








ISSN: 2617-6548

URL: www.ijirss.com



Academic excellence in innovation ecosystems: A predictive approach to university rankings and startup ecosystem performance

 Mateus Dall'Agnol^{1*},  Elizane Maria de Siqueira Wilhelm²,  José Roberto Cruze³,  Celso Bilynkievycz dos Santos⁴,  Luiz Alberto Pilatti⁵

^{1,2,3,5}*Federal Technological University of Paraná, Ponta Grossa, Brazil.*

⁴*State University of Ponta Grossa, Ponta Grossa, Brazil.*

Corresponding author: Mateus Dall'Agnol (Email: mateus.agnol@ifto.edu.br)

Abstract

This study investigates the extent to which institutional attributes derived from global university rankings (QS and THE) influence the performance of territorial innovation ecosystems, as measured by the Global Startup Ecosystem Report (GSER). By integrating 2,145 institutional records linked to dozens of cities featured in all three rankings, the analysis applies feature selection techniques, support vector machine (SVM) regression models, and clustering methods. The results indicate that employability, academic reputation, and internationalization are strongly associated with the dynamism of the startup ecosystem. However, unmodeled contextual factors, such as local innovation policies, venture capital networks, and technological infrastructure, also exert significant influence. The adopted methodological approach combines statistical rigor with predictive capacity, offering valuable insights for data-driven innovation ecosystem planning and institutional strategies aimed at developing startups. From a practical perspective, the findings provide clear guidance for policymakers, university leaders, and innovation stakeholders to design targeted strategies that enhance graduate employability, strengthen institutional reputation, and foster international collaborations, thereby improving the global competitiveness of cities' startup ecosystems. The study further outlines practical directions for policymakers and university leaders, particularly in emerging cities, and recommends future model enhancements incorporating data on technological output, international co-authorship networks, and regional R&D investment.

Keywords: Data-driven planning, Innovation ecosystems, Machine learning, Startup ecosystems, University rankings.

DOI: 10.53894/ijirss.v8i6.9769

Funding: This study received no specific financial support.

History: Received: 17 July 2025 / **Revised:** 19 August 2025 / **Accepted:** 21 August 2025 / **Published:** 10 September 2025

Copyright: © 2025 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

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

Authors' Contributions: Conceptualization, methodology, data curation, formal analysis, writing—original draft, Mateus Dall'Agnol (MDA); Methodology, supervision, validation, writing—review and editing, Elizane Maria de Siqueira Wilhelm (EMSW); Methodology, data science, statistical analysis, software, José Roberto Cruz e Silva (JRCS); Methodology, data science, statistical analysis, software, Celso Bilynkiewicz dos Santos (CBDS); Methodology, supervision, production engineering, original draft, Luiz Alberto Pilatti (LAP). All authors have read and agreed to the published version of the manuscript.

Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Publisher: Innovative Research Publishing

1. Introduction

Technological innovation is widely recognized as a key driver of economic growth and social development, serving as a competitive differentiator for nations and regions [1, 2]. Within this context, Higher Education Institutions (HEIs) play a fundamental strategic role in shaping and energizing territorial innovation ecosystems, directly contributing to knowledge generation and dissemination, strengthening collaborative networks, and promoting sustainable social transformations [3-5]. This dynamic becomes especially relevant given global challenges related to sustainable development, digital transformation, and the increasing international competitiveness of cities.

From the perspective of the quintuple helix model, which integrates universities, government, industry, civil society, and the environment [6]. HEIs assume an even more significant role by articulating diverse actors and interests, thereby fostering the development of innovative solutions aimed at contemporary socio-environmental demands [5]. Innovative ecosystems such as those in Tel Aviv, London, and San Francisco exemplify successful integration among HEIs, the productive sector, and structured public policies [7]. The synergy established between social and economic spheres, combined with sustainability and highly qualified human capital, favors the emergence of innovative and economically competitive environments [8].

The growing appreciation of innovation as a vector for ecosystem development has boosted the use of academic metrics to evaluate the role of HEIs in this process. International academic rankings, such as the QS World University Rankings and Times Higher Education (THE), include indicators such as academic reputation, research impact, and graduate employability. Concurrently, indices such as the Global Startup Ecosystem Report (GSER) provide benchmarks regarding city performance regarding startup density, connectivity, and economic dynamism [9].

However, a significant theoretical and empirical gap persists in the literature concerning the specific ways in which the academic performance of HEIs directly influences the outcomes and competitive positioning of cities in international innovation rankings [3]. Recent studies indicate that variables such as international collaboration, revenue from industry partnerships, and the impact of scientific publications exhibit strong correlations with the vitality of territorial innovation ecosystems [10]. These findings suggest that academic excellence may play an essential predictive role in cities' innovation performance, thereby enhancing global competitiveness [11].

In this context, the present study aims to analyze the prospective influence of academic metrics on the future performance of cities in international innovation rankings, contributing directly to the strategic formulation of public policies and the evidence-based management of innovation ecosystems. The methodological approach is structured around three core axes: (i) analysis of the correlation between QS/THE and GSER rankings; (ii) development of predictive models based on academic reputation and scientific output; and (iii) clustering of HEIs according to similar attributes, exploring their connection to regional entrepreneurial ecosystems.

The choice to employ machine learning algorithms, notably Support Vector Machine (SVM), is grounded in their superior capacity to capture nonlinear relationships, network effects, and complex interactions among variables [12]. These methods overcome common limitations of conventional econometric approaches and enable the identification of variables with greater predictive power over innovation rankings, an essential aspect for evidence-based policymaking and data-driven ecosystem planning.

This study makes a methodological contribution by integrating academic metrics with advanced analytical techniques, thereby addressing an existing gap in the literature. It also offers theoretical contributions by proposing an analytical model that explicitly connects institutional academic performance to the positioning of cities' competitive startup ecosystems in innovative and sustainable contexts. In doing so, it provides robust quantitative evidence for institutional leaders, policymakers, and decision-makers.

The findings guide strategic efforts to strengthen the role of HEIs as central agents within territorial innovation and sustainability ecosystems. Audretsch et al. [2] emphasize that understanding the interactions among institutional actors is essential for designing effective public interventions. Rosado-Cubero et al. [13] and Madaleno et al. [14] highlight that reinforcing university incubators with public policy support enhances the territorial impact of innovation.

The following sections detail the methodological procedures adopted, including data selection criteria, processing techniques, and predictive modeling strategies. Subsequently, the results are presented and critically discussed in light of the existing literature, followed by the study's limitations, suggestions for future research, and general conclusions.

2. Method

2.1. Data Collection

Data were collected in September 2024 directly from the official portals of the QS and THE rankings through structured dataset downloads. In the case of the GSER, information was manually extracted, following rigorous pre-defined criteria. This process included independent cross-verification by two researchers at different times, using detailed protocols to ensure accuracy by validating entries by year, city, and ranking position. These procedures ensured robustness, internal consistency, and replicability. The temporal range (2022–2024) was selected due to methodological consistency across rankings in these years, allowing for longitudinal analyses and appropriate comparability.

