analytics

MDPI

Article

Exploring University Ranking Stability and Predictability: Cross-Ranking Consistency, Persistence Baselines, and Machine Learning

Emil Andekin [1], Orazmukhamed Bekmurat [1], Vassiliy Serbin [1], Yersain Chinibayev [1], Kanibek Sansyzbay [2], Shingis Kadirkulov [3,*], Baurzhan Abzhanov [3] and Ayaulym Oralbekova [4,*]

[1] Institute of Automation and Information Technologies, Satbayev University, Almaty 050013, Kazakhstan; e.andekin@satbayev.university (E.A.); o.bekmurat@satbayev.university (O.B.); v.serbin@satbayev.university (V.S.); y.chinibayev@satbayev.university (Y.C.)
[2] Department of Intelligent Telecommunications, International Information Technologies University, Almaty 050060, Kazakhstan; k.sansyzbai@iitu.edu.kz
[3] Department of Mobilization Training of the Armed Forces, Other Troops and Military Formations, National Defense University of the Republic of Kazakhstan, Astana 020000, Kazakhstan; baurzhan.abzhanov@inbox.ru
[4] Department of Automation and Electric Power Engineering in Transport, International University of Transportation and Humanities, Almaty 050063, Kazakhstan
* Correspondence: shihgis070@gmail.com (S.K.); ayauka959@gmail.com (A.O.)

Abstract

Global university rankings differ in methodology, yet their published positions are frequently compared across systems and monitored over time. This study examines two dimensions of ranking stability: cross-sectional consistency across ranking systems and temporal persistence within QS, while assessing whether machine-learning models improve predictive accuracy beyond a persistence baseline. The analysis included 57 universities from 20 countries and seven regions. Cross-ranking consistency among QS-2026, THE-2026, and ARWU 2025 was evaluated using pairwise rank correlations, bootstrap confidence intervals, and normalized percentile positions. Temporal predictability was assessed on a balanced panel of 40 universities with QS positions for 2021–2024 and 2026 using an expanding-window design, a persistence baseline, linear regression, Random Forest, XGBoost, and LightGBM. Pairwise correlations were positive within observed overlaps, but the ranking systems were not interchangeable. The persistence baseline achieved the lowest development MAE. On the final two-year holdout, Random Forest (MAE 5.073) did not significantly outperform persistence (MAE 5.150; $p = 0.902$), and the latest QS position was the only predictor with consistently positive importance. Historical persistence errors bracketed the 2024–2026 result, indicating that predictability was driven mainly by ranking persistence, whereas additional machine-learning complexity provided limited benefits in this sample.

Keywords: university rankings; ranking stability; ranking predictability; cross-ranking consistency; temporal persistence; persistence baseline; temporal validation; machine learning; higher education analytics

Check for updates

Academic Editor: Domenico Ursino

Received: 29 June 2026
Revised: 13 August 2026
Accepted: 23 September 2026
Published: 9 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Global university rankings influence institutions' reputations, their international visibility, student recruitment, academic partnerships, and strategic planning. At the same time, these rankings should not be viewed as interchangeable indicators of educational

quality. Their results depend on different sets of indicators, weightings, data sources, and methodological priorities [1–5].

The QS World University Rankings, Times Higher Education, and the Academic Ranking of World Universities are among the best-known global ranking systems. QS places significant emphasis on reputation, citations, graduate employability, and internationalization. THE combines indicators of teaching, research environment, research quality, industry engagement, and international activity. ARWU is more focused on scientific output and major academic achievements [6–12]. Because of these differences, the same university may occupy markedly different positions in the three rankings.

Recent studies show that machine learning methods are already being used to analyze and predict university rankings. Rocha and Figueira [13] investigated the relationship between universities' ranking positions and their social-media strategies using text analysis, LSTM, and Random Forest. Their work addressed a related problem of digital activity, but did not directly predict future official ranking positions. Udupi et al. [14] applied exploratory data analysis and regression-based machine learning methods to predict global university rankings based on QS-related indicators. Wardley et al. [15] analyzed data from 49 Canadian universities for the years 2017–2021 and combined feature engineering methods with Random Forest, Gradient Boosting, and voting models [15]. Wisaeng and Muangmeesri [16] used Times Higher Education data on 1904 universities and evaluated hybrid feature selection alongside linear regression, Random Forest, and a multilayer perceptron.

Existing studies confirm that machine learning can be applied to university-ranking data. The remaining scientific question is therefore not whether ranking positions can be modeled, but how much of their apparent predictability is explained by the persistence of the ranking itself and whether more complex models provide stable incremental information beyond that persistence. At the same time, cross-ranking comparison raises a related question of cross-sectional consistency because QS, THE, and ARWU differ in methodology, scope, and coverage.

Accordingly, this study is organized around one overarching research question: To what extent are university rankings stable across ranking systems and over time, and do machine-learning models add predictive value beyond the persistence of the most recent QS position? This question is examined through three complementary subquestions: (RQ1) how consistently do QS, THE, and ARWU order the universities that appear in their pairwise overlaps; (RQ2) how persistent are QS positions across successive and two-year ranking transitions; and (RQ3) under temporally ordered evaluation, do machine-learning models provide a consistent advantage over a transparent persistence baseline?

The principal contribution of this study is empirical rather than methodological. The cross-ranking analysis evaluates one dimension of ranking stability, namely consistency across ranking systems, whereas the temporal QS analysis examines a second dimension, namely persistence and predictability across ranking editions. The results indicate that the most recent QS position contains the majority of the stable predictive signal within the analyzed sample, while the incremental contribution of more complex machine-learning models remains limited and temporally unstable. These findings are confined to the purposive sample considered in this study and should not be interpreted as evidence that official ranking methodologies can be replicated.

2. Related Work

2.1. University Ranking Systems and Their Impact

In recent years, global university rankings have become one of the most prominent tools for evaluating the performance of higher education institutions. Today, a university's ranking in the QS, THE, or ARWU affects not only its academic reputation but also its

internationalization efforts, the recruitment of international students, research collaboration, and participation in international projects [1,2]. For many universities, rankings are gradually becoming an integral part of strategic management and long-term planning.

The growing interest in rankings is linked to changes in the very model of higher education. Modern universities now compete not only at the local level but also within the global academic landscape [3,4]. As a result, indicators of publication activity, citation rates, research performance, and international collaboration are beginning to play a significantly greater role than they did in the past.

The QS World University Rankings, Times Higher Education (THE) World University Rankings, Academic Ranking of World Universities (ARWU), CWUR, and Webometrics are among the best-known global university ranking systems [6–10]. Despite their shared aim of evaluating universities, each system applies a distinct methodology and set of indicators. For example, QS emphasizes academic reputation, employer reputation, and internationalization, whereas THE uses a broader framework covering teaching, research environment, research quality, industry engagement, and international outlook. ARWU, in turn, focuses primarily on research output and major academic achievements [11,12].

Such differences mean that the same university can occupy significantly different positions in various rankings. This is due not only to differing sets of indicators but also to differences in the very principles used to calculate final scores [5]. For this reason, many researchers note that rankings cannot be viewed as fully interchangeable systems for evaluating university performance. Furthermore, the ranking methodologies themselves continue to evolve. The weights of indicators are being revised, new metrics are emerging, and the mechanisms for evaluating research and educational activities are changing. As a result, the analysis of rankings is gradually shifting from a simple comparison of positions to more complex tasks of monitoring, mathematical analysis, and forecasting.

2.2. Comparative and Mathematical Analysis of Rankings

One of the key areas of research in the field of university rankings is the application of mathematical methods to analyze the similarities and differences between various ranking systems [17–19]. Such studies emphasize that rankings reflect different aspects of university performance; therefore, a direct comparison of final positions does not always allow for an objective assessment of the actual standing of universities.

Studies on multidimensional ranking analysis examine approaches based on the use of multidimensional analysis, similarity metrics, and ranking aggregation [17–19]. Such methods make it possible to assess the degree of consistency between rankings and to identify stable patterns in the distribution of university positions.

Special attention is given to methods of ranking aggregation. The main idea behind such approaches is to combine several evaluation systems to obtain a more robust analytical model [19]. This is particularly important when different rankings use dissimilar indicators and various mechanisms for calculating final scores.

At the same time, most existing studies focus primarily on the static analysis of already published results. Many studies devote relatively little attention to analyzing changes in ranking positions over time and to building predictive models. Furthermore, the impact of changes in ranking methodologies on the robustness of analytical results is rarely examined.

2.3. Scientometrics and Metrics for Evaluating Scientific Activity

Contemporary university rankings are closely linked to the development of scientometrics and bibliometrics [20–25]. A significant portion of global rankings uses metrics such as citation rates, publication activity, and research performance of universities. As a result, the analysis of scientific metrics has become one of the key elements in evaluating university rankings.

Scientometric studies have examined citation indicators, the interpretation of citation counts, and the responsible use of quantitative metrics in the evaluation of scientific activity [20–25]. These studies show that bibliometric indicators can be used in institutional analysis; however, their significance depends on the data source, disciplinary context, and the purpose of the evaluation. It is emphasized that scientometric indicators allow for the formalization of part of universities' academic activity; however, their interpretation requires a cautious approach.

One of the most debated issues remains the dependence of results on the citation databases used and the mechanisms for indexing publications. Differences between databases can significantly affect citation metrics and, consequently, universities' final rankings in global rankings. Studies examining the limitations of quantitative metrics have played an important role in the development of scientometrics [22–24]. For example, the Leiden Manifesto highlights the need for more cautious use of quantitative indicators when evaluating scientific activity [22]. Similar conclusions are presented in works on citation analysis and research evaluation [23,24].

The limitations of indicators such as the h-index and citation counts are also actively discussed in the literature [23–25]. Despite their widespread use, such indicators do not always allow for an objective assessment of research quality, especially in interdisciplinary fields. Thus, the analysis of university rankings is gradually moving beyond a simple comparison of positions and requires the use of more complex methods for analyzing scientific information and research performance. Overall, an analysis of the research areas discussed shows that the tasks of monitoring and forecasting university rankings lie at the intersection of several fields, including scientometrics, mathematical analysis of rankings, machine learning, forecasting methods, and Big Data technologies. The conceptual relationship between these areas is illustrated in Figure 1.

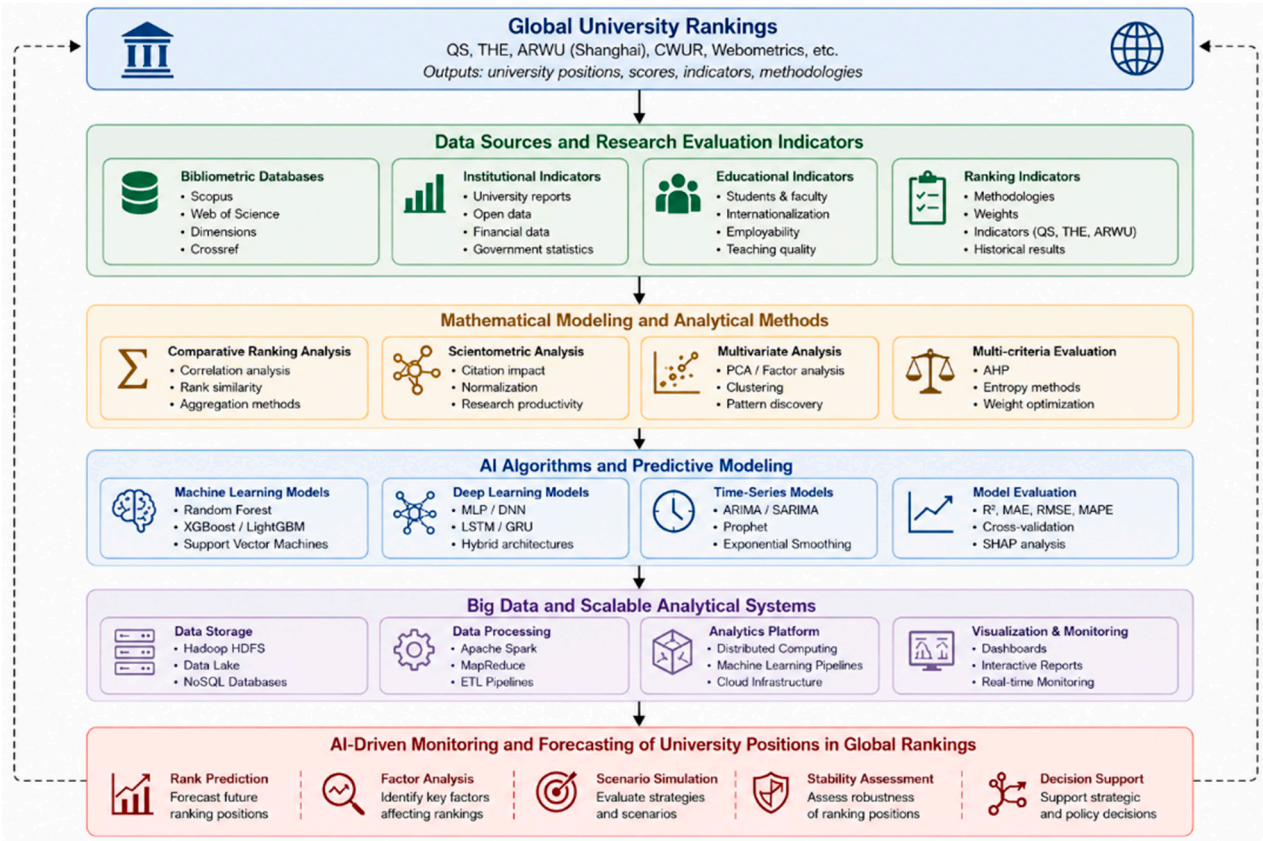

Figure 1. Conceptual framework of research areas related to the monitoring and forecasting of university rankings.

2.4. Fundamentals of Machine Learning and Predictive Modeling

The development of machine learning methods has significantly changed approaches to the analysis of multidimensional data. In recent years, AI algorithms have been actively used in forecasting, business intelligence, and the processing of large datasets [15,26–31].

The theoretical foundation of modern machine learning is rooted in statistical learning and predictive modeling. These approaches can identify complex relationships among features and construct predictive models from historical data. Random Forest, Gradient Boosting, XGBoost, and LightGBM are widely used for tabular-data analysis [15,26,27]. Their popularity is driven by their robustness to noisy data, their ability to handle numerous predictors, and their effectiveness in modeling nonlinear relationships.

Such algorithms are particularly relevant for analyzing university rankings, as ranking systems use dozens of different indicators, some of which have complex nonlinear relationships. Several studies note that gradient boosting methods yield more stable results compared to classical statistical models.

Software tools for implementing machine-learning models have also advanced considerably. Scikit-learn (version 1.6.1) was used in this study for data preprocessing, model fitting, and evaluation [28].

At the same time, researchers also emphasize the need for cautious interpretation of machine learning results [31]. This is particularly important in the context of university rankings, as changes in universities' rankings may depend not only on internal indicators but also on changes in the methodology of the rankings themselves.

2.5. Existing Research on Forecasting University Rankings

Machine-learning-based analysis of university rankings has developed more noticeably in recent years. Rocha and Figueira examined the relationship between higher education institution rankings and social-media strategies for 18 institutions [13]. Their analysis included topic categorization, LSTM [32], and Random Forest methods. However, the predicted outcomes were related mainly to social-media activity rather than to a future official university-ranking position. Tabassum et al. [33] developed a system for predicting university rankings based on historical data from Times Higher Education. The evaluation of the results was based on the deviation of the predicted rankings, recall, and accuracy at various acceptable error thresholds. The study demonstrates the possibility of forecasting future ranking results based on historical observations; however, it does not provide directly comparable values for MAE, RMSE, R^2, or Spearman's correlation coefficient for the QS time-series forecasting task.

Udupi et al. applied exploratory data analysis and machine-learning regression techniques to QS-related indicators [14]. Their study demonstrated that regression models can be used to estimate global university rankings. However, chronological expanding-window validation and direct comparison with a persistence baseline were not the main focus of the study.

Wardley et al. used observations from 49 Canadian universities covering the period from 2017 to 2021 [15]. The authors applied feature-engineering procedures together with Random Forest, Gradient Boosting, and voting-based models. The study provides useful evidence from a national university system, although its sample structure and classification-oriented evaluation differ from the temporal QS forecasting design used in the present work.

Wisaeng and Muangmeesri analyzed Times Higher Education data for 1904 universities [16]. Their framework combined hybrid feature selection with linear regression, Random Forest, and multilayer perceptron models. The study focused mainly on feature selection and predictive accuracy. The present work addresses a narrower question:

whether models trained only on previous ranking editions retain their performance on a later temporal holdout and provide a meaningful improvement over a persistence baseline. These studies show that university-ranking prediction is no longer an unexplored research topic. At the same time, the published evidence remains methodologically heterogeneous because the studies use different ranking systems, samples, prediction targets, features, and validation procedures. Their reported accuracy values therefore should not be compared directly. Table 1 summarizes the main methodological differences and specifies the contribution of the present study.

