, "Data mining: the textbook", Springer, 2015.

[2] M. Adnan, A.A.S. Alarood, M.I. Uddin and I.U.Rehman, "Utilizing grid search crossvalidation with adaptive boosting for augmenting performance of machine learning models", PeerJ Comput Sci., vol. 8, 2022, 29 p.

[3] E. Alyahyan and D. Dűştegőr, "Predicting academic success in higher education: literature review and best practices", Int. J. of Educational Technology in Higher Education, vol. 17, no. 3, 2020, 21 p.

[4] A.-L. Barabási and M. Pósfai, "Network Science", Cambridge University Press, 2016.

[5] P. Beja-Battis, “Overview of AdaBoost: Reconciling its views to better understand its dynamics”, arXiv:2310.18323v1, 2023, 39p.
[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, “Fast unfolding of communitiesin large networks”, J. of Statistical Mechanics: Theory and Experiment, vol. 2008, no. 10, 2008, 12 p.
[7] L. Breiman, J.H. Friedman, R.A. Olshen and Ch.J. Stone, “Classification and Regresion Trees”, Taylor & Francis, 1984.
[8] L. Breiman, “Random Forests”, Statistics Department, University of California, Berkeley, 2001, 33p.
[9] B.S. Duran and P.L. Odell, “Cluster Analysis: A Survey”, Springer, 2013.
[10] Y. Freund and R. Schapire, “Experiment with a new boosting algorithm”, Machine Learning: Proc. of the Thirteenth International Conf., pp.148 - 156.
[11] M. Girvan and M.E.J. Newman, “Community structure in social and biological networks”, PNAS, vol. 99, no. 12, 2002, pp. 7821-7826.
[12] B.W. Kerninghan and S. Lin, “An efficient heuristic procedure for partitioning graphs”, The Bell system technical journal, vol. 49, no. 2, 1970, pp. 291-307.
[13] T. Kohonen, “Self-Organized Formation of Topologically Correct Feature Maps”, Biological Cybernetics, vol. 43, no.1, 1982, pp. 59-69.
[14] J. Kuzilek, M. Hlosta and Z. Zdrahal, “Open university learning analytics dataset”, Sci Data, vol. 4, 2017.
[15] J. MacQueen, “Some methods for classification and analysis of multivariate observations”, Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281-297.
[16] “Membership of the UK Parliament”, https://commonslibrary.parliament.uk/research-briefings/sn01250/, accessed 2024-07-15.
[17] G. Molnár and Á. Kocsis, “Cognitive and non-cognitive predictors of academic success in higher education: a large-scale longitudinal study”, Stud. in Higher Educ., vol. 49, no. 9, 2024, pp. 1610-1624.
[18] O. Moscoso-Zea, A. Sampedro and S. Lujan-Mora, “Datawarehouse design for educational data mining”, Proc. of ITHET, 2016, pp. 1-6.
[19] A.-S. Nystrőm, C. Jackson and M.S. Karlsson, “What counts as success? Constructions of achievement in prestigious higher education programmes”, Research Papers in Education, vol. 34, no. 4, 2019, pp. 465-482.
[20] J.R. Quinlan, “C4.5: Programs for Machine Learning”, Morgan Kaufmann, 1993.
[21] P.J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”, Journal of Computational and Applied Mathematics, vol. 20, 1987, pp. 53-65.
[22] Ch. Stadtfeld, A. Vörös, T. Elmer, Z. Boda and I.J. Raabe, “Integration in emerging social networks explains academic failure and success”, PNAS, vol. 116, no. 3, 2019, pp. 792-797.
[23] T.G. Tareke, T.Z. Oo and K. Jozsa, “Bridging theoretical gaps to improve students´ academic success in higher education in the digital era: A systematic literature review”, Int. J. of Educational Research Open, vol. 9, article no. 100510, 2025, 12 p.
[24] V.A. Traag, L. Waltman and N.J. van Eck, “From Louvain to Leiden: guaranteeing well-connected communities”, Scientific Reports, vol. 9, article no. 5233, 2019, 12 p.
[25] M. Weatherton and E.E. Schussler, “Success for All? A Call to Re-examine How Student Success Is Defined in Higher Education”, CBE – Life Sciences Education, vol. 20:es3, 2021, pp. 1-13.
[26] L.N. Wood and Y.A. Breyer, “Success in Higher Education”, Springer Nature, 2017.
[27] X. Wu, V. Kumar, R. Quinlan et al. “Top 10 algorithms in data mining”, Knowl Inf Syst, vol. 14, 2008, pp. 1–37.
[28] Czechia in the data, “Šance na dostudování českých VŠ”, https://www.ceskovdatech.cz/graphs/vs2.php, accessed: 2025-06-29.