2.2. Data Processing and Analysis

The analysis followed the Knowledge Discovery in Databases (KDD) framework Fayyad [15] encompasses three key stages: preprocessing, data mining, and postprocessing. During preprocessing, data were integrated using the Pandas library McKinney [16], a widely used Python toolkit for data manipulation and statistical computation.

Datasets were initially loaded separately and merged using common attributes (City, Country, Year). For QS and THE, institutional names enabled additional integration by the university. As the GSER dataset lacks institutional identifiers, its integration was based solely on City, Country, and Year. Institutional names were standardized to their English-language forms to ensure statistical alignment. Ordinal ranks were converted into continuous scores using interval means, following established best practices in the literature [17, 18].

Derived variables, such as Mean Rank, were created to enhance interpretability. Institutions with homonymous names in different cities received unique identifiers (University, City, Country). Missing values, such as the QS Overall Score, were imputed using multiple linear regression models, employing available attributes from the same year as predictors. This approach enhanced estimation accuracy while minimizing bias. The QS Sustainability Score was excluded due to its limited longitudinal coverage. After integration, the remaining missing values were imputed using local means; 35 records with persistent missing data were excluded after a preliminary statistical assessment confirmed no significant bias, yielding a final sample of 2,145 observations.

Feature selection was conducted using Correlation-Based Feature Selection (CFS) Hall [19] and Wrapper Subset Evaluation [20]. These interpretable techniques were prioritized over dimensionality-reduction methods such as PCA or autoencoders to maintain transparency for applications in policy design and institutional benchmarking.

In the data mining phase, the K-Means algorithm Arthur and Vassilvitskii [21] was used to identify clusters of universities, while regression modeling was performed using the SVM algorithm (SMOReg) [12]. These methods enabled the identification of patterns and the construction of predictive models for territorial innovation ecosystem performance.

2.3. Predictive Modeling

Predictive modeling was conducted using SMOReg, which was selected for its suitability in capturing nonlinear relationships among multiple continuous predictors. Model validation was performed via 10-fold cross-validation, which offers a balanced trade-off between bias and variance while maintaining computational efficiency compared to bootstrap or leave-one-out methods [22].

Two models were developed and compared: Model A, using seven variables selected via the CFS technique, and Model B, using five variables selected via the Wrapper method. Hyperparameter tuning was conducted through grid search, testing various regularization parameters ($C = 0.1, 1.0, 10$) and kernel types (Linear, RBF, Polynomial). The model using a polynomial kernel (degree = 1, $C = 1.0$) demonstrated superior performance, with lower Root Mean Squared Error (RMSE) and higher R^2 , indicating strong generalizability.

The choice of SVM was justified by its robust handling of non-linearities and lower risk of overfitting with moderate-sized data sets, in line with recent empirical evidence in educational research and innovation ecosystems [23, 24].

The following section presents empirical results, including descriptive statistics, clustering outcomes, and performance metrics of predictive models.

3. Results

This section presents the main findings derived from data integration, exploratory analysis, predictive modeling, and clustering. The analyses are conducted sequentially, beginning with the characterization of the integrated dataset, followed by statistical exploration, correlation analysis, and the development of predictive models. Table 1 summarizes the data from the integration and preprocessing stage, including data exploration and selection.

Table 1.

Summary of data resulting from the integration and preparation of the dataset (exploration and selection of data).

Variables	GSER	THE	QS	Intersection
Categorical variables				
Year	3	3	3	3
Institution (HEI)	—	1868	1674	818
City	1156	1204	1040	491
Country	144	102	104	80
Numeric variables				
Rank variables	1	1	1	3
Score variables	1	6	10	17
Total				
Variables (total)	5	13	17	26
Records	2996	5141	4217	2145

The data presented reveals the structure and volume of the information used. The THE ranking contains the highest number of records (5,141), followed by QS (4,217) and GSER (2,996). The intersection across all three rankings totals 2,180 records, representing HEIs with complete and compatible data, and serving as the primary basis for the subsequent analyses. Notably, QS offers greater granularity, with 17 variables, compared to 13 in THE and 5 in GSER, reinforcing its influence in the predictive modeling process. This overview confirms the feasibility of integrating the datasets and justifies using feature selection techniques to mitigate collinearity and ensure statistical robustness.

Table 2 presents a comparative analysis of the academic performance of HEIs (QS/THE averages) and the positioning of cities in the GSER, highlighting patterns of alignment, asymmetries, and potential discrepancies that inform the subsequent analyses.

Table 2.

Comparative Analysis Between Academic Rankings and GSER Performance.

Rank	City	IES	Country	Rank Medio THE	Rank Medio QS	Média QS THE	Rank GSER 2022	Rank GSER 2023	Rank GSER 2024	Média Rank GSER
1	New York City	Columbia University	EUA	11.0	20.5	15.8	2.0	2.0	2.0	2.0
2	London	Imperial College London	GBR	10.7	6.3	8.5	3.0	3.0	3.0	3.0
3	Los Angeles	University of California	EUA	20.7	37.7	29.2	4.0	4.0	4.0	4.0
4	Boston	Boston University	EUA	68.0	104.3	86.2	5.0	5.0	5.0	5.0
5	Beijing	Peking University	CHN	16.7	15.7	16.2	6.0	6.0	6.0	6.0
6	Shanghai	Fudan University	CHN	54.0	38.3	46.2	7.0	7.0	7.0	7.0
7	Bangalore	Indian Institute of Science	IND	292.2	188.7	240.4	8.0	8.0	8.0	8.0
8	Tel Aviv	Tel Aviv University	ISR	225.5	243.3	234.4	9.0	10.0	9.0	9.3
9	Paris	Paris Sciences and Letters	FRA	44.7	31.3	38.0	10.0	9.0	10.0	9.8
10	Seattle	University of Washington	EUA	27.0	76.0	51.5	11.0	12.0	12.0	11.7
11	Berlin	Freie Universittes Berlin	ALE	87.0	112.5	99.8	12.0	12.0	13.0	12.5
12	New Delhi	Jamia Millia Islamia	IND	600.5	883.8	742.2	13.0	13.0	11.0	12.7
13	Tokyo	University of Tokyo	JAP	37.7	24.7	31.2	15.0	14.0	14.0	14.2
14	Chicago	The University of Chicago	EUA	12.0	10.3	11.2	14.0	15.0	15.0	14.7
15	Shenzhen	Southern University of Science and Technology	CHN	164.7	267.3	216.0	18.0	16.0	18.0	17.3
16	Washington	Georgetown University	EUA	134.0	275.3	204.7	19.0	18.0	19.0	18.7
17	Sao Paulo	University of So Paulo	BRA	225.5	107.0	166.3	16.0	17.0	23.0	19.0
18	Singapore City	National University of Singapore	SGP	19.7	10.0	14.8	22.0	20.0	16.0	19.3

19	Austin	University of Texas at Austin	EUA	49.0	65.7	57.3	20.0	21.0	17.0	19.3
20	Mumbai	Institute of Chemical Technology	IND	833.8	166.0	499.9	17.0	25.0	20.0	20.7

The table reveals a significant concentration of cities with high academic performance and strong presence in innovation ecosystems. London, New York, Los Angeles, and Boston stand out for the number of well-ranked HEIs and their leading positions in the GSER. In contrast, cities such as Tel Aviv, Bangalore, and Toronto, despite having a smaller number of highly ranked HEIs, feature prominently among the top innovation ecosystems, indicating that non-academic factors such as policy incentives, access to capital, and entrepreneurial density play a significant role in strengthening these environments.