Table 1. Comparative positioning, reported evaluation results, and methodological comparability of related university-ranking studies.

Study	Data and Prediction Target	Methods and Validation	Reported Evaluation Results	Comparability with the Present Study
Rocha and Figueira [13]	Social-media strategy and ranking context for 18 higher education institutions	Topic analysis, LSTM, and Random Forest	No directly comparable official-rank MAE, RMSE, R^2, or Spearman's ρ was reported	Different prediction target; not a forecast of a future official QS position
Tabassum et al. [33]	Historical THE data; 2011–2015 used for model development and 2016 used for evaluation	University-ranking prediction evaluated through rank deviations and threshold-based measures	Rank-deviation results, recall, and accuracy under different acceptable deviation thresholds; no directly comparable MAE or RMSE	Different ranking system and evaluation measures
Udupi et al. [14]	QS-related university data; 1225 observations	Regression-based machine learning with a random train–validation–test split	RMSE = 0.448; MAE = 0.382; MAPE = 154.19%; R^2 = 0.841	Different target scale and non-temporal validation; numerical values are not directly comparable
Wardley et al. [15]	49 Canadian universities, 2017–2021	Feature engineering, Random Forest, Gradient Boosting, and voting models	No directly comparable QS temporal-holdout metrics were reported on the ranking-position scale used in the present study	National sample, different prediction design, and different evaluation setting
Wisaeng and Muang-meesri [16]	THE data for 1904 universities	Hybrid feature selection, linear regression, Random Forest, and multilayer perceptron	No directly comparable QS temporal-holdout results were reported on the present ranking-position scale	Different ranking system, sample structure, predictors, and validation procedure
Present study	57 universities; balanced QS forecasting panel of 40 universities	One-year expanding-window development; exploratory two-year QS 2026 extrapolation; persistence and ML comparators	Persistence: development MAE 4.512; Random Forest: 4.680. Final QS 2026: RF 5.073 vs. persistence 5.150 (p = 0.902). Latest QS rank was the only stable predictor	Top-heavy panel; two-year final horizon; incremental predictive value assessed beyond persistence

Note. The numerical results are presented as they appear in the respective studies. They should not be interpreted as a direct ranking of model quality, as the studies differ in terms of the target variable, outcome scale, sample composition, rating system, and validation procedure. If a study did not report MAE, RMSE, R^2, or Spearman's correlation coefficient for a comparable official position forecasting task, the results are marked as directly incomparable.

The contribution of the present study is to characterize ranking stability and predictability rather than to present machine learning as inherently superior. The cross-ranking component quantifies how consistently overlapping universities are ordered across QS, THE, and ARWU, with explicit uncertainty estimates and normalized positions. The temporal component then asks how much of QS predictability is explained by persistence in the latest published position and whether machine-learning models add stable incremental

predictive value. Expanding-window validation, a persistence baseline, paired error tests, and permutation-based importance are used to answer this substantive question.

2.6. Scalability as a Future Extension

Distributed data-processing approaches, including the MapReduce programming model and Apache Spark, are discussed in this study only as potential technologies for future scaling; neither MapReduce nor Apache Spark was implemented or used in the empirical analysis [34–36]. Hybrid processing architectures may become relevant for future university-monitoring systems that integrate long time series, publication records, citation data, and regularly updated institutional metrics [37,38]. However, the dataset used in the present experiment comprised only 57 universities and 44 variables. No distributed platform or computing cluster was implemented, and no scalability, throughput, latency, or infrastructure comparison was performed. Big Data technologies are therefore discussed only as a possible direction for future scaling and not as an experimentally validated component of the present study.

2.7. Forecasting Theory and Time Series Methods

Modern forecasting problems rely heavily on time series analysis methods [16,39]. Classical statistical forecasting approaches continue to be actively used in contemporary research. Of particular significance is the study by the M4 Competition, which demonstrated that the quality of forecasting depends not only on the complexity of the model but also on the correctness of data processing and the choice of analytical strategy [16]. The work by Hyndman and Athanasopoulos examines the main methods of time series forecasting and approaches to assessing model quality [39].

Such studies provide an important foundation for developing forecasting models used in the analysis of university rankings.

2.8. Summary of the Research Gap

Previous studies have examined cross-ranking comparison and university-ranking forecasting, but these areas are often treated separately. A useful unifying perspective is ranking stability: cross-sectional consistency asks whether different ranking systems order the same universities similarly, whereas temporal persistence asks how strongly a university's later position can be anticipated from its recent position. The research gap addressed here is therefore not simply forecasting accuracy, but whether machine-learning complexity adds reproducible predictive information beyond this temporal persistence while cross-ranking uncertainty is quantified explicitly [40].

This study does not claim to overcome the limitations of small ranking datasets or to provide a universally generalizable forecasting model. Instead, it uses a purposive panel of 40 predominantly highly ranked universities as an exploratory test bed for three linked questions: cross-ranking consistency, temporal persistence, and the incremental predictive value of machine learning relative to a persistence baseline. This framing places the substantive behavior of ranking positions—not model complexity—at the center of the analysis.

3. Materials and Methods

3.1. Overall Study Design

The study was designed around the overarching question of ranking stability and predictability and comprised two complementary analytical components. The first examined cross-sectional consistency by comparing the published positions of universities in QS 2026, THE 2026, and ARWU 2025 within the available pairwise overlaps. The second examined

temporal persistence and predictability within QS using lagged ranking information and a limited set of institutional characteristics. Together, the two components assess whether ranking positions are relatively stable across systems at a given point in time and across successive editions.

The components were kept analytically separate because they address different dimensions of the overarching question and rely on different information sets. Variables used for descriptive and cross-ranking analysis were stored separately from predictive variables, and all predictive preprocessing was performed within the corresponding temporal training window. This separation prevented current ranking outcomes or target-period component indicators from entering the forecasting models.

Operationally, RQ1 was addressed using pairwise rank correlations, uncertainty estimates, absolute rank differences, and normalized percentile positions. RQ2 was addressed by examining QS position changes across consecutive and two-year transitions and by using the most recent QS position as a persistence benchmark. RQ3 was addressed by comparing the persistence baseline with linear regression and Random Forest implemented in scikit-learn (version 1.6.1), XGBoost implemented using XGBoost (version 3.4.1), and LightGBM implemented using LightGBM (version 4.6.0), under expanding-window temporal validation. QS was used for the temporal component because it contained the most complete historical series for 2021–2024 and the observed 2026 outcome [41].

The analysis was not intended to reproduce the proprietary calculation procedures of QS, THE, or ARWU. Ranking positions were treated as observed outcomes. Likewise, permutation importance was used only to characterize the predictive dependence of the fitted model on its inputs; it was not used to infer causal effects of institutional characteristics on university performance or ranking outcomes.

The general logic of the study is presented in Figure 2.

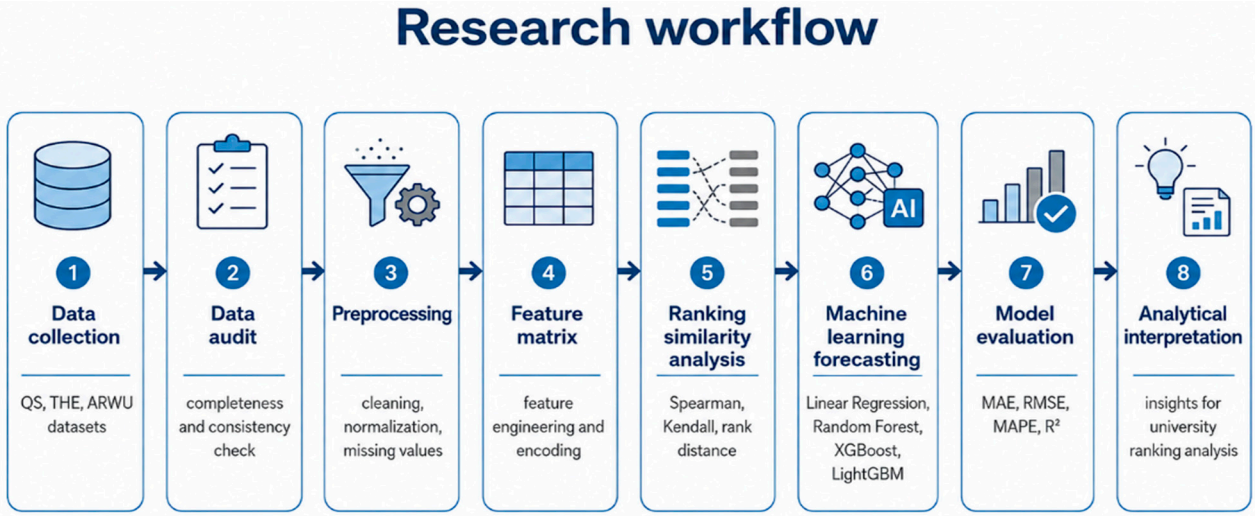

Figure 2. The diagram reflects the analytical procedures actually carried out in the study; distributed processing is considered only as a possible direction for further development.

3.2. Data Sources and the Formation of the Initial Sample

The empirical basis of the study was the structured dataset world_university_rankings_2021_2026_extended_compiled.csv, compiled by the authors to compare the rankings of universities in the QS World University Rankings, Times Higher Education (THE), and the Academic Ranking of World Universities (ARWU). The dataset included 57 universities and 44 variables and covered institutions from 20 countries and seven geographic regions, including 42 public and 15 private universities. The composition of the dataset and the completeness of ranking coverage are presented in Table 2.

Table 2. Composition and completeness of the initial dataset.

Indicator	Value
Dataset name	world_university_rankings_2021_2026_extended_compiled.csv
Dataset preparation	Compiled by the authors from publicly available ranking and institutional sources
Validation status	Internal consistency checks performed by the authors; no independent external audit
Number of universities	57
Number of variables	44
Number of countries	20
Number of regions	7
Public universities	42
Private universities	15
Universities with QS rank 2026	56
Universities with THE rank 2026	55
Universities with ARWU rank 2025	19
Universities represented in all three rankings	18
Universities included in the QS 2026 forecasting sample	40
Universities excluded from forecasting because of incomplete QS history	17

The dataset was based on the ranking editions listed in the source information as QS 2026, published in June 2025; THE 2026, published in October 2025; and ARWU 2025, published in August 2025. It included available overall rankings, total scores, and component indicators from QS and THE, as well as overall rankings and total scores from ARWU. The data were obtained from publicly available ranking materials and then consolidated into a single analytical table.

The authors do not claim ownership of the original QS, THE, or ARWU ranking data or of the associated ranking names, published tables, methodologies, or other proprietary materials. The original source materials and associated intellectual-property rights remain with Quacquarelli Symonds (QS), Times Higher Education (THE), and ShanghaiRanking Consultancy (ARWU), respectively. The author-compiled analytical dataset was created for research purposes by consolidating values available from the cited source materials; it is not an official dataset of any ranking organization and should not be treated as a substitute for the original sources. Readers seeking the original ranking data, current terms of access, or permission for reuse should consult the official QS, THE, and ARWU channels directly.

Institutional variables included the university's name, country, region, type of institution, year founded, total student enrollment, percentage of international students, age of the university, and specific indicators of academic achievement, including the number of Nobel Prize and Fields Medal laureates. For descriptive and cross-ranking analysis, derived variables were also created to show a university's representation in all three rankings and its average position across the available ranking systems.

The resulting dataset supported two complementary analyses. First, available QS, THE, and ARWU observations were used to examine cross-sectional ranking consistency. Second, temporal persistence and predictability were examined on a balanced subsample of 40 universities with complete QS positions for 2021–2024 and an observed QS 2026

position. Variables reflecting the final QS 2026 outcome and other target-period ranking indicators were excluded from the historical predictive feature set. The sampling procedure and inclusion criteria are presented in Section Population, Inclusion Criteria, and Sample Selection.

An overview of the initial dataset is presented in Table 2.

As shown in Table 2, the QS and THE rankings provided the most comprehensive coverage in the initial sample, whereas ARWU data were available for only 19 universities. Therefore, ARWU was used primarily for cross-ranking comparisons and was not considered a primary data source for the forecasting experiment. The subsample for the QS 2026 forecast included 40 universities for which QS rankings were available for 2021, 2022, 2023, and 2024, as well as a known QS ranking for 2026. The remaining 17 universities were excluded from the forecasting experiment due to the absence of at least one required historical QS value. Where relevant data were available, they were retained in the descriptive and cross-ranking analyses.

Population, Inclusion Criteria, and Sample Selection

The authors compiled the analytical dataset based on publicly available university ranking tables and institutional information. The study utilized the 2026 edition of the *QS World University Rankings*, published in June 2025; the 2026 edition of the *Times Higher Education World University Rankings*, published in October 2025; and the 2025 edition of the *Academic Ranking of World Universities*, published in August 2025. The official ranking pages cited in sources [6,7,9] were accessed on 14 April 2026. The exact dates when individual values were entered into the working CSV file were not recorded separately. Historical QS rankings for 2021–2024 were obtained from the ranking pages listed in the "historical_data_sources" field. University names were standardized, after which entries from different sources were matched based on the institution's name and country.

The dataset was compiled by the authors and underwent an internal consistency check; no independent audit or certification by an external organization was conducted. The internal review included assessing the completeness of variables, standardizing university names, matching records by organization name and country, identifying duplicate or ambiguous records in the consolidated table, and cross-checking the stored ranking values against the source pages cited in the dataset. These procedures were used to reduce inconsistencies that may have arisen during data collection and record consolidation. They should not be considered an independent external audit or official validation by ranking organizations.

Three working criteria were applied to ensure that universities were included in the final analytical dataset. First, the institution had to be unambiguously identified by name and country. Second, the university had to be ranked in at least one of the three systems under consideration: QS 2026, THE 2026, or ARWU 2025. Third, key institutional indicators necessary for descriptive analysis had to be available: country, region, type of institution, year founded, total number of students, percentage of international students, and specific indicators of academic achievement. Duplicate entries and institutions that could not be unambiguously matched across sources were excluded from the final dataset.

After applying these criteria, the final dataset consisted of 57 different universities from 20 countries and 7 regions, including 42 public and 15 private institutions. All 57 universities were ranked in at least one of the three ranking systems. After standardizing the names and matching the records, there were no duplicate records left in the final consolidated set. The surviving working materials contain the final dataset, but do not include a separate list of all the institutions considered prior to its compilation. Therefore, it is impossible to determine the exact number of universities that were initially reviewed and

excluded. Consequently, the resulting dataset should be viewed as a targeted analytical sample rather than a random or statistically representative sample of the entire population of QS, THE, or ARWU universities.

All 57 universities were included in the descriptive and cross-ranking analysis, provided that the necessary indicators were available. For the forecasting experiment, a stricter requirement for data completeness was applied. A university was included in the subsample for the QS 2026 forecast only if QS rankings for 2021, 2022, 2023, and 2024 were available, as well as a known QS ranking for 2026, which was used as the target variable. Forty universities met this condition. The remaining 17 institutions were excluded from the forecasting experiment due to the absence of at least one required historical QS ranking. However, they were retained in the descriptive and cross-ranking analysis if the relevant data were available.

Thus, the forecast subsample was formed based on the completeness of the historical QS data series, rather than by country, region, university type, ranking position, or model results. The study was exploratory in nature, and its results describe patterns observed within the analytical sample, primarily within the subset of 40 universities with complete data. The conclusions drawn should not be generalized to the entire population of institutions represented in the QS World University Rankings. Information on the ranking coverage of each university, the availability of historical data, and inclusion in the predictive subsample is provided in Table A1.

Table A1 presents details for each of the 57 universities included in the analytical dataset, including country, region, type of institution, availability of QS, THE, and ARWU data, completeness of the QS historical series, and the university's participation in pairwise cross-ranking and predictive analyses.

3.3. Data Completeness and Sample Limitations

The completeness of the data varied depending on the ranking system and the purpose of the analysis. The 2026 QS rankings were available for 56 of the 57 universities, while the 2026 THE rankings were available for 55 institutions. ARWU's 2025 coverage was notably limited: ranking data were available for only 19 universities. Eighteen institutions were represented in all three rankings simultaneously. Therefore, ARWU data were used primarily for pairwise cross-ranking comparisons and did not serve as the main basis for constructing the predictive model.

To analyze QS trends and conduct the forecasting experiment, a university had to have all the observations used by the model: QS rankings for 2021, 2022, 2023, and 2024, as well as a known QS ranking for 2026, which was used as the target variable. Forty universities met this criterion. The remaining 17 institutions were excluded from the forecasting experiment due to the absence of one or more required QS values. Where relevant data were available, they were retained in the descriptive and inter-ranking analyses. Missing historical rankings were not artificially reconstructed or replaced with estimated values.