Table A1 (see Appendix A) presents the three identified clusters reflecting distinct patterns among HEIs. Cluster 1 groups globally excellent institutions with high reputation, scientific impact, and graduate employability. Cluster 2 concentrates on emerging HEIs with increasing internationalization, though they are still consolidating their reputational standing. Cluster 3 consists of regional institutions with limited academic performance and low international visibility, yet relevant to their local ecosystems.

Figure 1 displays the Pearson correlation heatmap for the numerical variables from the GSER, THE, and QS rankings. Preliminary observations indicate that the variable *SCORE Employment Outcomes* (QS) exhibits the highest absolute correlation with the GSER rank, which motivated the implementation of a simple linear regression between these variables.

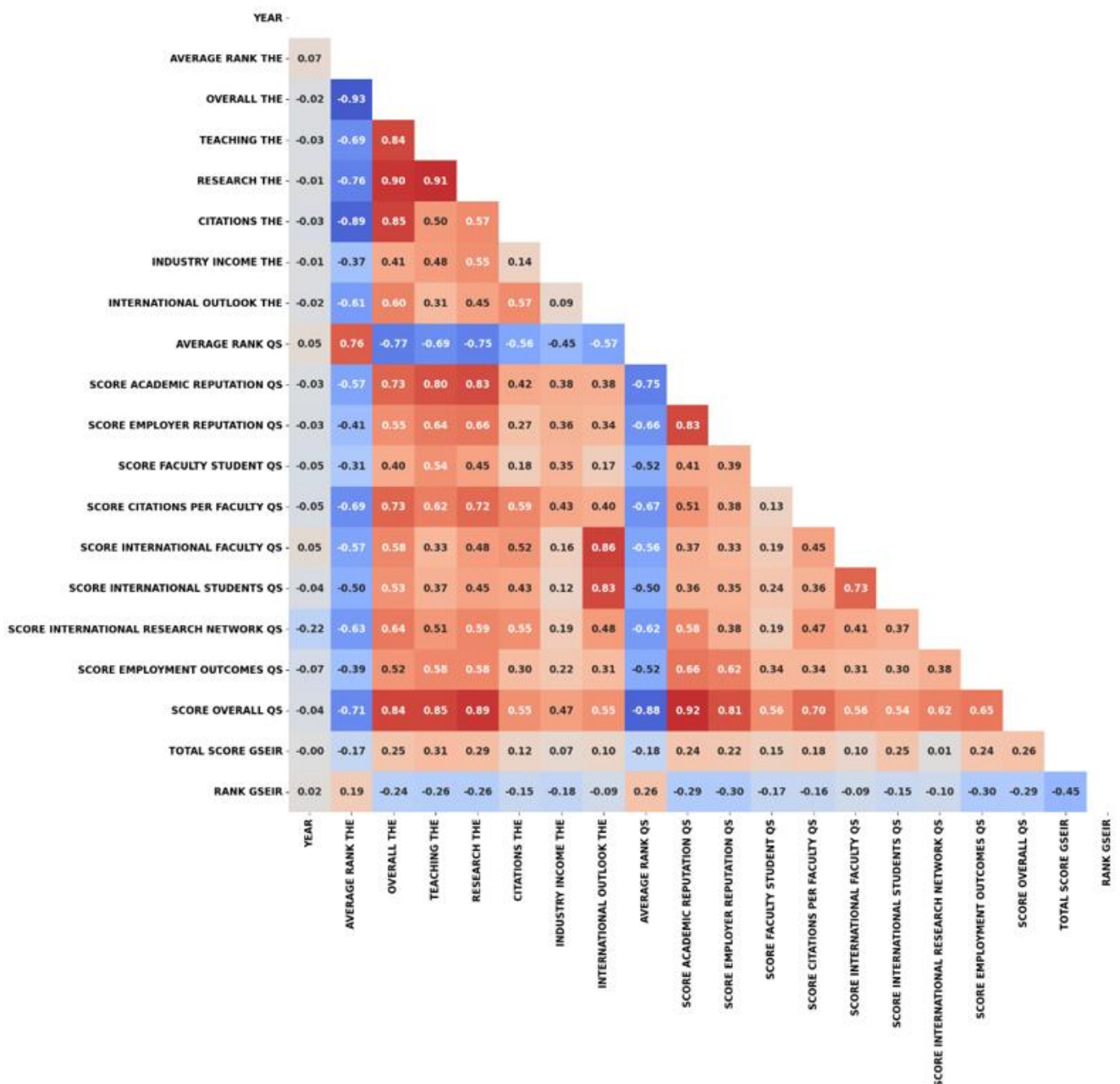


Figure 1.
Pearson Correlation Heatmap Between Academic Indicators and GSER Rank.

To explore the relationship between the GSER Rank and academic indicators, a simple linear regression was conducted using the variable *SCORE Employment Outcomes* (QS), which showed the highest correlation. The model yielded a regression coefficient (β_1) of -3.0836, with a 95% confidence interval ranging from -3.496 to -2.671. While the model is statistically significant ($p < 0.0001$), it demonstrates low explanatory power ($R^2 = 0.089$), indicating that only 8.9% of the variability in the GSER Rank is accounted for by this single predictor. This result highlights that the performance of territorial innovation ecosystems cannot be adequately captured by a single academic indicator, reinforcing the need for more robust multivariate models.

Given the limited explanatory power identified in the simple linear regression analysis, which, although statistically significant, reflects the inherent limitations of univariate models in capturing complex phenomena, variable selection was carried out using the CFS technique and the Wrapper method to identify predictors with greater explanatory potential. This methodological refinement represents a typical incremental stage in data mining, in which initial exploratory models are enhanced through more robust multivariate approaches. The variables selected through these two methods are detailed in Table 3.

Table 3.

Variables selected using the CFS technique and Wrapper method with the highest predictive power for the GSER Rank.

Models	A	B
Algorithm/Method	CFS Technique	Wrapper Method
Selected Independent Variables	Year,	Year,
	HEI,	City,
	City,	Country,
	Country,	International Research Network Score (QS),
	Industry Income (THE),	Employer Reputation Score (QS)
	Employer Reputation Score (QS),	
	Employment Outcomes Score (QS)	
Number of Predictive Variables	7	5

The multiple regression models, developed using the SMOReg algorithm and presented in Tables 4 and 5, demonstrate high predictive capacity. Based on attribute selection via the CFS technique, Model A exhibits greater explanatory power ($R^2 = 0.8766$). In contrast, Model B, developed using the Wrapper method, yields a lower mean absolute error (MAE = 40.28), indicating higher accuracy in individual predictions.

Table 4.

Model A: Multiple Regression developed with SMOReg and variables selected using the CFS method for GSER Rank prediction.

$$y = \beta_0 + \sum_{i=1}^n \beta_i \cdot x_i$$

Where:

y: value predicted by the model = GSER

β_0 : intercept of Model A (0.3676)

n: number of variables considered in the model

β_i : weight (or coefficient) associated with variable x_i

x_i : value of the i-th variable (normalized)

Selected Independent Variables (7):

Year, HEI, City, Country, Industry Income (THE), Employer Reputation Score (QS), Employment Outcomes Score (QS).

Example:

$y = 0.3676 + 0.001 \cdot \text{Year} - 0.052 \cdot \text{HEI=UNIVERSITY OF OXFORD} - 0.1134 \cdot \text{HEI=UNIVERSITY OF CAMBRIDGE} - 0.1403 \cdot \text{City=OXFORD} \dots$

=== Cross-validation ===	
=== Summary ===	
Correlation coefficient	0.9363
Mean absolute error	42.1041
Root mean squared error	88.6886
Relative absolute error	20.671 %
Root relative squared error	35.3651 %
Total Number of Instances	2145

Table 5.

Model B: Multiple Regression developed with SMOReg and variables selected using the Wrapper method for GSER Rank prediction.

$$y = \beta_0 + \sum_{i=1}^n \beta_i \cdot x_i$$

Where:

y: value predicted by the model = GSER

 β_0 : intercept of Model B (0.3662)

n: number of variables considered in the model

 β_i : weight (or coefficient) associated with variable x_i x_i : value of the i-th variable (normalized)

Selected Independent Variables (5):

Year, City, Country, International Research Network Score (QS), Employer Reputation Score (QS).

=== Cross-validation ===	
=== Summary ===	
Correlation coefficient	0.9251
Mean absolute error	40.2789
Root mean squared error	95.9247
Relative absolute error	19.7749 %
Root relative squared error	38.2506 %
Total Number of Instances	2145

Table 6 compares the performance metrics of the two multiple regression models developed using SMOReg, assessing their predictive capacity and the significance of the selected predictors based on academic variables.

Table 6.

Predictive Model Performance Summary.

Metric	Model A (CFS)	Model B (Wrapper)	Best Model
Correlation (R)	0.9363	0.9251	Model A
Coefficient of Determination (R ²)	0.8766	0.8558	Model A
Mean Absolute Error (MAE)	42.10	40.28	Model B
Root Mean Squared Error (RMSE)	88.69	95.92	Model A
Relative Absolute Error (RAE)	20.67%	19.77%	Model B
Root Relative Squared Error (RRSE)	35.37%	38.25%	Model A

Although the models exhibit high coefficients of determination (R²), the absolute and relative error metrics reveal practical limitations. In Model A, the RMSE reached 88.7 positions in the GSER ranking, while in Model B it was 95.9. The MAE exceeded 56 positions in both cases, indicating that the predictions should be interpreted as general trends rather than precise rankings.

Moreover, the elevated residuals suggest the influence of unmodeled factors, such as local innovation policies, venture capital networks, entrepreneurial culture, digital infrastructure, and the activity of innovation hubs and accelerators independent of HEIs. The absence of these variables reduces the predictive reach of the models. It reinforces the need to complement them with contextual analyses and territorial data for practical use in public policy design.

Figure 2 illustrates the dispersion between predicted and observed GSER rankings for Model A. The proximity of the data points to the reference line indicates good generalization capacity. However, significant deviations highlight unmodeled factors such as venture capital availability, innovation hubs, and local fiscal policies.

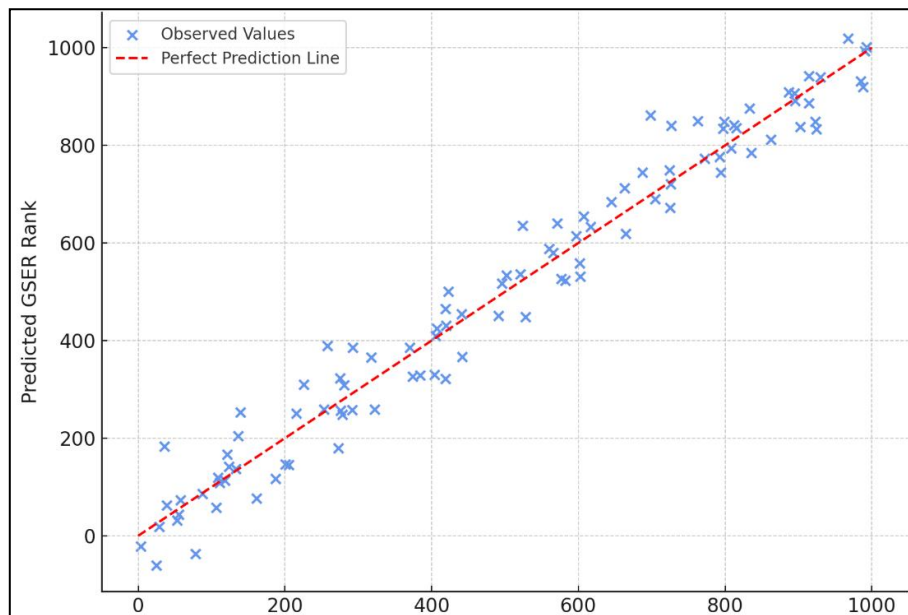


Figure 2.
Predicted vs. Observed GSER Rank (Model A).

The dashed line represents the perfect prediction ($y = x$). Deviations from this line indicate residual errors.

The dashed line represents perfect alignment. Distant points indicate deviations that can be explained by contextual variables not captured by the models.

These results confirm that territorial innovation performance is strongly associated with academic attributes, particularly institutional reputation, employability, and internationalization. The following section offers a critical discussion of these findings in light of international literature.

4. Discussion

The role of HEIs as strategic vectors in the consolidation of innovation ecosystems is widely recognized in the literature, particularly in models that incorporate the logic of the quintuple helix [11]. Recent studies show that attributes such as academic reputation, internationalization, and scientific impact directly influence the competitiveness of startup ecosystems [25].