The limited sample size must be taken into account when interpreting the results. The estimates presented describe correlations and ranking trends within the analytical sample and cannot be considered evidence of a general pattern applicable to all universities listed in QS, THE, or ARWU. Comparisons involving ARWU should be interpreted with particular caution, as they were based on a significantly smaller subsample. The results of the predictive experiment also apply only to the group of 40 universities with complete observations and should not be generalized to the entire population of institutions included in the QS rankings.

3.4. Separation of Variables for Descriptive Analysis and Forecasting

Prior to preprocessing and model building, the variables were divided into two disjoint sets. This division was necessary because the overall dataset contained both historical data and indicators relating to the target period. Combining them into a single forecasting matrix could have resulted in the use of information that would not yet be available at the time the forecast was generated.

The first set included current ranking results, their component indicators, and derived inter-ranking variables. These data were used only for descriptive analysis and pairwise comparisons of QS, THE, and ARWU. They were not used as predictors of the 2026 QS ranking position. The list of variables for this part of the analysis is presented in Table 3.

Table 3. Variables used solely for descriptive and cross-ranking analysis.

Variable Group	Variables	Purpose
Final ranking positions	qs_rank_2026, the_rank_2026, arwu_rank_2025	Description and pairwise comparison of university positions
Overall ranking scores	qs_score, the_score, arwu_score	Comparison of overall scores and final positions
QS 2026 component indicators	qs_academic_rep, qs_employer_rep, qs_faculty_student, qs_citations, qs_intl_faculty, qs_intl_students	Descriptive interpretation of the QS 2026 results
THE 2026 component indicators	the_teaching, the_research_env, the_research_quality, the_industry, the_intl_outlook	Descriptive interpretation of the THE 2026 results
Derived cross-ranking indicators	avg_rank_all3, ranked_in_all_3	Description of ranking coverage and average position across the available systems

The second set contained only the variables intended for the experiment to predict the QS 2026 ranking. Historical QS positions for 2021–2024 were used to generate lagged predictors corresponding to a specific time period. At each stage, the model received only the latest available QS ranking, the previous change in ranking (if the necessary data were available), an indicator of whether a change was available, and the length of the forecast interval. The four annual QS columns were not included in the model simultaneously. A limited set of institutional characteristics reflecting the university's scale and overall profile was added to these. The set of predictive variables is presented in Table 4.

Table 4. Predictors used in the experiment to predict the QS 2026 ranking.

Predictor Group	Variables	Role in the Forecasting Model
Lagged QS information	qs_rank_previous, qs_rank_change_previous, qs_rank_change_available, forecast_gap_years	Most recent available QS rank, preceding rank change, availability of the preceding change, and length of the forecast interval
Institutional history	founded	Stable historical characteristic of the institution
Institutional context	region, university_type	Regional and organizational characteristics

As shown in Tables 3 and 4, the ARWU rankings and scores, the THE rankings and scores, and the current QS component indicators were excluded from the predictive feature matrix. Therefore, missing ARWU values were not imputed for model training. ARWU data were used only in descriptive and pairwise inter-ranking analyses; the corresponding

sample sizes are reported together with the cross-ranking results in Section 4.2. The QS, THE, and ARWU indicators were not combined into a single predictive feature space.

For a preliminary analysis, the original wide table was transformed into a "university–year" panel. The variable "qs_rank_previous" reflected the most recent QS ranking available prior to the target issue. The variable "qs_rank_change_previous" was calculated as the difference between the two most recent available QS rankings; a negative value indicated an improvement. For the first training transition from QS 2021 to QS 2022, there was no earlier ranking available. Therefore, the change variable was assigned a default value of 0, and the qs_rank_change_available metric took the value 0. For all subsequent transitions, this metric was equal to 1. The forecast_gap_years variable was set to 1 for the QS 2023 and QS 2024 [42] editions and to 2 for the final QS 2026 edition.

The annual columns "qs_rank_2021", "qs_rank_2022", "qs_rank_2023", and "qs_rank_2024" were not included in the forecasting models simultaneously. Prior to training, they were converted into lagged predictors corresponding to a specific time period. At each stage, the model received only the most recent QS ranking available at that time, the previous change in ranking (if it could be calculated), an indicator of whether that change was available, and the length of the forecast interval. Therefore, the problem of multicollinearity—which arises when four closely related annual rankings are included simultaneously in a single forecasting matrix—does not apply to the revised model. Accordingly, feature interpretation was based on the derived predictors listed in Table 4 rather than on the individual annual QS ranking columns.

Variable institutional indicators for which annual observation dates were missing—including total student enrollment and the proportion of international students—were excluded from the main time-series model. This restriction was applied to avoid including information that might not have been available during the early stages of the time series.

3.5. Data Preprocessing

Preprocessing was performed separately at each stage of the evaluation. Only the predictors listed in Table 4 were included in the prediction matrix. The current ranking results, component indicators, and inter-ranking variables presented in Table 3 were used for descriptive and inter-ranking analysis but were not fed into the prediction models.

Numeric predictors included qs_rank_previous, qs_rank_change_previous, qs_rank_change_available, forecast_gap_years, and founded. They were standardized using the mean and standard deviation calculated solely from the corresponding training time window. The categorical variables region and university_type were converted using one-hot encoding. Although scale normalization is not mandatory for tree-based models, a uniform preprocessing scheme was used to ensure consistent data preparation across all models considered, including linear regression.

The predictive subsample included 40 universities and contained complete values for all predictors listed in Table 4. After transforming the data into a "university–ranking edition" panel format, the analytical dataset included 160 records and contained no missing values among either the predictors or the target variable. Therefore, in the presented predictive experiment, no values were actually imputed. The Simple Imputer component was retained in the Pipeline as a safeguard. For numerical variables, it was configured to use the median, and for categorical variables, the most frequent category; however, no observations in the analyzed sample were altered.

All preprocessing operations were performed using the Pipeline framework implemented in scikit-learn (version 1.6.1). At each time step, imputation rules, means, and standard deviations of numerical features, as well as the levels of categorical variables, were determined solely based on the corresponding training set. The pre-configured trans-

formations were then applied to the test set without recalculating the parameters. The training and test windows were defined separately for each temporal evaluation stage and are described in Section 3.7.

The target variable was retained on the original QS ranking scale and was not standardized or otherwise transformed. Therefore, model predictions, MAE, and RMSE were interpreted directly in terms of ranking positions. This procedure also prevented information from the test set from being incorporated into feature preprocessing or model training.

3.6. Mathematical Methods for Comparing Rankings

To compare QS, THE, and ARWU, we used methods of rank correlation and distance estimation between university positions. Since university rankings are ordinal in nature, ordinary linear correlation is not always sufficient. Therefore, the focus was on Spearman's and Kendall's coefficients. These methods allow us to assess not only the numerical proximity of values but also the degree of consistency in the order of institutions across different ranking systems.

Spearman's [43] rank correlation coefficient was calculated as follows:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{1}$$

where ρ is Spearman's rank correlation coefficient; d_i is the difference between the ranks of the same university in the two rankings, and n is the number of universities included in both ranking systems being compared.

Kendall's [44] coefficient was used as an additional measure of rank consistency:

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)} \tag{2}$$

where τ is Kendall's coefficient, C is the number of consistent pairs, D is the number of inconsistent pairs, and n is the number of items in the comparison. This metric allows us to assess the extent to which two ranking systems order universities in the same way.

All comparisons were performed separately for each pair of ranking systems using observations for which data were available in both systems. Therefore, the number of universities differed across the pairwise comparisons. Two-sided p-values were calculated for Spearman's and Kendall's rank correlation coefficients. The uncertainty of the estimates was characterized using 95% confidence intervals obtained from 20,000 paired bootstrap resamples at the university level. For reproducibility, the random seed was fixed at 42.

Because QS, THE, and ARWU publish ranking lists of different sizes, the comparison of absolute positions was supplemented with normalized percentile positions. For university i in ranking system j, the normalized position was calculated as:

$$P_{ij} = 100\left(1 - \frac{r_{ij} - 1}{N_j - 1}\right) \tag{3}$$

where r_{ij} is the published position of university i in ranking system j, and N_j is the total number of institutions included in the corresponding ranking edition. A value of 100 corresponds to the first position, whereas values approaching 0 indicate positions near the lower end of the published ranking list. The calculations used $N = 1501$ for QS 2026, $N = 2191$ for THE 2026, and $N = 1000$ for ARWU 2025, according to the corresponding official ranking editions.

For each ranking pair, the mean absolute difference between the published positions and the mean absolute difference between the normalized percentile positions were cal-

culated. The latter measure was expressed in percentage points. Since conversion to the percentile scale is a monotonic transformation, Spearman's and Kendall's rank correlation coefficients remain unchanged after normalization. The normalized measure was introduced to provide a more appropriate comparison of ranking systems with different levels of coverage.

Additionally, the absolute distance between the positions of the same university in the two rankings was calculated:

$$D_i = \left| rank_i^{(A)} - rank_i^{(B)} \right| \tag{4}$$

where $rank_i^{(A)}$ and $rank_i^{(B)}$ are the positions of the university in the rankings, respectively.

The average distance between the two ranking systems was calculated as follows:

$$\overline{D} = \frac{1}{n} \sum_{i=1}^{n} \left| rank_i^{(A)} - rank_i^{(B)} \right| \tag{5}$$

The smaller the value of $\overline{D}$, the closer the two ranking systems place the same universities. If the value of the mean deviations is large, this may indicate significant methodological differences between the rankings. Such an analysis is important because QS, THE, and ARWU use different sets of indicators and evaluate reputation, research, and international indicators differently.

Cosine similarity can also be used to analyze the multidimensional similarity of universities:

$$\cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|} \tag{6}$$

where A and B are the vectors of normalized indicators for two universities. This approach allows us to compare not only the final rankings but also the overall profiles of universities across several criteria.

Machine Learning Algorithms. To predict ranking positions, we used a persistence baseline, linear regression, Random Forest, XGBoost, and LightGBM [26,27]. The persistence baseline assumes that a university's position at the target edition is equal to its most recent observed QS position. It serves as the primary benchmark for testing whether additional model complexity provides incremental predictive value.

The persistence model was used as the baseline for comparison. It assumes that the university's future ranking position is equal to its most recent available position:

$$\hat{y}_{i,t+1} = y_{i,t} \tag{7}$$

where $\hat{y}_{i,t+1}$ is the predicted university ranking i for the next available period, and $y_{i,t}$ is the last known ranking. In this study, the 2024 QS rank can serve as such a baseline when forecasting the 2026 QS rank.

Linear regression was used as a simple, interpretable model:

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{8}$$

where $\hat{y}_i$ is the predicted ranking position of the university, i; x_{ij} is the value of the feature j for the university i; β_0 is the intercept; β_j is the model coefficient for the feature j, and p is the number of features.

Random Forest was used to account for nonlinear relationships between predictors, while XGBoost and LightGBM were considered as alternative tree-based gradient-boosting

methods. Interpretation was not based on impurity-derived feature importance because such estimates can be unstable when predictors are correlated and can favor variables with many possible split points. Instead, the predictive dependence of the selected Random Forest model on each original predictor was assessed using grouped permutation importance on temporally held-out samples.

Instead, the predictive dependence of the selected Random Forest model on each original predictor was assessed using grouped permutation importance on temporally held-out samples. The procedure is described in Section Grouped Permutation Importance and Stability Assessment. These values quantify incremental predictive information within the fitted model and sample; they should not be interpreted as estimates of institutional influence or causal effects on ranking outcomes.

The forecasting models and evaluation metrics used in the study are summarized in Table 5.

Table 5. Machine-learning models and metrics used for evaluating forecasting performance.

Component	The Method Used	Purpose of Application
MAE	Mean Absolute Error	Primary measure of prediction error and primary criterion for model and hyperparameter selection
RMSE	Root Mean Squared Error	Complementary measure giving greater weight to larger prediction errors
MAPE	Mean Absolute Percentage Error	Supplementary descriptive percentage error; not used for model or hyperparameter selection
R^2	Coefficient of Determination	Additional descriptive measure of explained variation on the final temporal holdout
Spearman's ρ	Spearman's Rank Correlation Coefficient	Assessment of preservation of the relative ordering of universities

As shown in Table 5, model evaluation was not limited to a single error measure. Both numerical forecasting accuracy and preservation of the relative ordering of universities were considered. This was done because, in ranking analysis, it is important not only to achieve an exact numerical match in position but also to preserve the overall order of the universities.

MAPE was retained only as a supplementary descriptive measure and was not used for model or hyperparameter selection. Because numerical rank appears in the denominator, percentage errors can be disproportionately large for universities near the top of the ranking. MAE was the primary selection criterion, while RMSE and Spearman's rank correlation were complementary measures. R^2 was reported descriptively for the exploratory QS 2026 two-year extrapolation [45,46]. The principal model-comparison results are reported in Section 4.4.

Statistical Assessment of Changes in QS Positions

To assess changes in QS positions between the 2024 and 2026 editions, the analysis included only universities for which ranking positions were available in both years. For university i, the paired change in ranking position was defined as:

$$\Delta r_i = r_{i,2026} - r_{i,2024} \tag{9}$$

where $r_{i,2024}$ and $r_{i,2026}$ denote the published QS positions of the same university in the 2024 and 2026 editions, respectively. Since a smaller numerical rank represents a stronger

position, a negative value of Δr_i indicates improvement, a positive value indicates a decline, and a value of zero indicates no change.

The descriptive analysis included the mean and median change, as well as the numbers of universities whose positions improved, remained unchanged, or declined. Because the paired differences were not assumed to follow a normal distribution, a two-sided Wilcoxon signed-rank test was used to examine whether the observed changes were systematically centered away from zero. Universities with zero differences were retained in the descriptive counts but did not contribute to the signed-rank statistic [47].

Uncertainty in the mean change was estimated using 20,000 paired bootstrap resamples at the university level. In each resample, universities were selected with replacement while their QS 2024 and QS 2026 positions remained paired. The 2.5th and 97.5th percentiles of the bootstrap distribution were used as the limits of the 95% confidence interval. A random seed of 42 was used for reproducibility. The statistical procedure and its results are summarized in Table A2.

3.7. Model Training and Evaluation Taking Temporal Order into Account

Since the target variable pertained to a later QS edition, random train–test splitting, shuffled k-fold cross-validation, and randomly ordered block validation were not used for model evaluation. Instead, the expanding-window scheme presented in Table 6 was applied. At each temporal origin, the validation set contained the same 40 universities, and the QS 2026 observations remained unavailable until preprocessing rules, predictors, and model hyperparameters had been finalized.

Table 6. Expanding-window design used for time-aware evaluation.

Evaluation Origin	Training Information	Validation Target	Training Panel Observations	Validation Universities	Role
Origin 1	QS transitions available through 2022	QS 2023	40	40	Development validation
Origin 2	QS transitions available through 2023	QS 2024	80	40	Development validation
Final origin	QS transitions available through 2024	QS 2026	120	40	Exploratory two-year temporal extrapolation

Note. The number of training observations refers to university-transition records in the panel dataset. The QS 2026 observations were not used during model development or hyperparameter selection.

The QS data were transformed from a broad table of universities into a "university–year" panel format. Each panel record corresponded to a specific university and the time the forecast was generated. The target variable was the QS ranking in the next available edition of the rankings, and the predictors used were the most recent QS ranking available at that time and the institutional variables listed in Table 4. The analysis was conducted on the same balanced subsample of 40 universities for which QS rankings were available for 2021, 2022, 2023, 2024, and 2026. Using the same set of universities at each time step made the results of individual stages more comparable.

Next, an expanding time window approach was applied. At each step, the training set included only observations available up to the validation year. In the first stage, the model was trained on the transition from 2021 to 2022, after which the forecast for 2023 was evaluated. In the second stage, the training set was expanded to include data available through 2023, and validation was performed on the QS 2024 positions. In the final stage, all valid observations available through 2024 were used for training, while the QS 2026

positions were retained as a separate temporal test sample. The validation scheme is presented in Table 6.

To characterize ranking volatility across the available QS editions, the magnitude of position changes was calculated within the same balanced panel of 40 universities used for the time-aware model evaluation [5]. For each university i and transition from edition t_1 to edition t_2, the absolute change was defined as:

$$V_{i,t_1,t_2} = \left| r_{i,t_2} - r_{i,t_1} \right| \tag{10}$$

where r_{i,t_1} and r_{i,t_2} denote the published QS positions of university i in the earlier and later editions, respectively.

The mean absolute change for each transition was then calculated as:

$$\overline{V}_{t_1,t_2} = \frac{1}{n} \sum_{i=1}^{n} V_{i,t_1,t_2} \tag{11}$$

where n is the number of universities included in the balanced panel.