The data presented in Table 1 demonstrate the robustness of the empirical foundation, reflecting the integration of QS, THE, and GSER rankings and the need for dimensionality reduction, as summarized in Table 4. This step enabled the selection of the most relevant variables for predictive analysis, highlighting the centrality of attributes such as employability and internationalization.

The exploratory analysis presented in Table 2 indicates that, although there is a general alignment between academic performance and GSER ranking, this relationship is neither universal nor linear. Cities such as New York (1st) and London (2nd) exemplify ecosystems in which academic excellence, reflected in low QS/THE means (15.8 and 8.5, respectively), is directly associated with high innovation performance. This pattern confirms the propositions of Hausberg and Korreck [26] who emphasize the centrality of academic knowledge and intellectual capital density in forming robust innovation ecosystems. Coad et al. [27] highlight that global innovation hubs tend to emerge in contexts where world-class universities act as strategic anchors.

However, cases such as São Paulo (17th), with a significantly higher QS/THE average (166.3), show that other factors, such as targeted public policies, business density, and technological infrastructure, also play a decisive role in shaping innovation ecosystems. This is also supported by Yusuf [28], who acknowledges that in specific contexts, elements external to academia can compensate for institutional academic limitations.

The regional asymmetries observed in Table 2 become particularly evident when analyzing cities like Bangalore (7th) and Tel Aviv (8th), which maintain robust innovation ecosystems despite relatively low academic scores (240.4 and 234.4, respectively). This pattern supports the analyses of Breznitz and Zhang [29] and Mason et al. [30] who emphasize the central role of entrepreneurial culture, access to venture capital, and favorable regulatory environments as compensating factors for the absence of elite HEIs. Sandström et al. [31] reinforce that in dynamic markets, the density of startups and the effectiveness of innovation policies can overcome academic limitations, strengthening regional ecosystems.

In contrast, the cases of Beijing (5th) and Shanghai (6th) illustrate an opposite model, in which ecosystem performance is deeply rooted in the convergence of academic excellence, expressed in low QS/THE means (16.2 and 46.2), and significant public and private investment in science, technology, and innovation, as shown by Ma [32], Wang et al. [33] and Lerman et al. [8]. These findings reinforce that while different trajectories may lead to the consolidation of innovation ecosystems, the combination of academic capital, financial resources, and institutional policies remains critical for sustaining such environments.

The temporal variation observed in the GSER, particularly the stability of cities like New York (2.0) and London (3.0), aligns with what Guerrero et al. [10] describe as mature innovation ecosystems, supported by robust institutional frameworks and consistent policies for talent attraction and investment. On the other hand, fluctuations in cities such as Berlin (11th) and Singapore (18th) may reflect situational shifts, such as recent adjustments in innovation policies or changes in talent retention, phenomena discussed by Lerman et al. [8] in European and Asian contexts.

The decline of São Paulo, from 16th in 2022 to 23rd in 2024, corroborates earlier analyses Ma [32] that point to persistent structural challenges in the Brazilian institutional environment. It also confirms evidence from Castañón and Bustamante [34] who highlight intensified regional competition and regulatory barriers as limiting factors for the strengthening of innovation ecosystems in the Global South.

The New Delhi (12th) and Mumbai (20th) cases reveal a pronounced asymmetry between academic performance, as indicated by QS/THE averages of 742.2 and 499.9, respectively, and GSER performance. This suggests that innovation ecosystems can thrive even without highly ranked HEIs, as identified in emerging innovation ecosystems like those in the Global South, where robust institutional frameworks and regional policy coordination often outweigh moderate academic performance. This finding corroborates the analyses of and Etzkowitz [35], Duan et al. [36] and Javanmardi [37] who emphasize the central role of extrinsic factors such as talent attraction policies, robust government incentives, and the availability of venture capital.

Similarly, Wang et al. [38] and Gao et al. [39] argue that in emerging markets, elements like entrepreneurial density, regulatory flexibility, and global connectivity can offset academic limitations, enabling highly competitive ecosystems.

In the Brazilian context, this pattern is reflected in the analysis of São Paulo, which performs worse in the GSER than international hubs. This suggests a need for greater alignment between HEIs and industry and the expansion of internationalization. This need is comparable to the model observed in Seattle (10th), where the University of Washington (QS/THE average: 51.5) plays a central role as a local ecosystem anchor, linking academic excellence with entrepreneurial dynamics and innovation policies, as shown by Cantner et al. [11].

Table A1 (Appendix A) complements this analysis by clustering HEIs into three distinct groups, reflecting structural patterns identified in the literature on innovation ecosystems. Cluster 1, labeled “Institutions of Excellence with Global Integration,” includes HEIs with a high reputation, substantial scientific impact, and high employability, acting as anchors in leading innovation ecosystems. This configuration aligns with the propositions of Etzkowitz and Leydesdorff [4] regarding the centrality of HEIs in triple and quintuple helix models.

Cluster 2, identified as “Emerging Institutions with an International Profile,” consists of HEIs on a path of international expansion, with increasing impact on their ecosystems but still limited by moderate academic performance and reputation. This profile aligns with Cai et al. [3] and Coad et al. [27] who highlight the role of emerging HEIs as intermediate catalysts in ecosystem dynamics.

Finally, Cluster 3, “Regional Institutions with Limited Performance,” includes HEIs with low indicators for reputation, internationalization, and scientific impact, whose influence on global innovation ecosystems is reduced. However, they remain key players in their local socioeconomic contexts as Guerrero et al. [10] argued.

These findings are supported by the analysis in Figure 3, which shows strong correlations between employability (QS Employment Outcomes Score) and GSER rank, as well as between academic reputation and internationalization. These results reinforce arguments in the literature Wang et al. [38] that, while multiple factors contribute to the vitality of territorial innovation ecosystems, attributes directly associated with academic performance, particularly employability, institutional reputation, and internationalization, play a central role in shaping these environments.

The predictive models summarized in Tables 4 and 5 and compared in Table 6 corroborate and deepen the exploratory analysis and clustering findings. Model A, developed with variables selected through the CFS method, has a higher explanatory power ($R^2 = 0.8766$), demonstrating its robustness in identifying general patterns between academic attributes and the performance of the startup ecosystem. Model B, based on the Wrapper method, while having a slightly lower R^2 , achieves the lowest mean absolute error (MAE = 40.28), suggesting higher precision for individual predictions. This complementarity aligns with discussions in the literature on trade-offs between generalization and local precision in machine learning models applied to complex systems [40, 41].

Both models confirm that employability, institutional reputation, and internationalization are the most determining variables for the performance of innovation ecosystems. This finding is consistent with the propositions of Wang et al. [33] who emphasize that attributes tied to the ability of HEIs to generate qualified human capital, international visibility, and high-impact scientific production are central to building globally competitive ecosystems.