In this analysis, ranking volatility refers to the observed magnitude of changes in university positions rather than to variance estimated from a continuous time series. The mean absolute change, median absolute change, maximum absolute change, and corresponding 95% bootstrap confidence interval were calculated for the QS 2021–2022, QS 2022–2023, QS 2023–2024, and QS 2024–2026 transitions. The confidence intervals were estimated using 20,000 paired bootstrap resamples at the university level. A fixed random seed of 42 was used to ensure reproducibility. The QS 2024–2026 transition covered two calendar years and was reported descriptively, but it was not included in the test comparing the one-year transitions. The results are presented in Table A3.

The first three transitions covered one calendar year and were therefore compared using the Friedman test for related samples. When the overall test was statistically significant, paired Wilcoxon signed-rank tests with the Holm adjustment for multiple comparisons were used for post hoc analysis. The QS 2024–2026 transition covered two calendar years and was reported descriptively, but it was not included in the statistical comparison of the one-year transitions.

The QS 2023 and QS 2024 time series were used solely for model development and hyperparameter tuning. The Random Forest, XGBoost, and LightGBM configurations were evaluated using the same expanding time window scheme presented in Table 6. A full search of the specified parameter grid was performed, rather than a random search.

For Random Forest, 18 configurations were tested; for XGBoost, 32; and for LightGBM, 24. Each configuration was evaluated separately on the QS 2023 and QS 2024 test sets. The configurations were ranked based on the arithmetic mean of the two MAE values. RMSE and Spearman's rank correlation coefficient were considered as supplementary metrics but were not used as the primary selection criterion.

The full search ranges and selected configurations are presented in Table A4. The selected Random Forest configuration included 500 trees, a maximum depth of 4, at least 2 observations in the terminal node, and the use of all available features at each split. For XGBoost, 150 trees, a maximum depth of 2, a learning rate of 0.03, and full use of rows and features were selected. The LightGBM configuration included 150 trees, 5 leaves, a learning rate of 0.03, and at least 5 observations in each child node.

XGBoost and LightGBM were included as restricted comparison models, not because the available sample size was considered sufficient for unrestricted gradient boosting. In the search space presented in Table A4, the tree depth, number of leaves, and training rate were intentionally restricted. The selected XGBoost model used shallow trees with

a maximum depth of 2, and the selected LightGBM model used only 5 leaves. Random Forest was also restricted to a maximum depth of 4 and a minimum of two observations in the terminal leaf.

At successive temporal origins, the expanding training set contained 40, 80, and 120 university-transition records, as shown in Table 6. These records still represented the same 40 institutions and were not treated as observations from 120 independent universities. Accordingly, the ensemble-model results are interpreted as an exploratory comparison with simpler approaches rather than as evidence of broad generalizability.

After model selection, the hyperparameters were fixed. The models were retrained on all valid panel observations available through QS 2024 and evaluated once on the QS 2026 outcome. QS 2026 was not used in preprocessing, feature selection, hyperparameter tuning, or model selection. Linear regression and the persistence baseline required no hyperparameter tuning.

The working dataset did not contain QS 2025 positions, and these values were neither reconstructed nor imputed. Consequently, the final evaluation from QS 2024 to QS 2026 spans two calendar years, whereas the QS 2023 and QS 2024 development evaluations represent one-year transitions. The variable forecast_gap_years records this difference but does not eliminate the structural mismatch in forecasting horizon. Therefore, the QS 2026 result is treated throughout the manuscript as an exploratory two-year temporal extrapolation rather than as equivalent to a conventional one-period forecast.

To provide a horizon-matched robustness check without introducing additional model complexity, the persistence baseline was also evaluated descriptively over the historical two-year transitions QS 2021–QS 2023 and QS 2022–QS 2024 within the same balanced panel of 40 universities. Because the persistence forecast is simply the rank observed at the forecast origin, its MAE is exactly the mean absolute change between the origin and target editions. A leakage-free machine-learning backtest for QS 2021–QS 2023 would require model-training observations preceding 2021, which are not available in the dataset; therefore, we did not retrospectively fit machine-learning models using information that would postdate the forecast origin. The horizon-matched persistence results are reported in Table A6.

All preprocessing operations were configured separately within each training time window and then applied to the corresponding test year. This applied to numerical transformations and one-hot encoding of categorical variables. The means and standard deviations of numerical features, the levels of categorical variables, and other preprocessing parameters were calculated only based on the corresponding training time window and were not recalculated using the test set [48].

For the two stages of the preliminary time-based validation, the average value of the metric was calculated as follows:

$$\overline{M}_{\text{dev}} = \frac{M_{2023} + M_{2024}}{2} \tag{12}$$

where M_{2023} and M_{2024} are the metric values obtained from the QS 2023 and QS 2024 editions, respectively. The result from the QS 2026 edition was reported separately and was not included in this average.

To test whether the Random Forest model selected during development provided incremental predictive value beyond persistence, a paired comparison was conducted on the exploratory QS 2026 two-year extrapolation for the same 40 universities. For university i, the difference in absolute errors was calculated as follows:

$$d_i = |y_i - \hat{y}_{\text{RF},i}| - |y_i - \hat{y}_{\text{Baseline},i}| \tag{13}$$

where y_i is the actual ranking of university i in the QS 2026; $\hat{y}_{RF,i}$ is the ranking forecast obtained using the Random Forest model; $\hat{y}_{Baseline,i}$ is the forecast from the baseline model, based on the most recent available QS ranking. A negative value of d_i indicates a smaller absolute error for the Random Forest model, whereas a positive value indicates an advantage for the baseline model.

A 95% confidence interval for the mean pairwise difference was calculated based on 20,000 pairwise bootstrap resamples at the university level. Additionally, a two-sided pairwise permutation test was applied, involving 100,000 random sign permutations. The null hypothesis that the mean pairwise difference in absolute errors is zero was tested. Pairwise comparisons were conducted after the model configurations were finalized and were not used for model selection or for tuning the model's hyperparameters.

Grouped Permutation-Importance Estimation and Stability Assessment

To interpret the Random Forest model selected during the development phase, grouped permutation importance was calculated separately on the QS 2023, QS 2024, and QS 2026 test samples. The evaluation was performed only after training the model on observations available prior to the corresponding target ranking release. The results of this analysis were not used for feature selection, hyperparameter tuning, or model selection.

For predictor j, the permutation importance was defined as the increase in the mean absolute error on the test sample following a random permutation of the values of that feature:

$$I_j = \mathrm{MAE}_{\mathrm{permuted},j} - \mathrm{MAE}_{\mathrm{original}} \quad (14)$$

where I_j is the permutation importance of predictor j; $\mathrm{MAE}_{\mathrm{permuted},j}$ is the MAE value after permuting the values of this predictor; $\mathrm{MAE}_{\mathrm{original}}$ is the model's original MAE value before permuting the feature.

A positive value of I_j indicates that permuting the feature worsened predictive accuracy and, consequently, that the fitted model depended on that predictor in the corresponding temporal holdout. Values close to zero indicate little measurable incremental predictive contribution within that holdout. Negative values can occur when prediction error decreases slightly after permutation and were interpreted as instability rather than as negative substantive importance [49].

The permutation was applied to the original predictors, not to individual transformed columns. For the categorical variables "region" and "university_type", all one-hot columns corresponding to a single original variable were permuted together. This prevented the importance of a single categorical feature from being distributed across multiple dummy variables.

The uncertainty of the estimates was calculated using 2000 paired bootstrap-permutation resamples at the university level. In each iteration, universities in the test sample were first selected with replacement, after which a single predictor was permuted within the resulting sample. The 2.5th and 97.5th percentiles of the resulting distribution were used as the boundaries of the 95% confidence interval.

The stability of feature rankings was examined separately at the QS 2023, QS 2024, and QS 2026 temporal origins. After ranking the mean permutation-importance values for each pair of origins, Spearman's correlation coefficient was calculated. Features that were constant within a temporal holdout could not be evaluated using within-origin permutation importance and were marked as not estimable.

3.8. Conceptual Architecture for Future Scaling

The empirical analysis was implemented as a standard Python workflow for tabular data. It included data loading, completeness checks, preprocessing, construction of analyti-

cal variables, pairwise comparison of rankings, model training, temporal validation, and visualization. For a dataset of 57 universities and 44 variables, this workflow was sufficient.

In the future, a more comprehensive monitoring system could organize these operations into separate modules for data loading, validation, feature preparation, cross-ranking analysis, forecasting, and reporting. A modular structure could simplify regular data updates and the addition of new ranking editions or institutional indicators.

Distributed processing technologies may only become relevant when the system is scaled up to significantly larger longitudinal or publication datasets [34,35,38]. In this study, a Hadoop or Spark cluster was not implemented, and scalability, latency, throughput, and differences in computing infrastructures were not evaluated. Therefore, the presented architecture should be viewed as a conceptual direction for future scaling rather than as a validated experimental result.

3.9. Ethical and Methodological Considerations

The study utilized publicly available aggregated institutional-level data. The study did not involve human participants and did not involve the processing of personal, clinical, or other sensitive information at the individual level. Therefore, approval from an ethics committee and informed consent were not required. However, the results should be interpreted with methodological caution, as a ranking position does not constitute a comprehensive assessment of the quality of education or research and may change due to revisions in methodology, weightings, scope, and the composition of the institutions included in the ranking [22,24].

At the same time, the results must be interpreted with caution. University rankings are not a comprehensive or universal measure of educational quality. They reflect only those aspects of a university's activities that are included in the specific methodology of the ranking system. Furthermore, a change in a university's ranking may be due not only to its own performance but also to changes in methodology, the emergence of new participants in the ranking, a recalculation of weights, and shifts in the global competitive environment. These limitations are consistent with the general criticism of the overreliance on quantitative indicators in the evaluation of scientific and educational activities [24,26].

Therefore, the proposed mathematical and AI models should be viewed as analytical support tools rather than as a definitive mechanism for evaluating university performance. This approach helps avoid oversimplification and makes the results of the study more accurate from both a scientific and a practical standpoint.

4. Results

4.1. General Data Characteristics

In the first stage, the structure and completeness of the final analytical dataset were analyzed. The dataset included 57 universities that met the criteria described in Section Population, Inclusion Criteria, and Sample Selection and had a final ranking in at least one of the three ranking systems under consideration. All 57 institutions met this condition. However, the completeness of the data varied: QS 2026 rankings were available for 56 universities, THE 2026 rankings for 55, and ARWU 2025 rankings for 19 universities.

Since ARWU covered a significantly smaller portion of the sample, its data were used primarily for pairwise cross-ranking comparisons rather than as the basis for a predictive experiment. Therefore, the results in this section characterize only the analytical sample and should not be regarded as a description of the distribution of all universities represented in the QS, THE, or ARWU rankings.

The coverage of the analytical sample across the three ranking systems is shown in Figure 3.

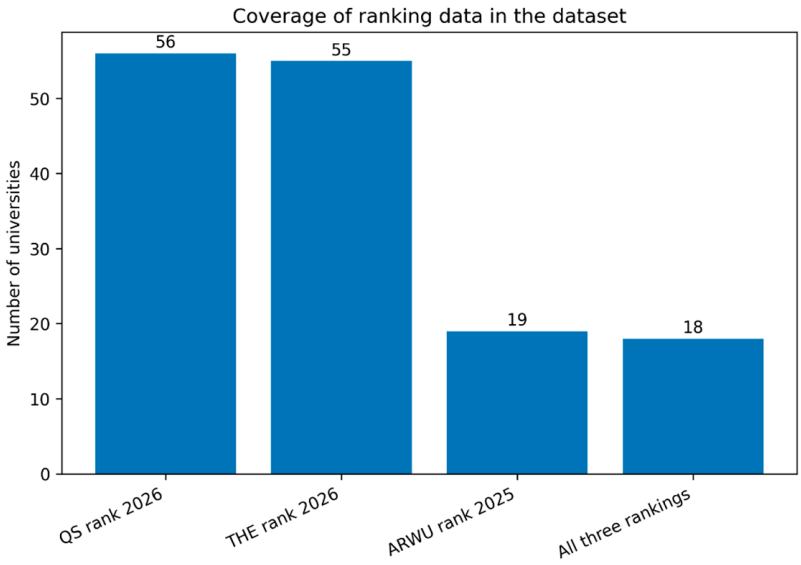

Figure 3. Coverage of the source dataset across the QS, THE, and ARWU ranking systems.

Summary statistics for the main ranking indicators are presented in Table 7.

Table 7. Descriptive statistics of the main ranking indicators.

Indicator	*n*	Average	Median	Minimum	Maximum
QS rank 2026	56	45.96	28.00	1	488
THE rank 2026	55	48.91	27.00	1	351
ARWU rank 2025	19	18.42	10.00	1	81
QS score	56	90.81	91.20	62.40	100.00
THE score	55	77.19	79.40	46.40	97.00
ARWU score	19	65.09	64.20	33.40	100.00

As shown in Table 7, QS and THE data were available for nearly the entire analytical sample, whereas the ARWU covered only 19 universities. The lower numerical value of the average ARWU ranking should not be interpreted as evidence that universities generally performed better in this ranking. Ranking systems publish lists of varying sizes, and the ARWU metric was calculated for a significantly smaller and more selective subset. Therefore, the subsequent cross-ranking analysis took into account both differences in absolute rankings and normalized percentile positions.

The QS 2026 sample was noticeably concentrated at the top of the ranking distribution. Of the 56 universities with an available QS 2026 ranking, 43 institutions—or 76.8%—were in the Top 50, and 52 institutions—or 92.9%—were in the Top 100. The median rank was 28, while the highest rank was 488. Thus, a small number of universities with lower rankings formed a long right tail, but the majority of institutions were concentrated at the top of the published rankings.

In the forecast subsample, the concentration was even more pronounced. Of the 40 universities with a complete QS history, 39 institutions—or 97.5%—were ranked in the Top 50, and all 40 universities were ranked within the Top 100. The median QS 2026 ranking was 19, and the lowest ranking in this subsample was 51. Concentration measures for both samples are presented in Table A5 and Figure A1.

This sampling structure narrows the range of ranking positions across which forecasting quality was assessed. Therefore, an MAE of about five positions should be interpreted as an estimate of error for a group of predominantly highly ranked universities, rather

than as the expected accuracy for the entire QS universe. The results obtained cannot be generalized to universities in significantly lower ranking ranges.

4.2. Pairwise Comparison of QS, THE, and ARWU

Pairwise comparisons were conducted only for universities that appeared in both of the ranking systems under consideration. Consequently, the sample sizes varied: the QS–THE comparison included 54 universities, the QS–ARWU comparison included 18, and the THE–ARWU comparison included 19. Spearman's and Kendall's correlation coefficients, their 95% bootstrap confidence intervals, two-tailed p-values, and measures of the difference in absolute and normalized percentile ranks are presented in Table 8. Point estimates of the coefficients and confidence intervals are shown in Figure 4.

Table 8. Pairwise comparison of QS, THE, and ARWU using rank correlations and normalized percentile positions.

Ranking Comparison	n	Spearman (ρ), 95% CI	(p_ρ)	Kendall (τ), 95% CI	(p_τ)	Mean Absolute Rank Difference	Mean Absolute Percentile Difference, p_p
QS 2026–THE 2026	54	0.815 [0.667, 0.904]	6.32×10^{-14}	0.642 [0.502, 0.758]	1.03×10^{-11}	20.796	1.087
QS 2026–ARWU 2025	18	0.810 [0.456, 0.952]	4.59×10^{-5}	0.634 [0.315, 0.856]	1.05×10^{-4}	6.944	0.681
THE 2026–ARWU 2025	19	0.819 [0.572, 0.936]	1.77×10^{-5}	0.641 [0.406, 0.839]	1.51×10^{-4}	8.842	0.941

Note. Confidence intervals were estimated using 20,000 paired bootstrap resamples at the university level. Normalized percentile positions were calculated using the published ranking-list sizes: QS 2026 ($N = 1501$), THE 2026 ($N = 2191$), and ARWU 2025 ($N = 1000$). Higher percentile values indicate stronger ranking positions. "p_p" denotes percentage points.

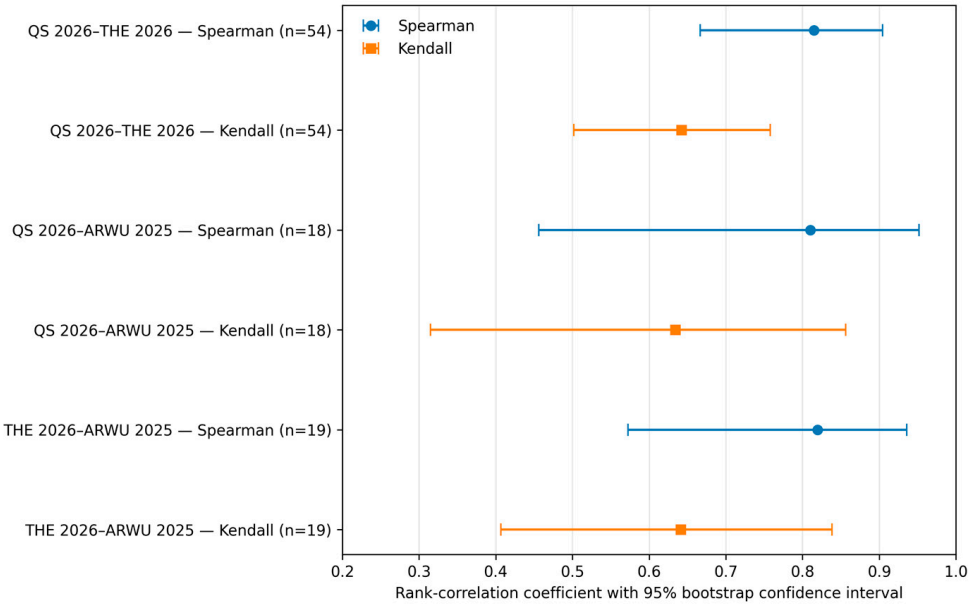

Figure 4. Spearman's and Kendall's rank correlation coefficients [43,44] with 95% bootstrap confidence intervals for pairwise comparisons of QS 2026, THE 2026, and ARWU 2025. Confidence intervals were calculated based on 20,000 paired bootstrap resamples at the university level.

The point estimates and 95% bootstrap confidence intervals for the pairwise rank correlation coefficients are shown in Figure 4.

In the limited subsets of universities that appeared in all of the compared rankings, Spearman's correlation coefficients ranged from 0.810 to 0.819, and Kendall's correlation coefficients ranged from 0.634 to 0.642. All corresponding two-tailed p-values were less than 0.001. This indicates a positive correlation in the ranking of universities within the available

pairwise samples. However, the precision of the estimates varied considerably. The QS–THE comparison included 54 universities and had a comparatively narrower confidence interval. Comparisons with ARWU were conducted on only 18 or 19 institutions, so the confidence intervals were significantly wider. For example, for the QS–ARWU pair, the 95% confidence interval for the Spearman correlation coefficient ranged from 0.456 to 0.952, and for the Kendall correlation coefficient, from 0.315 to 0.856. Therefore, the point estimates cannot be interpreted in isolation from their uncertainty.

The average absolute difference between QS and THE was 20.796 positions. After accounting for the different lengths of the published rankings, the corresponding difference was 1.087 percentage points. For QS–ARWU, the normalized difference was 0.681 percentage points, and for THE–ARWU, it was 0.941. This indicates that universities from the overlapping subsets were generally located in close proximity within the published ranking distributions. However, this result applies only to the selected institutions, which were primarily located at the top of the rankings.

Thus, the analysis does not demonstrate overall consistency among the complete QS, THE, and ARWU rankings. It merely shows the presence of a positive correlation within the available pairwise subsets. Results involving ARWU should be considered preliminary due to its limited coverage. The ranking systems also cannot be considered interchangeable, as they differ in methodology, overall coverage, and the rankings of individual universities [47,50].

4.3. Changes in QS Rankings Between 2024 and 2026

Changes in university rankings were assessed based on available QS data for 2024 and 2026. Rankings for both years were available for 44 universities. The change was calculated as the difference between the 2026 QS ranking and the 2024 QS ranking; therefore, a negative value indicated an improvement, a positive value indicated a decline in ranking, and a zero value indicated no change.

Of the 44 universities, 17 improved their rankings, 15 fell in the rankings, and the rankings of 12 institutions remained unchanged. The average change was −0.955 positions, and the median value was 0. The distribution of the observed changes is shown in Figure 5, and the results of the statistical analysis are presented in Table A2.

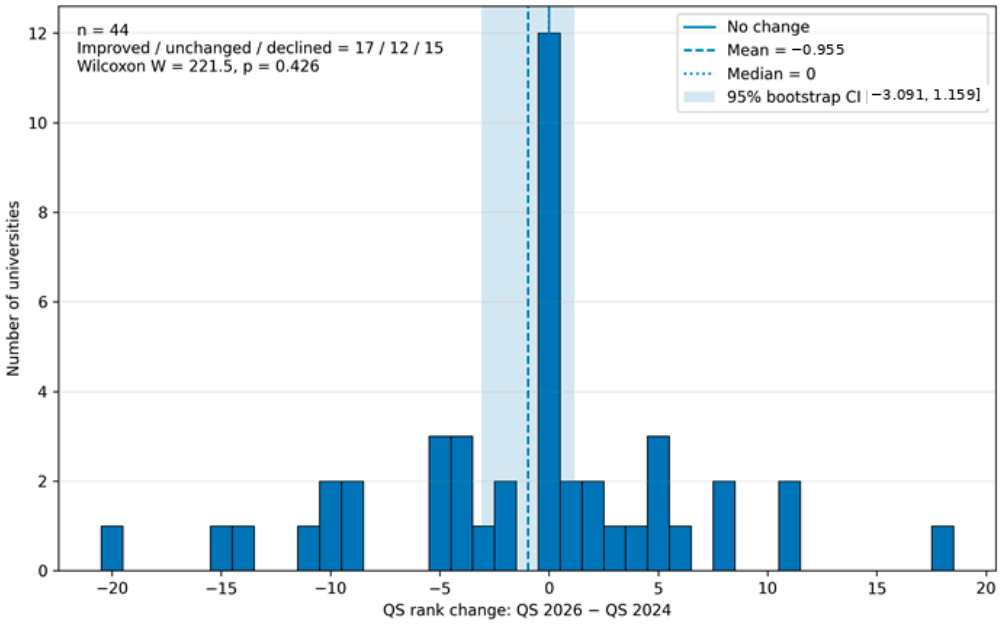

Figure 5. Distribution of paired changes in QS rankings between the 2024 and 2026 editions for 44 universities.

In Figure 5, the change was calculated as the QS 2026 ranking minus the QS 2024 ranking; negative values indicate an improvement, positive values indicate a decline, and zero indicates no change. The dashed line shows the mean change, the dotted line shows the median, and the shaded interval corresponds to the 95% paired bootstrap confidence interval for the mean. The two-sided Wilcoxon signed-rank test yielded W = 221.5 and $p = 0.426$.

The slightly negative mean did not confirm a systematic improvement between the two cohorts. Furthermore, the 95% paired bootstrap confidence interval for the mean change ranged from −3.091 to 1.159 positions and included zero. Therefore, the mean change should be considered a descriptive point estimate rather than evidence of a general directional change in QS positions.

At the same time, individual universities experienced changes of varying magnitude. The most notable improvements in the available sample were observed at Fudan University—rising from 50th to 30th place—the University of Hong Kong—rising from 26th to 11th place—and NTU Singapore—rising from 26th to 12th place. Conversely, UC Berkeley dropped from 10th to 28th place, UCLA from 29th to 40th, and LSE from 45th to 56th.

Such changes at the individual university level should not be interpreted as a direct improvement or deterioration in the quality of the institution. A university's published ranking may change due to its own performance metrics, revisions to the ranking methodology, changes in the weighting of indicators, the inclusion of new participants, and general dynamics in the competitive environment. The results apply only to the 44 universities for which paired observations are available and should not be generalized to the entire set of institutions in the QS ranking.

Ranking Volatility Across the Observed QS Transitions

In addition, we analyzed ranking volatility across the balanced sample of 40 universities used for the forecasting experiment. Unlike the analysis above, which examined the direction of change in the QS rankings over 2024–2026 for all 44 available paired observations, this analysis focused on the magnitude of change over several successive transition periods using a constant sample of universities. The results are presented in Table A3.

The mean absolute change was 2.425 positions between QS 2021 and QS 2022 and 2.300 positions between QS 2022 and QS 2023. A substantially larger change was observed between QS 2023 and QS 2024, for which the mean absolute change reached 6.725 positions. The corresponding median absolute changes were 2.0, 2.0, and 4.5 positions. Differences among the three one-year transitions were statistically significant according to the Friedman test, $\chi^2(2) = 12.434$, $p = 0.002$. Holm-adjusted post hoc comparisons showed no significant difference between QS 2021–2022 and QS 2022–2023 ($p = 0.311$), whereas the absolute changes for QS 2023–2024 were larger than those for QS 2021–2022 ($p = 0.0015$) and QS 2022–2023 ($p < 0.001$).

The mean absolute change from QS 2024 to QS 2026 was 5.150 positions. This value corresponds to a two-year calendar interval and should not be interpreted as directly equivalent to the one-year changes. Its 95% bootstrap confidence interval ranged from 3.600 to 6.825 positions. These results show that ranking movements were not equally large across editions. The variation also helps explain the changing performance of the persistence baseline across the evaluated temporal origins. Because the baseline predicts the target position using the latest available position, its MAE is numerically equal to the mean absolute change for the corresponding transition. Accordingly, the persistence MAE was 2.300 for QS 2023, 6.725 for QS 2024, and 5.150 for the exploratory QS 2026 two-year extrapolation.

As a horizon-matched robustness check, the persistence baseline produced a two-year MAE of 3.025 positions for QS 2021–QS 2023, 6.875 for QS 2022–QS 2024, and 5.150 for QS 2024–QS 2026. The corresponding median absolute errors were 2, 5, and 4 positions, respectively. Thus, the final two-year extrapolation had an error between the two historical two-year transitions rather than being directly comparable with the one-year development origins. This descriptive robustness analysis is summarized in Table A6.

4.4. Temporal Predictability and Model Performance

A temporally ordered forecasting experiment was conducted on the balanced subsample of 40 universities described in the Population, Inclusion Criteria, and Sample Selection subsection and Section 3.3. The raw rankings were transformed into a university–year panel. The models used only the lagged QS positions and institutional predictors listed in Table 4; target-period ranking outcomes and descriptive variables from Table 3 were excluded from the predictive matrix.

Model performance was assessed using the expanding-window scheme described in Section 3.7 and Table 6. QS 2023 and QS 2024 served as one-year development origins for model and hyperparameter selection. QS 2026 was evaluated only after all model configurations had been frozen and is reported as an exploratory two-year temporal extrapolation from QS 2024. The hyperparameter search ranges, the number of tested configurations, and the settings selected based on the results of the QS 2023 and QS 2024 time steps are presented in Table A4.

The results are presented in Table 9.

Table 9. Model performance under expanding-window development validation and the exploratory QS 2026 two-year temporal extrapolation.

Model	MAE, QS 2023	MAE, QS 2024	Mean Development MAE	MAE, QS 2026	RMSE, QS 2026	(R^2), QS 2026	Spearman's (ρ), QS 2026
Persistence baseline	2.300	6.725	4.512	5.150	7.311	0.652	0.872
Linear Regression	2.900	7.128	5.014	5.146	6.670	0.711	0.887
Random Forest	2.676	6.684	4.680	5.073	7.027	0.679	0.854
XGBoost	2.731	6.758	4.744	4.964	7.094	0.672	0.864
LightGBM	3.001	6.765	4.883	5.359	7.450	0.639	0.826

Note. QS 2023 and QS 2024 are one-year development evaluations used for model and hyperparameter selection. QS 2026 is an exploratory two-year temporal extrapolation from QS 2024 and was not used for hyperparameter tuning.

MAPE was not used to rank the models or select their hyperparameters and was therefore not included in the main comparison columns in Table 9.

The persistence baseline was the strongest benchmark during model development, with the lowest mean MAE across QS 2023 and QS 2024 (4.512 positions). Among the machine-learning models, Random Forest had the lowest mean development MAE (4.680), only 0.168 positions higher than the baseline. Random Forest was therefore retained as the development-selected machine-learning comparator for the paired analysis, not as a model already shown to be superior to persistence.

On the exploratory QS 2026 two-year extrapolation, XGBoost had the lowest point MAE (4.964), while linear regression had the lowest RMSE (6.670) and the highest Spearman correlation (0.887). Random Forest had an MAE of 5.073 and the persistence baseline an MAE of 5.150. No model therefore showed a consistent advantage across development origins, the final horizon, and evaluation metrics. This pattern supports treating machine learning as an incremental comparison rather than as the central source of predictability [51].

Model configurations were selected using QS 2023 and QS 2024 only; QS 2026 was retained as an exploratory two-year temporal extrapolation and was not used for model selection.

Figure 6 compares the observed QS 2026 positions with predictions from the Random Forest model selected during development. The MAE values of the evaluated models across the QS 2023 and QS 2024 development origins and the exploratory QS 2026 two-year temporal extrapolation are compared in Figure 7. Figure 6 compares the observed QS 2026 positions with predictions from the Random Forest model selected during development. Most points lie near the line of perfect agreement, although some universities show larger deviations. The figure is presented only as a diagnostic view of the exploratory two-year extrapolation and was not used for model selection.

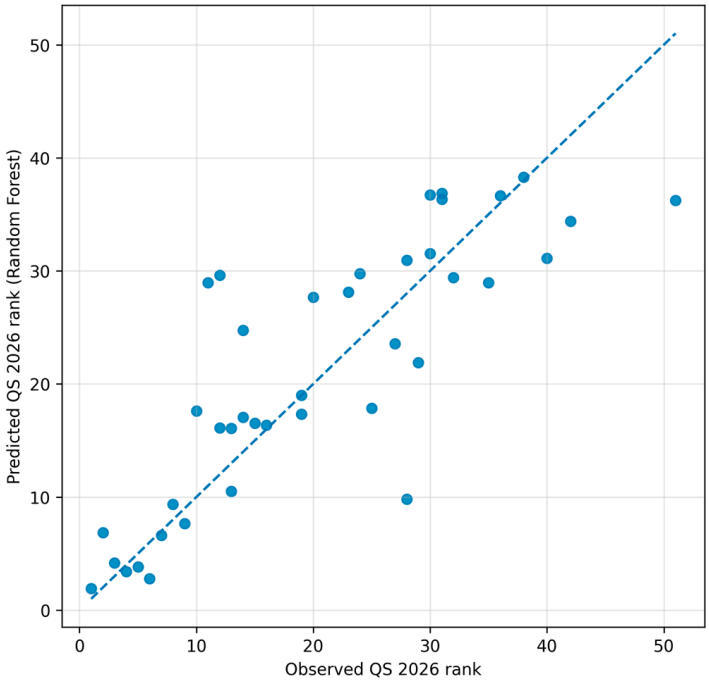

Figure 6. Observed and Random-Forest-predicted QS 2026 positions for the exploratory two-year temporal extrapolation in the 40-university panel. The dotted line indicates perfect agreement.

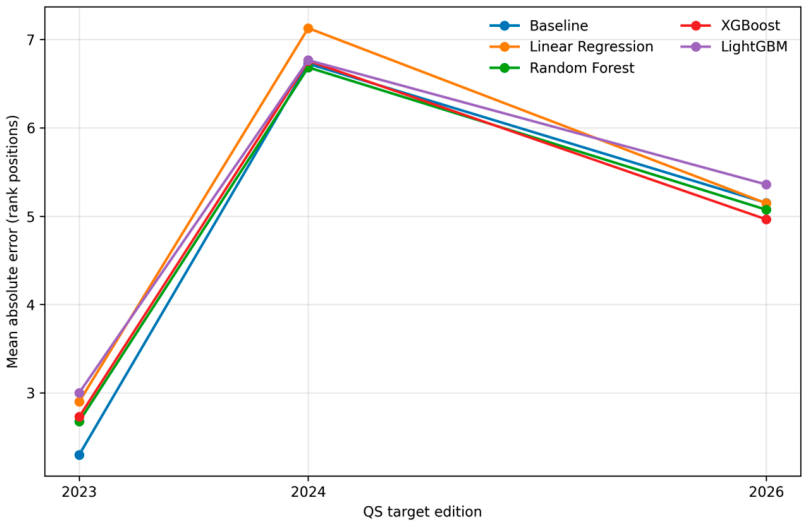

Figure 7. MAE of the evaluated models at the one-year QS 2023 and QS 2024 development origins and the exploratory two-year QS 2026 temporal extrapolation.

Overall, the forecasting results point to persistence rather than model complexity as the main source of predictability in this sample. The most recent QS position remained a strong benchmark, differences between the persistence baseline and machine-learning models were generally small, and no machine-learning model demonstrated a stable advantage across temporal origins and metrics. The paired Random Forest–baseline comparison in Section 4.5 therefore evaluates incremental predictive value beyond persistence.

4.5. Persistence Baseline Versus Random Forest: Paired Comparison

Random Forest was selected as the machine-learning comparator for the paired analysis because it had the lowest mean development MAE among the machine-learning models considered. This selection did not imply superiority over the persistence baseline. The two models were therefore compared on the same 40 universities in the exploratory QS 2026 two-year extrapolation. Persistence had an MAE of 5.150 positions and Random Forest an MAE of 5.073, a point difference of only 0.077 positions in favor of Random Forest. Table 10 reports the paired bootstrap confidence interval, permutation-test result, and university-level win counts.