This observation has significant implications for drawing up public policies to promote an innovation ecosystem. In particular, it suggests that strategies to strengthen HEIs by improving academic quality, expanding internationalization, and enhancing graduate employability may positively affect cities' performance in global innovation rankings.

Furthermore, the results show that while multiple trajectories can lead to the development of innovative ecosystems, the presence of universities with strong institutional attributes is a high-impact competitive advantage. This supports the argument that public policies combining the academic dimension with broader territorial strategies, including support for applied research, startup development, favorable regulatory environments, and integration with productive sectors, have greater potential to create sustainable innovation environments.

This configuration aligns with the need to orchestrate internal capabilities and external assets, as Priyono and Hidayat [42] propose in their typology of dynamic capabilities within open innovation, where resource-strategy-capability alignment is essential to design responsive ecosystems.

Finally, the predictive models developed offer a practical and strategic tool for anticipating trends and supporting evidence-based decision-making in innovation policy and ecosystem planning. The ability to forecast, with a high degree of accuracy, the future performance of cities in the GSER based on institutional attributes of local universities enables public managers and territorial planners to act more proactively and assertively. The methodological approach adopted in this study, combining exploratory analysis, statistical modeling, and machine learning techniques, contributes to consolidating an applied research agenda oriented toward strengthening the scientific foundations of innovation ecosystems. Furthermore, this analytical framework facilitates the alignment between higher education policies and broader innovation agendas, especially in contexts where resources are limited and strategic prioritization is essential. By providing a replicable and data-driven methodology, this study contributes to academic literature and practical governance frameworks. Integrating machine learning techniques with multivariate modeling expands the frontier of applied research in innovation ecosystems, reinforcing the strategic importance of universities as key drivers of knowledge-based economic development in contemporary cities.

5. Conclusion

This study confirmed that institutional variables derived from international academic rankings, such as QS and THE, exert a statistically significant influence on the performance of startup ecosystems, as measured by the GSER. Although the magnitude of this association is moderate, the findings highlight that attributes such as graduate employability, institutional reputation, and internationalization are directly correlated with cities' economic and innovation dynamism. The integrated application of data mining techniques, feature selection, and predictive regression revealed robust patterns connecting the activities of HEIs to the strengthening of territorial innovation ecosystems.

Beyond providing a retrospective understanding of these relationships, the results also contribute to anticipating trends within innovation ecosystems. The predictive approach based on machine learning and multivariate modeling offers a set of analytical tools for medium- and long-term ecosystem planning, enabling the formulation of public policies that are more responsive to ongoing technological and institutional transformations.

From a methodological perspective, the findings suggest that the choice between predictive models depends on the analytical goal: when the aim is to maximize explanatory power, the model developed through CFS-based feature selection proves more effective; conversely, if the objective is to minimize absolute error in individual predictions, the wrapper-based model yields superior performance. This duality offers helpful methodological guidance for researchers and institutional decision-makers, enabling the selection of models better aligned with specific analytical and strategic needs.

This study also promotes understanding of the interactions between academic performance and the innovation ecosystem by filling a methodological gap in the literature, namely, integrating academic metrics with indicators from entrepreneurial ecosystems. In addition to contributing to developing more accurate explanatory models, the findings provide actionable insights for public policy design, university governance, and the global strengthening of territorial innovation ecosystems.

As an innovative contribution, this research distinguishes itself from the existing literature by quantitatively and predictively integrating institutional metrics from academic rankings with indicators of territorial performance in innovation. While prior studies have examined the role of HEIs in territorial development, few have employed machine learning techniques to investigate how academic attributes explain structural variations in rankings such as the GSER. The combination of statistical feature selection with predictive modeling represents a significant methodological advance over traditional approaches, allowing the capture of complex interdependencies between higher education and startup ecosystems. This approach provides empirical evidence to support data-driven policymaking and reinforces the strategic role of HEIs in the vitality of contemporary innovation ecosystems.

5.1. Contributions to the Literature

This study offers a novel contribution to the literature by empirically demonstrating the predictive power of academic attributes, such as institutional reputation, internationalization, and graduate employability, on the performance of startup ecosystems. While previous research has underscored the strategic role of HEIs in regional development, few studies have operationalized this relationship using predictive modeling techniques grounded in machine learning. By integrating data from global academic rankings (QS, THE) with ecosystem-level innovation indicators (GSER), this study advances a quantitative framework to identify statistically significant and contextually relevant patterns. Using feature selection algorithms (CFS and Wrapper) and predictive regression models (SMOReg) provides methodological robustness and allows nuanced analysis across distinct institutional configurations.

Furthermore, the article bridges a methodological gap by linking the traditional indicators of academic excellence with entrepreneurial ecosystem metrics, an intersection that remains underexplored in empirical studies. This integration between domains strengthens theoretical discussions about the role of HEIs in triple and quintuple helix structures, as well as offering analytical tools for data-driven decision-making in policy to promote innovation ecosystems. In doing so, the study reinforces the centrality of academia in territorial development and introduces a replicable, scalable model for forecasting innovation outcomes based on institutional academic performance. These findings are particularly relevant for scholars, policymakers, and university leaders seeking to align academic strategies with broader innovation agendas.

5.2. Limitations and Future Research Directions

Although the developed models yielded robust results, the scope of the sample, composed primarily of HEIs located in developed regions, highlights the need for future investigations that broaden the geographic diversity of the institutions analyzed, with more representative inclusion of HEIs from Global South countries. Such expansion would enable a more

comprehensive understanding of innovation ecosystem dynamics across different sociopolitical, economic, and institutional contexts.

Additionally, future studies may enrich the analytical complexity by incorporating new data sources that complement traditional institutional rankings. The inclusion of technological production indicators (e.g., number of patents registered per city or institution, based on databases such as WIPO or USPTO), data on scientific collaboration networks (e.g., international co-authorships from platforms like Scopus or SciVal), and regional-level metrics on R&D investment, startup density, or the quality of digital infrastructure could substantially improve the explanatory power regarding the performance of the startup ecosystem. These variables can be operationalized using public and institutional databases such as the Global Innovation Index, Startup Genome, OECD regional statistics, or municipal data from innovation ecosystem observatories.

An additional limitation concerns the use of the GSER as a dependent variable. Although widely referenced and influential, the GSER methodology lacks complete transparency, with limited disclosure of data sources, weighting criteria, and changes between editions. This opacity may affect the replicability and consistency of longitudinal analyses, potentially introducing latent biases or distortions in the measured innovation performance of cities. Future research should consider triangulating GSER scores with complementary or alternative indicators of entrepreneurial dynamism to strengthen the robustness of the findings. It is also important to highlight that the GSER does not provide a fully replicable methodological framework, and its scores may vary according to opaque or ad hoc criteria, which presents limitations for comparative modeling and scientific generalization.