Table 10. Paired comparison of Random Forest and the persistence baseline on the exploratory QS 2026 two-year temporal extrapolation.

Comparison	(*n*)	Baseline MAE	Random Forest MAE	Mean Paired Difference, RF–Baseline	95% Bootstrap CI	Paired Permutation (*p*)-Value	RF Wins/Ties/ Baseline Wins
Random Forest vs. persistence baseline	40	5.150	5.073	−0.077	[−1.287, 1.031]	0.902	15/0/25

Note. The pairwise difference was defined as the absolute prediction error of the Random Forest model minus the absolute prediction error of the persistence baseline. Negative values indicate a lower absolute error for Random Forest, whereas positive values indicate a lower absolute error for the persistence baseline. The 95% confidence interval was calculated using 20,000 paired bootstrap resamples at the university level. The two-tailed *p*-value was obtained using 100,000 random paired sign permutations.

The 95% confidence interval included zero, and the paired permutation test found no statistically significant difference between the models ($p = 0.902$). Random Forest therefore did not demonstrate a statistically significant advantage over persistence on the exploratory QS 2026 two-year extrapolation. Its slightly lower point MAE should be interpreted as a small numerical difference rather than evidence of a clear improvement.

University-level results lead to the same conclusion. Random Forest had a lower absolute error for 15 of the 40 universities, whereas persistence did so for 25. The proportion of forecasts with an absolute error of no more than five positions was also higher for persistence (67.5% versus 57.5%). As a supplementary descriptive measure, MAPE was 32.12% for persistence and 36.86% for Random Forest. Because the forecasting sample is concentrated at small numerical rank values, these percentage errors should be interpreted cautiously and were not used for model selection [52].

Figure 8 shows both improvements and deteriorations of varying magnitudes. Therefore, the small average difference does not imply a consistent advantage for Random Forest for most universities.

Random Forest's relative performance also varied across temporal origins. Its MAE was 0.376 positions higher than the persistence baseline for QS 2023, 0.041 positions lower for QS 2024, and 0.077 positions lower for the exploratory QS 2026 two-year extrapolation. Because both the direction and magnitude of the difference varied, Random Forest should not be considered a more stable forecasting strategy; within this sample, its predictive performance was broadly comparable to the persistence baseline.

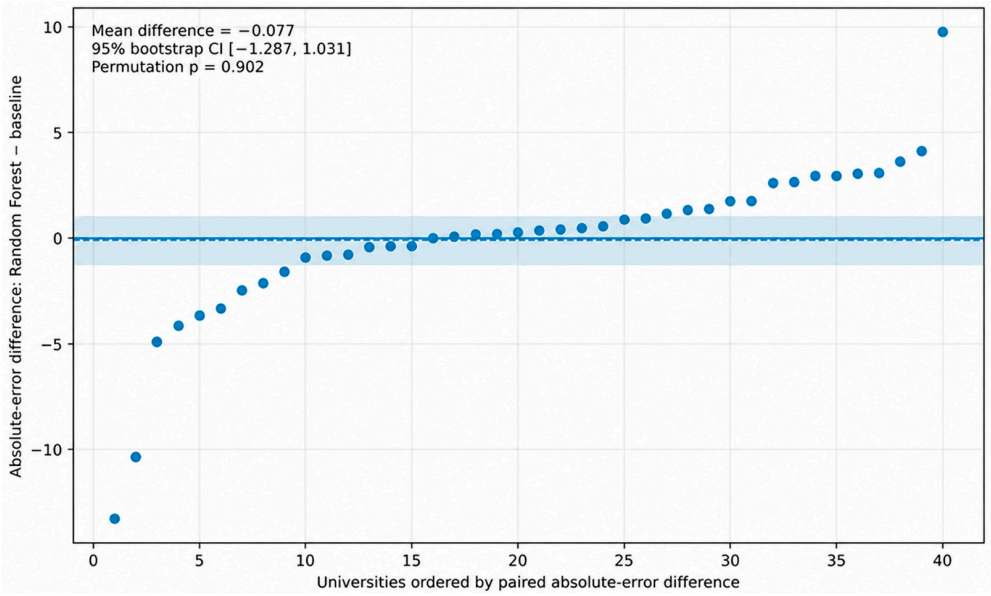

Figure 8. University-level paired differences in absolute prediction error between Random Forest and the persistence baseline on the exploratory QS 2026 two-year temporal extrapolation. Negative values favor Random Forest, whereas positive values favor persistence. The solid horizontal line at zero indicates equal absolute errors for the two models, the dashed horizontal line represents the mean paired difference, and the shaded area represents its 95% paired bootstrap confidence interval.

4.6. Permutation-Based Feature Importance and Temporal Stability

Permutation-based importance was calculated for the Random Forest model separately at the QS 2023, QS 2024, and QS 2026 temporal origins. The results are presented in Table 11 and visualized in Figure 9. Importance is expressed as the increase in validation MAE after permuting one original predictor.

Table 11. Permutation-based feature importance and temporal stability for the Random Forest model.

Feature	Variable	QS 2023 ΔMAE [95% CI]	QS 2024 ΔMAE [95% CI]	QS 2026 ΔMAE [95% CI]	Rank Sequence
Previous QS rank	qs_rank_previous	11.508 [8.244, 15.117]	8.057 [4.291, 11.970]	8.288 [4.854, 11.969]	1/1/1
Previous QS-rank change	qs_rank_change_previous	0.000 [−0.000, 0.000]	−0.024 [−0.071, 0.021]	−0.042 [−0.276, 0.182]	5/4/5
Previous-change availability	qs_rank_change_available	Not estimable	Not estimable	Not estimable	–/–/–
Forecast interval	forecast_gap_years	Not estimable	Not estimable	Not estimable	–/–/–
Year founded	founded	0.030 [−0.126, 0.195]	0.007 [−0.149, 0.164]	0.117 [−0.216, 0.448]	3/2/2
Region	region	0.102 [−0.069, 0.311]	−0.225 [−0.482, 0.002]	0.032 [−0.241, 0.317]	2/5/3
University type	university_type	0.003 [−0.015, 0.021]	−0.009 [−0.036, 0.016]	0.017 [−0.014, 0.049]	4/3/4

Note. Importance is expressed as the increase in validation MAE after permuting one original predictor. Positive values indicate greater model dependence on the feature. Confidence intervals were obtained from 2000 paired bootstrap-permutation resamples at the university level. The variables qs_rank_change_available and forecast_gap_years were constant within each temporal holdout, and their within-origin importance was therefore not estimable. The rank sequence refers to QS 2023/QS 2024/QS 2026.

The most recent available QS position, represented by qs_rank_previous, was the only predictor with a consistently positive importance estimate and a confidence interval that remained above zero at all three origins. Its mean importance was 11.508 ranking positions for QS 2023, 8.057 for QS 2024, and 8.288 for QS 2026. It was ranked first at every temporal origin.

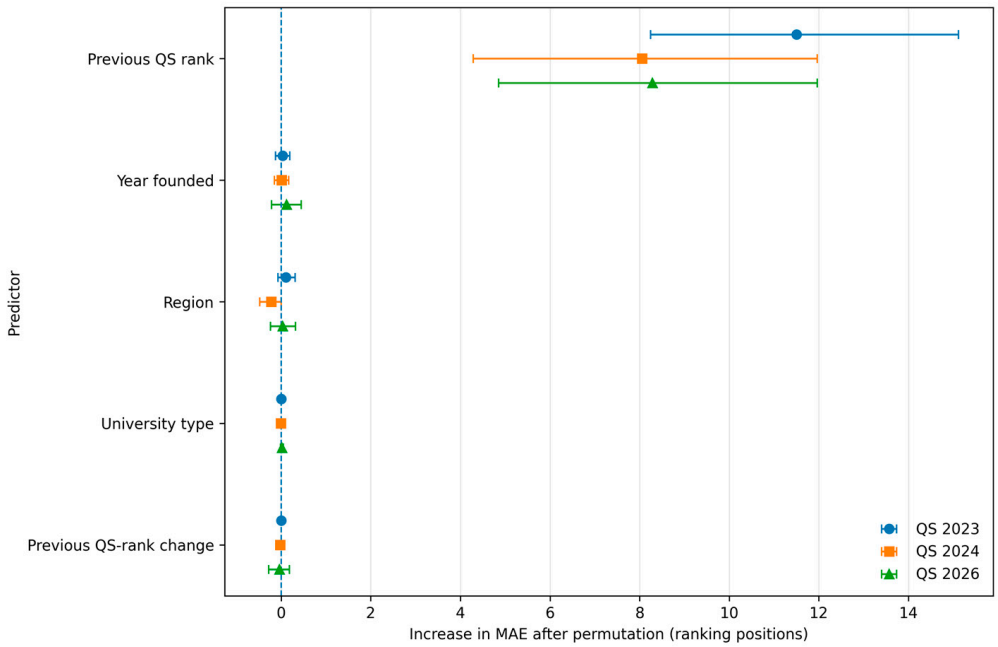

Figure 9. Grouped permutation importance of the Random Forest predictors at the QS 2023, QS 2024, and QS 2026 temporal origins. Importance is expressed as the increase in validation MAE after predictor permutation. Error bars represent 95% paired bootstrap-permutation confidence intervals based on 2000 resamples. The dashed vertical line at zero indicates no change in MAE after permutation; positive values indicate greater model dependence on the predictor, whereas negative values indicate that prediction error decreased slightly after permutation. Variables that were constant within each temporal holdout are not displayed.

The remaining estimable predictors had much smaller and less stable importance values. Confidence intervals for founded, region, university_type, and qs_rank_change_previous included zero at the evaluated origins. These results mean that the variables did not provide stable incremental predictive information to this fitted Random Forest model after the latest QS position and other predictors were considered. They do not show that the corresponding institutional characteristics are unrelated to university performance or ranking outcomes [53].

The variables qs_rank_change_available and forecast_gap_years were constant within each temporal holdout. Their values could therefore not be changed by within-origin permutation, and their importance was reported as not estimable. This does not demonstrate that the variables have no effect; it means that their contribution cannot be separated using variation within a single target edition.

Table 11 presents all the original predictors used by the predictive model. One-hot columns representing the same categorical variable were estimated jointly. Values represent the change in MAE, not normalized proportions, so their sum should not equal 1.

Within this fitted model and sample, university_type did not provide stable incremental predictive information beyond the latest QS position and the remaining predictors. Its mean grouped permutation importance was 0.003 for QS 2023, −0.009 for QS 2024, and 0.017 for QS 2026, and all corresponding 95% confidence intervals included zero.

This is strictly a predictive statement. It does not imply that public and private universities have identical institutional characteristics, trajectories, or causal determinants of ranking performance. The analysis was not designed to estimate an independent effect of institutional type.

The rank of qs_rank_previous was fully stable: it occupied the first position at all three temporal origins. Stability was weaker for the remaining predictors. Spearman's

correlation between the feature rankings was 0.400 for QS 2023 versus QS 2024, 0.900 for QS 2023 versus QS 2026, and 0.700 for QS 2024 versus QS 2026. These coefficients are descriptive because only five predictors could be ranked within each origin.

Figure 9 shows the point estimates and confidence intervals. The large separation between qs_rank_previous and the other predictors indicates that the Random Forest forecast was driven mainly by the most recent observed ranking position. The smaller predictors clustered around zero and changed order across the temporal holdouts.

The results for university type are reported in Table 11 and visualized in Figure 9.

The results therefore do not support a broad claim that several historical or institutional variables were consistently important. A more accurate interpretation is that the model depended strongly and consistently on the latest available QS position, whereas the additional predictors had small or temporally unstable contributions within the analyzed sample.

5. Discussion

The central result of this study is not that machine-learning models substantially improve QS forecasting. Rather, the observed predictability of QS positions in the analyzed panel was driven primarily by persistence in the ranking itself. The most recent QS position was the only predictor with stable positive permutation importance at every temporal origin, the persistence baseline achieved the lowest mean development MAE, and Random Forest did not show a statistically significant advantage over persistence on the exploratory QS 2026 two-year extrapolation. This shifts the interpretation from model performance to the limits and sources of ranking predictability.

The cross-ranking and temporal components should be understood as complementary rather than juxtaposed analyses. The QS–THE–ARWU comparison evaluates cross-sectional consistency: whether different ranking systems place overlapping universities in a similar order at approximately the same period. The QS temporal analysis evaluates persistence and predictability: whether later positions can be anticipated from recent QS positions and whether machine learning adds information beyond that persistence. Both components therefore address different dimensions of ranking stability.

Data completeness varied significantly between the three ranking systems. QS 2026 and THE 2026 rankings were available for the majority of universities in the analytical sample, while ARWU 2025 rankings were available for only 19 institutions. This uneven coverage is due both to differences in the coverage of the ranking systems and to the sample composition of the compiled dataset. Therefore, comparisons involving ARWU were based on significantly smaller overlaps and require more cautious interpretation [5,17–19].

Pairwise analysis revealed a positive relationship between the order of universities within the limited groups of institutions simultaneously represented in the compared rankings. Spearman coefficients ranged from 0.810 to 0.819, but the precision of these estimates depended significantly on the size of the intersection. The QS–THE result was obtained for 54 universities, while the QS–ARWU and THE–ARWU comparisons included only 18 and 19 institutions, respectively. The wider bootstrap confidence intervals for comparisons with ARWU indicate significant uncertainty, despite the relatively large point coefficients.

Rank correlation and absolute distance between rankings characterize different data properties. A high Spearman coefficient indicates that the two ranking systems rank universities in roughly similar order, but do not require their assigned numerical positions to be close. Therefore, a high rank correlation can simultaneously be observed with a relatively large average absolute difference in the published rankings.

The use of normalized percentile positions also changed the interpretation of differences. For example, the average absolute difference between QS and THE, equal to

approximately 20.8 published positions, corresponded to approximately 1.09 percentage points after accounting for the different ranking sizes. This does not make the ranking systems interchangeable, but it does show that comparing only absolute positions can exaggerate the differences between systems with different list lengths.

Therefore, the obtained results should be viewed as exploratory evidence for selected overlapping groups of universities. They do not prove the same level of consistency among QS, THE, and ARWU across the entire population of ranked institutions. This limitation is particularly important for ARWU, which in this sample was represented by a small group of predominantly highly ranked universities [49].

A pairwise comparison of QS 2024 and QS 2026 revealed no statistically confirmed directional change within the sample of 44 universities. Although the mean change was −0.955 positions, the median was 0, the Wilcoxon signed-rank test yielded $p = 0.426$, and the bootstrap confidence interval of the mean included zero. Therefore, at the overall level, improvements and deteriorations largely balanced each other.

This result does not mean that individual university positions were stable. Several institutions moved more than ten places in either direction, so the overall average can mask university-level volatility. The finding supports use of the previous ranking as a transparent persistence benchmark, but it does not imply that persistence will be accurate for every institution or edition.

A broader analysis of the balanced 40-university panel showed that position volatility varied across transitions. The mean absolute change was 2.425 positions for QS 2021–2022, 2.300 for QS 2022–2023, and 6.725 for QS 2023–2024. The differences among the one-year transitions were statistically significant, showing that temporal prediction difficulty was not constant across editions.

The two-year QS 2024–QS 2026 transition requires separate interpretation from the one-year development origins. The final persistence MAE of 5.150 should therefore be read as an exploratory two-year extrapolation, not as a conventional one-period forecast. The horizon-matched robustness check supports this distinction: persistence MAE was 3.025 for QS 2021–QS 2023 and 6.875 for QS 2022–QS 2024, placing the QS 2024–QS 2026 value between the two historical two-year transitions. Thus, the final horizon was neither directly comparable with the one-year origins nor obviously anomalous within the limited two-year history.

This horizon check also clarifies the role of forecast_gap_years. Encoding the interval length allows the model matrix to record that the 2026 target is two years from the last observed QS position, but it does not make a two-year extrapolation equivalent to a one-year forecast. Because the available series begins in 2021, a leakage-free machine-learning backtest for the 2021–2023 horizon would require pre-2021 training observations; we therefore limited the historical two-year robustness analysis to the persistence benchmark rather than introducing a retrospectively contaminated machine-learning comparison.

The baseline-versus-machine-learning comparison is consequently one of the most informative findings. The persistence baseline had the lowest mean MAE during development. On QS 2026, Random Forest improved the point MAE by only 0.077 positions, with a 95% paired bootstrap confidence interval of −1.287 to 1.031 and $p = 0.902$ in the paired permutation test. Moreover, the persistence baseline produced a lower absolute error for 25 of the 40 universities. These results provide no evidence of a consistent practical or statistical advantage from the additional Random Forest complexity.