Furthermore, social network analysis (SNA) techniques could be employed to map interinstitutional collaboration structures, while deep learning approaches may be used to explore temporal and nonlinear patterns in greater depth. These methodological advances would support the development of hybrid models with enhanced predictive and contextual capabilities, thereby expanding the potential use of the findings to guide innovation policies adapted to the realities of cities at different stages of development.

In addition to broadening the scope of variables and methods, future research could investigate the longitudinal evolution of innovation ecosystems, focusing on how academic attributes influence their long-term resilience and adaptability. Incorporating temporal dynamics, such as changes in university rankings, fluctuations in regional innovation funding, or the rise and fall of local startup hubs, could help reveal causal relationships and feedback loops that shape the trajectories of innovation ecosystems. This temporal dimension is particularly relevant for cities in transitional or emerging contexts, where volatility in academic and entrepreneurial environments may catalyze or constrain sustainable development. In addition, comparative studies that examine how different governance structures mediate the relationship between higher education and innovation outcomes would offer valuable insights into the institutional arrangements that enable or hinder the impact of universities on ecosystem vitality.

References

- [1] E. G. Carayannis, D. F. Campbell, and E. Grigoroudis, "Helix trilogy: The triple, quadruple, and quintuple innovation helices from a theory, policy, and practice set of perspectives," *Journal of the Knowledge Economy*, vol. 13, no. 3, pp. 2272-2301, 2022. <https://doi.org/10.1007/s13132-021-00813-x>
- [2] D. B. Audretsch, M. Belitski, and R. Caiazza, "Start-ups, innovation and knowledge spillovers," *Journal of Technology Transfer*, vol. 46, no. 6, pp. 1995-2016, 2021. <https://doi.org/10.1007/s10961-021-09846-5>
- [3] Y. Cai, J. Ma, and Q. Chen, "Higher Education in Innovation Ecosystems," *Sustainability*, vol. 12, no. 11, p. 4376, 2020. <https://doi.org/10.3390/su12114376>
- [4] H. Etzkowitz and L. Leydesdorff, "The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university-industry-government relations," *Research Policy*, vol. 29, no. 2, pp. 109-123, 2000. [https://doi.org/10.1016/S0048-7333\(99\)00055-4](https://doi.org/10.1016/S0048-7333(99)00055-4)
- [5] M. Guerrero, D. Urbano, and E. Gajón, "Entrepreneurial university ecosystems and graduates' career patterns: do entrepreneurship education programmes and university business incubators matter?," *Journal of Management Development*, vol. 39, no. 5, pp. 753-775, 2020. <https://doi.org/10.1108/JMD-10-2019-0439>
- [6] E. G. Carayannis, T. D. Barth, and D. F. Campbell, "The quintuple helix innovation model: Global warming as a challenge and driver for innovation," *Journal of Innovation and Entrepreneurship*, vol. 1, no. 1, p. 2, 2012.
- [7] G. S. E. R. Startup Genome, "Global startup ecosystem report 2023. San Francisco," Startup Genome, Global Startup Ecosystem Report, 2023.
- [8] L. V. Lerman, W. Gerstlberger, M. F. Lima, and A. G. Frank, "How governments, universities, and companies contribute to renewable energy development? A municipal innovation policy perspective of the triple helix," *Energy Research & Social Science*, vol. 71, p. 101854, 2021.
- [9] WIPO, *Global innovation index 2023: Innovation in the face of uncertainty*. Geneva: WIPO, 2023.
- [10] M. Guerrero, F. Herrera, and D. Urbano, "Strategic knowledge management within subsidised entrepreneurial university-industry partnerships," *Management Decision*, vol. 57, no. 12, pp. 3280-3300, 2019.
- [11] U. Cantner, J. A. Cunningham, E. E. Lehmann, and M. Menter, "Entrepreneurial ecosystems: A dynamic lifecycle model," *Small Business Economics*, vol. 57, no. 1, pp. 407-423, 2021.
- [12] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199-222, 2004.
- [13] A. Rosado-Cubero, A. Hernández, F. J. B. Jiménez, and T. Freire-Rubio, "Does gender affect entrepreneurship? Evidence from Spanish and Argentinian business incubators," *Journal of Business Research*, vol. 170, p. 114326, 2024.
- [14] M. Madaleno, M. Nathan, H. Overman, and S. Waights, "Incubators, accelerators and urban economic development," *Urban Studies*, vol. 59, no. 2, pp. 281-300, 2022.
- [15] U. M. Fayyad, "Diving into databases: SQL is helpless in the face of massive, accumulating data stores," *Database Programming and Desing*, vol. 11, pp. 24-34, 1998.