Although XGBoost produced the lowest point MAE on the final QS 2026 extrapolation, it did not show a consistent advantage during the development origins or across the other metrics. This reinforces the broader conclusion that selecting a more complex learner does not, in itself, overcome the strong temporal inertia of the ranking positions in this sample.

The feature-importance results support the same interpretation. The latest QS position dominated grouped permutation importance at QS 2023, QS 2024, and QS 2026, while the confidence intervals for the other estimable predictors included zero and their rank ordering was less stable. The appropriate conclusion is therefore that these variables did not contribute stable incremental predictive information within this model and sample after recent QS position was taken into account.

This finding must remain strictly predictive. It does not imply that region, university type, institutional age, or other characteristics have no influence on university performance or ranking outcomes. Establishing such effects would require longitudinal measurements of the relevant indicators, an identification strategy suited to causal inference, and a substantially broader sample.

For practical monitoring, the persistence baseline is therefore an essential reference model rather than merely a preliminary comparator. A more complex forecasting workflow should be preferred only if it demonstrates reproducible incremental accuracy across multiple later editions, ranking ranges, and institutional groups. Under the present conditions, machine-learning models are better viewed as supplementary diagnostic tools than as replacements for the transparent persistence benchmark.

This result is consistent with the relatively good quality of the baseline model. Random Forest relied primarily on the latest known ranking position, while additional lagged and institutional variables provided little consistent information within the available sample. Permutation importance reflects the dependence of the trained model on the predictors and should not be interpreted as evidence of a causal influence of the features on university performance.

The dominance of the most recently available QS position also defines the practical limits of the fitted model. It may support near-term persistence monitoring and help identify universities whose later positions deviate from recent ranking dynamics. However, it cannot estimate how specific changes in citation impact, faculty–student ratio, academic reputation, employer reputation, or internationalization would affect future QS positions.

Historical values of these structural indicators for individual ranking editions were unavailable and were therefore intentionally not included in the forecast matrix. Consequently, the current model cannot be used for counterfactual policy modeling or for assessing the effect of specific policy interventions. Such an analysis would require longitudinal data on these indicators with precisely recorded observation dates and a separate causal or scenario-based research design.

University type illustrates the importance of this distinction. The predictive subsample contained 28 public and 12 private universities, and grouped permutation importance for university_type remained close to zero across the evaluated origins. Within the fitted Random Forest model, this indicates no stable incremental predictive contribution after recent QS position and the other predictors were considered.

It should not be interpreted as evidence that university status is unrelated to institutional performance or ranking dynamics. The private-university subgroup was small, the forecasting panel was concentrated near the top of the QS distribution, and the study was designed for predictive evaluation rather than causal attribution. A larger and more balanced longitudinal design would be required to estimate systematic differences between public and private institutions.

For institutional analytics, the combined results support a cautious workflow: use cross-ranking comparison to understand how consistently systems order overlapping institutions, use persistence as the primary temporal benchmark, and treat machine-learning predictions as supplementary unless they demonstrate reproducible incremental value.

The analysis should not be used to replicate proprietary ranking methodologies or to claim precise prediction of official ranking outcomes.

An additional limitation is the concentration of the forecasting sample in the upper QS ranges. Of the 40 universities, 39 were in the Top 50 and none ranked below 51st in QS 2026. Consequently, MAE values of around five positions on the exploratory two-year extrapolation describe error within a narrow group of highly ranked institutions and should not be generalized to lower ranking ranges.

This sample composition may have partially influenced the relatively good performance of the baseline model and the dominant permutation importance of the last QS ranking. However, the current analysis does not allow us to separately determine how much of the obtained performance is related to the ranking range itself and how much is related to the actual temporal stability of the universities under consideration. Therefore, the result should be considered dependent on the sample composition, rather than proof that the same accuracy will be obtained for universities in the Top 300, Top 500, or lower ranking ranges.

Further testing is necessary on a larger, more balanced sample of universities from different ranking groups. Separately reporting errors for the Top 50, Top 100, Top 300, and lower ranking ranges will allow us to determine whether the forecasting quality varies across different parts of the ranking scale.

The study has several limitations. The predictive experiment used only 40 universities with complete QS histories and was heavily concentrated in the upper ranking range. The final QS 2026 evaluation spans two calendar years because QS 2025 is absent from the working dataset; forecast_gap_years records this fact but does not remove the horizon mismatch. The horizon-matched persistence check is descriptive and covers only two historical two-year transitions. These constraints limit both the precision and generalizability of the forecasting results.

Another limitation relates to the uneven completeness of the ranking data. ARWU 2025 rankings were available for only 19 universities, while 18 institutions were represented simultaneously in QS, THE, and ARWU. Therefore, comparisons involving ARWU were conducted on a significantly smaller subsample and require particularly careful interpretation. Historical rankings for THE and ARWU for 2021–2024 were also unavailable, leading the time and forecast sections of the study to focus primarily on QS.

Taken together, the results are best interpreted as exploratory evidence about ranking stability and the limits of incremental predictability. They do not establish a generally superior forecasting algorithm. Instead, they show that, within this panel, recent ranking position carries most of the stable predictive signal and that additional model complexity yields small and temporally inconsistent gains.

Future validation should therefore prioritize broader longitudinal coverage rather than additional algorithmic complexity alone. More ranking editions, universities from lower and middle ranking ranges, and time-stamped institutional and scientometric predictors would make it possible to test whether machine learning adds reproducible value beyond persistence and to separate prediction from causal or policy-oriented analysis.

6. Conclusions

This study examined university ranking stability and predictability through two complementary dimensions: cross-sectional consistency across QS, THE, and ARWU, and temporal persistence within QS. The analytical sample comprised 57 universities, while the temporally balanced forecasting panel contained 40 universities with QS positions for 2021–2024 and 2026.

Across the available pairwise overlaps, QS, THE, and ARWU showed positive rank correlations, but the systems were not interchangeable and the ARWU comparisons were based on only 18–19 universities. The cross-ranking results therefore provide evidence of partial cross-sectional consistency within the observed overlaps, not a claim of equivalence across the full ranking populations.

The main temporal finding is that QS predictability was driven primarily by persistence. The persistence baseline had the lowest mean development MAE, and on the exploratory QS 2026 two-year extrapolation Random Forest achieved an MAE of 5.073 compared with 5.150 for persistence. The paired difference was only -0.077 positions, its 95% bootstrap confidence interval included zero, and the paired permutation test was not significant ($p = 0.902$). The persistence baseline also had a lower absolute error for 25 of the 40 universities.

The horizon-matched robustness analysis reinforces the need for careful interpretation of QS 2026. Persistence MAE was 3.025 for QS 2021–QS 2023, 6.875 for QS 2022–QS 2024, and 5.150 for QS 2024–QS 2026. The final result should therefore be regarded as an exploratory two-year temporal extrapolation whose difficulty lies between the two historical two-year transitions, not as equivalent to the one-year development origins.

Grouped permutation importance provides the same substantive message. The most recent QS position was the only predictor with stable positive importance across all temporal origins. The remaining estimable variables did not provide stable incremental predictive information within the fitted Random Forest model. This result is strictly predictive and does not imply that region, university type, institutional age, or other characteristics lack substantive or causal influence on university performance.

Accordingly, the principal contribution is evidence about the source and limits of ranking predictability rather than evidence for the superiority of machine learning. Within this small, top-heavy panel, additional model complexity produced limited and temporally unstable gains beyond a transparent persistence benchmark. Machine-learning models may still be useful as supplementary diagnostics, but replacing persistence with a more complex workflow would require reproducible improvement on larger and more diverse longitudinal samples.

The results remain exploratory. The forecasting panel was purposive: 39 of 40 universities were in the QS Top 50 in 2026, ARWU coverage was limited, historical THE and ARWU series were unavailable, and the final QS horizon was two years because QS 2025 was absent. Future work should prioritize broader temporal and institutional coverage and time-stamped predictors so that incremental predictive value can be assessed more reliably and causal questions can be studied separately.

Author Contributions: Conceptualization, E.A. and O.B.; methodology, K.S. and V.S.; software, O.B. and Y.C.; validation, E.A., A.O. and S.K.; formal analysis, O.B., B.A. and K.S.; investigation, V.S., S.K. and Y.C.; resources, A.O. and K.S.; data curation, B.A.; writing—original draft preparation, E.A., S.K.; writing—review and editing, O.B., A.O. and K.S.; visualization, S.K., V.S. and Y.C.; supervision, K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ML	Machine Learning
QS	Quacquarelli Symonds World University Rankings
THE	Times Higher Education World University Rankings
ARWU	Academic Ranking of World Universities
CWUR	Center for World University Rankings
RF	Random Forest
XGBoost	eXtreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
MAPE	Mean Absolute Percentage Error
R^2	Coefficient of Determination
CV	Cross-Validation
CSV	Comma-Separated Values
RDD	Resilient Distributed Dataset
SHAP	SHapley Additive exPlanations
NLP	Natural Language Processing

Appendix A

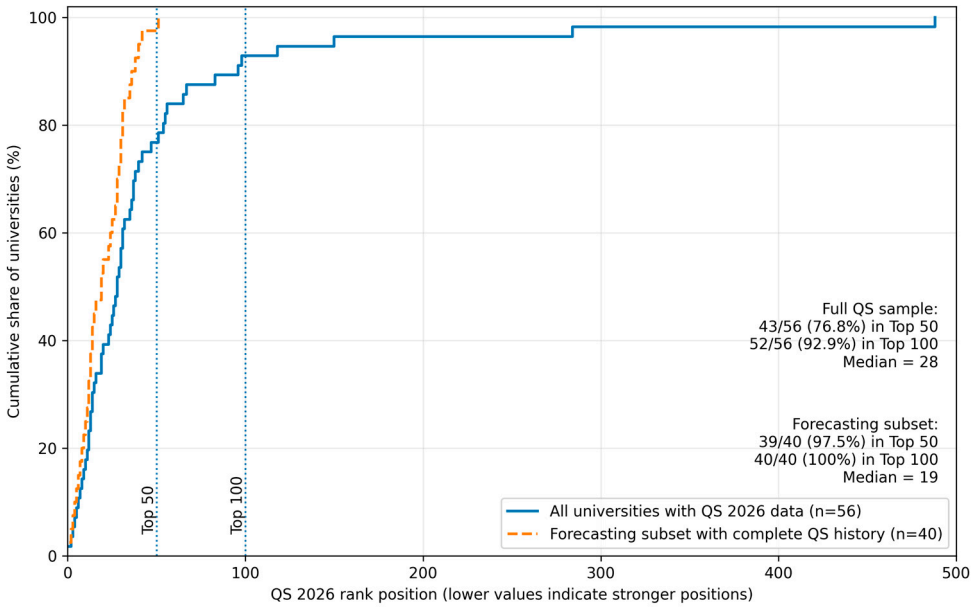

Figure A1. Empirical cumulative distributions of QS 2026 rankings in the full descriptive sample and the forecast subsample. The vertical lines indicate the boundaries of the Top 50 and Top 100. The forecast subsample is significantly more concentrated in the upper ranking range than the full QS sample. Lower numerical values correspond to higher rankings.

Table A1. Composition of the analytical sample, ranking coverage, historical QS data availability, and eligibility for the forecasting experiment by university.

No.	University	Country	Region	University Type	QS 2026 Rank	THE 2026 Rank	ARWU 2025 Rank	QS 2021 Available	QS 2022 Available	QS 2023 Available	QS 2024 Available	QS 2026 Available	Included in QS–THE Comparison	Included in QS–ARWU Comparison	Included in THE–ARWU Comparison	Included in QS 2026 Forecasting Sample	Reason for Inclusion in the Initial Dataset	Reason for Exclusion from Forecasting, If Applicable
1	MIT	USA	North America	Private	1	2	3	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria	Not applicable
2	Imperial College London	UK	Europe	Public	2	8	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
3	Stanford University	USA	North America	Private	3	5	2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria	Not applicable
4	University of Oxford	UK	Europe	Public	4	1	6	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria	Not applicable
5	Harvard University	USA	North America	Private	5	5	1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria	Not applicable
6	University of Cambridge	UK	Europe	Public	6	3	4	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria.	Not applicable
7	ETH Zurich	Switzerland	Europe	Public	7	7	22	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria	Not applicable
8	NUS Singapore	Singapore	Asia	Public	8	17	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable

Table A1. *Cont.*

No.	University	Country	Region	University Type	QS 2026 Rank	THE 2026 Rank	ARWU 2025 Rank	QS 2021 Available	QS 2022 Available	QS 2023 Available	QS 2024 Available	QS 2026 Available	Included in QS–THE Comparison	Included in QS–ARWU Comparison	Included in THE–ARWU Comparison	Included in QS 2026 Forecasting Sample	Reason for Inclusion in the Initial Dataset	Reason for Exclusion from Forecasting, If Applicable
9	UCL	UK	Europe	Public	9	9	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
10	Caltech	USA	North America	Private	10	7	9	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria	Not applicable
11	Univ of Chicago	USA	North America	Private	13	15	10	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria	Not applicable
12	Princeton University	USA	North America	Private	15	3	7	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria	Not applicable
13	Univ of Hong Kong	Hong Kong	Asia	Public	11	35	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
14	NTU Singapore	Singapore	Asia	Public	12	46	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
15	Cornell University	USA	North America	Private	13	18	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
16	Yale University	USA	North America	Private	16	9	14	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria	Not applicable

Table A1. *Cont.*

No.	University	Country	Region	University Type	QS 2026 Rank	THE 2026 Rank	ARWU 2025 Rank	QS 2021 Available	QS 2022 Available	QS 2023 Available	QS 2024 Available	QS 2026 Available	Included in QS–THE Comparison	Included in QS–ARWU Comparison	Included in THE–ARWU Comparison	Included in QS 2026 Forecasting Sample	Reason for Inclusion in the Initial Dataset	Reason for Exclusion from Forecasting, If Applicable
17	Columbia University	USA	North America	Private	14	20	8	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria	Not applicable
18	Univ of Pennsylvania	USA	North America	Private	12	14	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
19	Univ of Melbourne	Australia	Oceania	Public	19	16	33	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria	Not applicable
20	Peking University	China	Asia	Public	14	13	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
21	Tsinghua University	China	Asia	Public	20	12	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
22	UC Berkeley	USA	North America	Public	28	9	5	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria	Not applicable
23	Johns Hopkins University	USA	North America	Private	24	21	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
24	Univ of Sydney	Australia	Oceania	Public	25	43	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable

Table A1. *Cont.*

No.	University	Country	Region	University Type	QS 2026 Rank	THE 2026 Rank	ARWU 2025 Rank	QS 2021 Available	QS 2022 Available	QS 2023 Available	QS 2024 Available	QS 2026 Available	Included in QS–THE Comparison	Included in QS–ARWU Comparison	Included in THE–ARWU Comparison	Included in QS 2026 Forecasting Sample	Reason for Inclusion in the Initial Dataset	Reason for Exclusion from Forecasting, If Applicable
25	Duke University	USA	North America	Private	26	22	NA	Yes	No	Yes	No	Yes	Yes	No	No	No	Met the initial sample eligibility criteria	Incomplete historical QS series: QS ranks for 2022 and 2024 were unavailable
26	Univ of Toronto	Canada	North America	Public	29	22	24	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria	Not applicable
27	Univ of Michigan	USA	North America	Public	23	22	21	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Met the initial sample eligibility criteria	Not applicable
28	Kings College London	UK	Europe	Public	31	44	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
29	Univ of Manchester	UK	Europe	Public	35	63	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
30	McGill University	Canada	North America	Public	32	44	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
31	Seoul National University	South Korea	Asia	Public	31	62	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
32	Fudan University	China	Asia	Public	30	44	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable

Table A1. *Cont.*

No.	University	Country	Region	University Type	QS 2026 Rank	THE 2026 Rank	ARWU 2025 Rank	QS 2021 Available	QS 2022 Available	QS 2023 Available	QS 2024 Available	QS 2026 Available	Included in QS–THE Comparison	Included in QS–ARWU Comparison	Included in THE–ARWU Comparison	Included in QS 2026 Forecasting Sample	Reason for Inclusion in the Initial Dataset	Reason for Exclusion from Forecasting, If Applicable
33	Univ of Tokyo	Japan	Asia	Public	28	26	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
34	Univ of Edinburgh	UK	Europe	Public	27	30	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
35	CUHK Hong Kong	Hong Kong	Asia	Public	36	47	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
36	UCLA	USA	North America	Public	40	22	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
37	UNSW Sydney	Australia	Oceania	Public	19	67	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
38	Paris-Saclay University	France	Europe	Public	NA	17	13	No	No	No	No	No	No	No	Yes	No	Met the initial sample eligibility criteria	Incomplete QS series: QS ranks for 2021–2024 and the target QS 2026 rank were unavailable
39	LSE	UK	Europe	Public	56	52	NA	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Met the initial sample eligibility criteria	Incomplete historical QS series: the QS rank for 2023 was unavailable