- [16] W. McKinney, *Python for data analysis: Data wrangling with pandas, numpy, and ipython*. Sebastopol, CA: O'Reilly Media, 2012.
- [17] L. Bornmann, L. Leydesdorff, and R. Mutz, "The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits," *Journal of Informetrics*, vol. 7, no. 1, pp. 158-165, 2013.
- [18] A. Robitzsch, "Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods," *Frontiers in Education*, vol. 5, p. 589965, 2020.
- [19] M. A. Hall, "Correlation-based feature selection for machine learning," Doctoral Thesis, University of Waikato University of Waikato, 1999.
- [20] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273-324, 1997.
- [21] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," New Orleans: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, 2006.
- [22] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer, 2009.
- [23] Z. Wang, Q. He, S. Xia, D. Sarpong, A. Xiong, and G. Maas, "Capacities of business incubator and regional innovation performance," *Technological Forecasting and Social Change*, vol. 158, p. 120125, 2020.
- [24] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.
- [25] A. Colombelli, E. D'Amico, and E. Paolucci, "When computer science is not enough: Universities knowledge specializations behind artificial intelligence startups in Italy," *The Journal of Technology Transfer*, vol. 48, no. 5, pp. 1599-1627, 2023.
- [26] J. P. Hausberg and S. Korreck, "Business incubators and accelerators: A co-citation analysis-based, systematic literature review," *Journal of Technology Transfer*, vol. 45, no. 1, pp. 151-176, 2020.
- [27] A. Coad, U. Kaiser, and J. Kuhn, "Spin doctors vs the spawn of capitalism: Who founds university and corporate startups?," *Research Policy*, vol. 50, no. 10, p. 104347, 2021. <https://doi.org/10.1016/j.respol.2021.104347>
- [28] A. Yusuf, "An appraisal of research in Nigeria's university sector," *Journal of Research in National Development*, vol. 10, no. 2, pp. 321-330, 2012.
- [29] S. M. Breznitz and Q. Zhang, "Fostering the growth of student start-ups from university accelerators: An entrepreneurial ecosystem perspective," *Industrial and Corporate Change*, vol. 28, no. 4, pp. 855-873, 2019.
- [30] C. Mason, M. Anderson, T. Kessl, and M. Hruskova, "Promoting student enterprise: Reflections on a university start-up programme," *Local Economy*, vol. 35, no. 1, pp. 68-79, 2020.
- [31] C. Sandström, K. Wennberg, M. W. Wallin, and Y. Zherlygina, "Public policy for academic entrepreneurship initiatives: A review and critical discussion," *Journal of Technology Transfer*, vol. 43, no. 5, pp. 1232-1256, 2018.
- [32] J. Ma, "Developing joint R&D institutes between Chinese universities and international enterprises in China's innovation system: A case at Tsinghua University," *Sustainability*, vol. 11, no. 24, p. 7133, 2019.
- [33] Q. Wang, L.-N. Zhang, and Q. Lian, "Innovation facilitated by universities: Balancing enterprise and regional demands," *Discrete Dynamics in Nature and Society*, vol. 2022, no. 1, p. 2817232, 2022.
- [34] L. d. C. Á. Castañón and R. P. Bustamante, "Open innovation from the university to local enterprises: Conditions, complexities, and challenges," *Telos: Revista de Estudios Interdisciplinarios en Ciencias Sociales*, vol. 23, no. 3, pp. 692-709, 2021. <https://doi.org/10.36390/telos233.12>
- [35] H. Etzkowitz, "Incubation of incubators: Innovation as a triple helix of university-industry-government networks," *Science and Public Policy*, vol. 29, no. 2, pp. 115-128, 2002. <https://doi.org/10.3152/147154302781781056>
- [36] X. Duan, P. Sun, X. Wang, and B. Zhan, "Evolutionary game analysis of industry-university-research cooperative innovation in digital media enterprise cluster based on GS Algorithm," *Wireless Communications and Mobile Computing*, vol. 2022, no. 1, p. 5701917, 2022. <https://doi.org/10.1155/2022/5701917>
- [37] S. Javanmardi, "Identifying factors influencing Iranian innovation ecosystem and determining their links," *Sustainable Futures*, vol. 4, p. 100081, 2022. <https://doi.org/10.1016/j.sftr.2022.100081>
- [38] J. Wang, Y. Song, M. Li, C. Yuan, and F. Guo, "Study on low-carbon technology innovation strategies through government-university-enterprise cooperation under carbon trading policy," *Sustainability*, vol. 14, no. 15, p. 9381, 2022. <https://doi.org/10.3390/su14159381>
- [39] Q. Gao, L. Cui, Y. K. Lew, Z. Li, and Z. Khan, "Business incubators as international knowledge intermediaries: Exploring their role in the internationalization of start-ups from an emerging market," *Journal of International Management*, vol. 27, no. 4, p. 100861, 2021. <https://doi.org/10.1016/j.intman.2021.100861>
- [40] R. Jain and W. Xu, "Artificial Intelligence based wrapper for high dimensional feature selection," *BMC Bioinformatics*, vol. 24, no. 1, p. 392, 2023.
- [41] F. Mohtasham, M. Pourhoseingholi, S. S. Hashemi Nazari, K. Kavousi, and M. R. Zali, "Comparative analysis of feature selection techniques for COVID-19 dataset," *Scientific Reports*, vol. 14, no. 1, p. 18627, 2024.
- [42] A. Priyono and A. Hidayat, "Dynamic capabilities for open innovation: A typology of pathways toward aligning resources, strategies and capabilities," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 8, no. 4, p. 206, 2022. <https://doi.org/10.3390/joitmc8040206>

Appendix A

Table 1.
Sample Description and Cluster Profiles Based on Centroids

Data Scope	Variable	Class	General		Cluster 1	Cluster 2	Cluster 3
			N	2145	403	556	1186
			%	100	18.8	25.9	55.3
Institutional	University	Mode	University of California		University of California	London School of Economics and Political Science	University of the Andes

Ranks	Rank THE	Mode	601-800	201-250	201-250	801-1000
	Rank QS	Mode	1001-1200	141	801-1000	1001-1200
Geographic	City	Mode	London	Paris	London	Tokyo
	Country	Mode	United States	United States	United Kingdom	United States
QS Indicators	Rank Medio QS	μ	555.2871	66.0062	398.3812	778.9721
		\pm	380.3953	24.8972	237.6108	324.1145
	SCORE Overall QS	μ	29.8072	62.4858	33.1197	17.1503
		\pm	21.3762	17.3748	14.0491	10.1604
	SCORE Academic Reputation QS	μ	27.6510	66.0062	25.1525	15.7892
		\pm	25.4091	24.8972	17.0212	13.3448
	SCORE Employer Reputation QS	μ	27.0848	60.2367	24.7120	16.9322
		\pm	27.1020	29.6208	20.4358	18.8715
	SCORE Faculty Student QS	μ	31.9474	58.4091	28.1586	24.7319
		\pm	28.8400	32.3254	24.6179	23.8656
	SCORE Citations per Faculty QS	μ	31.8552	62.5776	38.3541	18.3692
		\pm	29.2455	26.8955	25.1344	21.9856
	SCORE International Faculty QS	μ	34.5255	56.5870	70.3262	10.2455
		\pm	35.9706	34.6321	29.3647	13.1668
	SCORE International Students QS	μ	32.0971	51.2880	60.9343	12.0570
		\pm	32.8731	32.4481	30.4014	15.7403
	SCORE International Research Network QS	μ	49.4666	78.6637	61.2759	34.0093
		\pm	31.9498	21.3864	28.0231	27.0086
	SCORE Employment Outcomes QS	μ	27.6508	55.5384	25.1079	19.3667
		\pm	24.5425	27.6026	19.2564	17.9290
GSER	Rank GSER	μ	264.6065	131.1067	284.9190	300.4468
		\pm	250.6681	144.3492	240.8820	268.1428
	Total Score GSER	μ	19.1818	34.0285	20.5722	13.4850
		\pm	38.2457	51.1223	47.2410	24.6235
THE Indicators	Rank Medio THE	μ	645.3449	127.2431	384.8273	943.5261
		\pm	472.9963	121.9529	224.8308	404.898
	Overall THE	μ	39.4668	64.9821	46.0363	27.7170
		\pm	18.3621	13.1396	9.5729	11.3450
	Teaching THE	μ	32.2977	55.6198	30.3994	25.2628
		\pm	15.1770	15.7605	8.2857	8.0082
	Research THE	μ	30.9104	59.8258	32.3640	20.4037
		\pm	18.9195	18.1535	11.1739	9.0296
	Citations THE	μ	56.4558	79.1632	70.7143	42.0554
		\pm	26.5720	17.5170	18.0168	23.3038
	Industry Income THE	μ	50.8978	67.2037	48.9217	46.2835
		\pm	17.7317	20.0606	14.5419	14.8236
	International Outlook THE	μ	56.1937	69.0126	79.5776	40.8753
		\pm	22.5368	18.5925	13.090	13.1959