Table A1. *Cont.*

No.	University	Country	Region	University Type	QS 2026 Rank	THE 2026 Rank	ARWU 2025 Rank	QS 2021 Available	QS 2022 Available	QS 2023 Available	QS 2024 Available	QS 2026 Available	Included in QS–THE Comparison	Included in QS–ARWU Comparison	Included in THE–ARWU Comparison	Included in QS 2026 Forecasting Sample	Reason for Inclusion in the Initial Dataset	Reason for Exclusion from Forecasting, If Applicable
40	Kyoto University	Japan	Asia	Public	51	38	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
41	Univ of Amsterdam	Netherlands	Europe	Public	55	61	NA	No	No	No	No	Yes	Yes	No	No	No	Met the initial sample eligibility criteria	Incomplete historical QS series: QS ranks for 2021, 2022, 2023, and 2024 were unavailable
42	KAIST	South Korea	Asia	Public	83	134	NA	Yes	Yes	Yes	No	Yes	Yes	No	No	No	Met the initial sample eligibility criteria	Incomplete historical QS series: the QS rank for 2024 was unavailable
43	Australian National Univ	Australia	Oceania	Public	30	78	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
44	Carnegie Mellon University	USA	North America	Private	65	27	NA	Yes	No	No	No	Yes	Yes	No	No	No	Met the initial sample eligibility criteria	Incomplete historical QS series: QS ranks for 2022, 2023, and 2024 were unavailable
45	NYU	USA	North America	Private	38	27	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable

Table A1. *Cont.*

No.	University	Country	Region	University Type	QS 2026 Rank	THE 2026 Rank	ARWU 2025 Rank	QS 2021 Available	QS 2022 Available	QS 2023 Available	QS 2024 Available	QS 2026 Available	Included in QS–THE Comparison	Included in QS–ARWU Comparison	Included in THE–ARWU Comparison	Included in QS 2026 Forecasting Sample	Reason for Inclusion in the Initial Dataset	Reason for Exclusion from Forecasting, If Applicable
46	Monash University	Australia	Oceania	Public	37	88	NA	No	No	No	Yes	Yes	Yes	No	No	No	Met the initial sample eligibility criteria	Incomplete historical QS series: QS ranks for 2021, 2022, and 2023 were unavailable
47	Delft Univ of Technology	Netherlands	Europe	Public	47	51	NA	No	No	No	Yes	Yes	Yes	No	No	No	Met the initial sample eligibility criteria	Incomplete historical QS series: QS ranks for 2021, 2022, and 2023 were unavailable
48	Univ of British Columbia	Canada	North America	Public	42	41	NA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Met the initial sample eligibility criteria	Not applicable
49	Technical Univ of Munich	Germany	Europe	Public	37	34	45	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	Met the initial sample eligibility criteria	Incomplete historical QS series: the QS rank for 2022 was unavailable
50	LMU Munich	Germany	Europe	Public	54	32	42	No	No	No	No	Yes	Yes	Yes	Yes	No	Met the initial sample eligibility criteria	Incomplete historical QS series: QS ranks for 2021, 2022, 2023, and 2024 were unavailable
51	Univ of Sao Paulo	Brazil	Latin America	Public	96	143	81	No	No	No	No	Yes	Yes	Yes	Yes	No	Met the initial sample eligibility criteria	Incomplete historical QS series: QS ranks for 2021, 2022, 2023, and 2024 were unavailable

Table A1. *Cont.*

No.	University	Country	Region	University Type	QS 2026 Rank	THE 2026 Rank	ARWU 2025 Rank	QS 2021 Available	QS 2022 Available	QS 2023 Available	QS 2024 Available	QS 2026 Available	Included in QS–THE Comparison	Included in QS–ARWU Comparison	Included in THE–ARWU Comparison	Included in QS 2026 Forecasting Sample	Reason for Inclusion in the Initial Dataset	Reason for Exclusion from Forecasting, If Applicable
52	Univ of Cape Town	South Africa	Africa	Public	150	178	NA	No	No	No	No	Yes	Yes	No	No	No	Met the initial sample eligibility criteria	Incomplete historical QS series: QS ranks for 2021, 2022, 2023, and 2024 were unavailable
53	IIT Bombay	India	Asia	Public	118	351	NA	No	No	No	No	Yes	Yes	No	No	No	Met the initial sample eligibility criteria	Incomplete historical QS series: QS ranks for 2021, 2022, 2023, and 2024 were unavailable
54	KFUPM Saudi Arabia	Saudi Arabia	Middle East	Public	67	234	NA	No	No	No	No	Yes	Yes	No	No	No	Met the initial sample eligibility criteria	Incomplete historical QS series: QS ranks for 2021, 2022, 2023, and 2024 were unavailable
55	Politecnico di Milano	Italy	Europe	Public	98	NA	NA	No	No	No	No	Yes	No	No	No	No	Met the initial sample eligibility criteria	Incomplete historical QS series: QS ranks for 2021, 2022, 2023, and 2024 were unavailable
56	Univ of Warsaw	Poland	Europe	Public	284	251	NA	No	No	No	No	Yes	Yes	No	No	No	Met the initial sample eligibility criteria	Incomplete historical QS series: QS ranks for 2021, 2022, 2023, and 2024 were unavailable

Table A1. *Cont.*

No.	University	Country	Region	University Type	QS 2026 Rank	THE 2026 Rank	ARWU 2025 Rank	QS 2021 Available	QS 2022 Available	QS 2023 Available	QS 2024 Available	QS 2026 Available	Included in QS–THE Comparison	Included in QS–ARWU Comparison	Included in THE–ARWU Comparison	Included in QS 2026 Forecasting Sample	Reason for Inclusion in the Initial Dataset	Reason for Exclusion from Forecasting, If Applicable
57	Sunway University	Malaysia	Asia	Private	488	NA	NA	No	No	No	No	Yes	No	No	No	No	Met the initial sample eligibility criteria	Incomplete historical QS series: QS ranks for 2021, 2022, 2023, and 2024 were unavailable

Note. "Yes" indicates that the required value or pair of ranking values was available; "No" indicates that at least one required value was missing. Inclusion in the QS 2026 forecasting sample required QS ranks for 2021, 2022, 2023, and 2024 and an observed QS 2026 rank. Missing ranking values are reported as "NA" and were not replaced by zero, mean, median, or reconstructed values. The initial dataset is a purposive analytical sample and is not statistically representative of the full QS-ranked population. Ranking releases: QS 2026 (June 2025), THE 2026 (October 2025), and ARWU 2025 (August 2025). Official ranking webpages were accessed on 14 April 2026.

Table A2. Statistical assessment of paired QS-position changes between 2024 and 2026.

Comparison	n	Improved	Unchanged	Declined	Mean Change	Median	Wilcoxon W	*p*-Value	95% Bootstrap CI
QS 2026 minus QS 2024	44	17	12	15	−0.955	0	221.5	0.426	[−3.091, 1.159]

Note. Negative changes indicate improved ranking positions, while positive changes indicate declines. The Wilcoxon test was two-sided. The confidence interval was estimated using 20,000 paired bootstrap resamples at the university level with a random seed of 42.

Table A3. Ranking volatility across the observed QS transitions in the balanced 40-university panel.

Transition	Calendar Interval	n	Mean Signed Change	Mean Absolute Change	95% Bootstrap CI for Mean Absolute Change	Median Absolute Change	Maximum Absolute Change
QS 2021–QS 2022	1 year	40	−0.325	2.425	[1.850, 3.025]	2.000	8
QS 2022–QS 2023	1 year	40	0.350	2.300	[1.675, 2.975]	2.000	8
QS 2023–QS 2024	1 year	40	−0.475	6.725	[4.750, 8.850]	4.500	26
QS 2024–QS 2026	2 years	40	−1.200	5.150	[3.600, 6.825]	4.000	20

Note. Signed change was calculated as the later QS position minus the earlier QS position; negative values indicate improvement. Confidence intervals were estimated using 20,000 paired bootstrap resamples at the university level with a fixed random seed of 42. The QS 2024–QS 2026 transition covers two calendar years and should not be interpreted as directly equivalent to a one-year transition.

Table A4. Hyperparameter search spaces and selected configurations for the tree-based forecasting models.

Model	Search Space	Number of Configurations	Selected Configuration	MAE QS 2023	MAE QS 2024	Mean Development MAE
Random Forest	n_estimators = {500}; max_depth = {None, 4, 8}; min_samples_leaf = {1, 2, 4}; max_features = {1.0, sqrt}	18	n_estimators = 500; max_depth = 4; min_samples_leaf = 2; max_features = 1.0	2.676	6.684	4.680
XGBoost	n_estimators = {150, 300}; max_depth = {2, 3}; learning_rate = {0.03, 0.08}; subsample = {0.8, 1.0}; colsample_bytree = {0.8, 1.0}	32	n_estimators = 150; max_depth = 2; learning_rate = 0.03; subsample = 1.0; colsample_bytree = 1.0	2.731	6.758	4.744
LightGBM	n_estimators = {150, 300}; num_leaves = {5, 10, 20}; learning_rate = {0.03, 0.08}; min_child_samples = {5, 10}	24	n_estimators = 150; num_leaves = 5; learning_rate = 0.03; min_child_samples = 5	3.001	6.765	4.883

Note. Configurations were selected using the mean MAE across the QS 2023 and QS 2024 development origins. The exploratory QS 2026 two-year temporal extrapolation was not used during tuning. Fixed, non-tuned settings included random_state = 42; reg_lambda = 1.0 for XGBoost and LightGBM; and objective = reg:squarederror for XGBoost. Linear regression and the persistence baseline were not tuned.

Table A5. Concentration of QS 2026 positions in the descriptive and forecasting samples.

Sample	n	Mean	Median	Q1	Q3	Maximum	Top 50, n (%)	Top 100, n (%)	Skewness
All universities with QS 2026 data	56	45.964	28	13.00	43.25	488	43 (76.8%)	52 (92.9%)	4.509
Forecasting subset with complete QS history	40	20.325	19	10.75	30.00	51	39 (97.5%)	40 (100.0%)	0.373

Note. Lower numerical positions indicate stronger ranking outcomes. The forecasting subset includes universities with complete QS observations for 2021, 2022, 2023, 2024, and 2026.

Table A6. Horizon-matched persistence robustness across two-year QS transitions in the balanced 40-university panel.

Transition	Calendar Interval	n	Persistence MAE	Median Absolute Error	Maximum Absolute Error
QS 2021–QS 2023	2 years	40	3.025	2	11
QS 2022–QS 2024	2 years	40	6.875	5	24
QS 2024–QS 2026	2 years	40	5.150	4	20

Note. The persistence forecast for each target edition equals the QS position observed at the corresponding two-year forecast origin; therefore, persistence MAE is numerically identical to the mean absolute rank change over that transition. The first two rows are historical horizon-matched robustness checks, while QS 2024–QS 2026 is the final exploratory two-year temporal extrapolation. This analysis is descriptive and was not used for model selection or hyperparameter tuning.

References

1. Hazelkorn, E. *Rankings and the Reshaping of Higher Education: The Battle for World-Class Excellence*, 2nd ed.; Palgrave Macmillan: London, UK, 2015. [CrossRef]
2. Dill, D.D.; Soo, M. Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems. *High. Educ.* **2005**, *49*, 495–533. [CrossRef]
3. Marginson, S. University rankings and social science. *Eur. J. Educ.* **2014**, *49*, 45–59. [CrossRef]

4. Altbach, P.G. The globalization of college and university rankings. *Change* **2012**, *44*, 26–31. [CrossRef]
5. Johnes, J. University rankings: What do they really show? *Scientometrics* **2018**, *115*, 585–606. [CrossRef]
6. Times Higher Education. World University Rankings 2026. Available online: https://www.timeshighereducation.com/world-university-rankings (accessed on 14 April 2026).
7. QS Quacquarelli Symonds. QS World University Rankings 2026. Available online: https://www.topuniversities.com/world-university-rankings (accessed on 14 April 2026).
8. ShanghaiRanking Consultancy. Academic Ranking of World Universities 2025. Available online: https://www.shanghairanking.com/rankings/arwu/2025 (accessed on 14 April 2026).
9. Cybermetrics Lab. Ranking Web of Universities: Webometrics. Available online: https://www.webometrics.online/ (accessed on 14 April 2026).
10. Liu, N.C.; Cheng, Y. The academic ranking of world universities. *High. Educ. Eur.* **2005**, *30*, 127–136. [CrossRef]
11. Shin, J.C.; Toutkoushian, R.K. The past, present, and future of university rankings. In *University Rankings: Theoretical Basis, Methodology and Impacts on Global Higher Education*; Springer: Dordrecht, The Netherlands, 2011; pp. 1–16. [CrossRef]
12. Daraio, C.; Bonaccorsi, A.; Simar, L. Rankings and university performance: A conditional multidimensional approach. *Eur. J. Oper. Res.* **2015**, *244*, 918–930. [CrossRef]
13. Rocha, B.; Figueira, Á. Post, predict, and rank: Exploring the relationship between social media strategy and higher education institution rankings. *Informatics* **2025**, *12*, 6. [CrossRef]
14. Udupi, P.K.; Dattana, V.; Netravathi, P.S.; Pandey, J. Predicting global ranking of universities across the world using machine learning regression technique. *SHS Web Conf.* **2023**, *156*, 04001. [CrossRef]
15. Wardley, L.J.; Rajabi, E.; Amin, S.H.; Ramesh, M. A Machine Learning Approach Feature to Forecast the Future Performance of the Universities in Canada. *Mach. Learn. Appl.* **2024**, *16*, 100548. [CrossRef]
16. Wisaeng, K.; Muangmeesri, B. University Rankings Prediction Using Hybrid Feature Selection Based on Machine Learning Methods. *Int. J. Anal. Appl.* **2025**, *23*, 112. [CrossRef]
17. Pavel, A.P. Global university rankings—A comparative analysis. *Procedia Econ. Financ.* **2015**, *26*, 54–63. [CrossRef]
18. Zhang, Y.; Xiao, Y.; Wu, J.; Lu, X. Comprehensive world university ranking based on ranking aggregation. *Comput. Stat.* **2021**, *36*, 1139–1152. [CrossRef]
19. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
20. Waltman, L. A review of the literature on citation impact indicators. *J. Inf.* **2016**, *10*, 365–391. [CrossRef]
21. Bornmann, L.; Leydesdorff, L. Scientometrics in a changing research landscape. *EMBO Rep.* **2014**, *15*, 1228–1232. [CrossRef] [PubMed]
22. Hicks, D.; Wouters, P.; Waltman, L.; de Rijcke, S.; Rafols, I. Bibliometrics: The Leiden Manifesto for research metrics. *Nature* **2015**, *520*, 429–431. [CrossRef] [PubMed]
23. Bornmann, L.; Daniel, H.-D. What do citation counts measure? A review of studies on citing behavior. *J. Doc.* **2008**, *64*, 45–80. [CrossRef]
24. Aksnes, D.W.; Langfeldt, L.; Wouters, P. Citations, citation indicators, and research quality: An overview of basic concepts and theories. *Res. Eval.* **2019**, *28*, 185–195. [CrossRef]
25. Waltman, L.; van Eck, N.J. The inconsistency of the h-index. *J. Am. Soc. Inf. Sci. Technol.* **2012**, *63*, 406–415. [CrossRef]
26. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
27. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3146–3154.
28. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
29. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
30. Kelleher, J.D.; Mac Namee, B.; D'Arcy, A. *Fundamentals of Machine Learning for Predictive Data Analytics*; MIT Press: Cambridge, MA, USA, 2015.
31. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **2012**, *55*, 78–87. [CrossRef]
32. Babenko, T.; Kolesnikova, K.; Bakhtiyarova, Y.; Yeskendirova, D.; Sansyzbay, K.; Sysoyev, A.; Kruchinin, O. Hybrid GNN–LSTM Architecture for Probabilistic IoT Botnet Detection with Calibrated Risk Assessment. *Computers* **2026**, *15*, 26. [CrossRef]
33. Tabassum, A.; Hasan, M.; Ahmed, S.; Tasmin, R. University ranking prediction system by analyzing influential global performance indicators. In Proceedings of the 9th International Conference on Knowledge and Smart Technology, Chonburi, Thailand, 1–4 February 2017. [CrossRef]

34. Zaharia, M.; Chowdhury, M.; Das, T.; Dave, A.; Ma, J.; McCauley, M.; Franklin, M.J.; Shenker, S.; Stoica, I. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation, San Jose, CA, USA, 25–27 April 2012; pp. 15–28.
35. Dean, J.; Ghemawat, S. MapReduce: Simplified data processing on large clusters. *Commun. ACM* **2008**, *51*, 107–113. [CrossRef]
36. Zaharia, M.; Xin, R.S.; Wendell, P.; Das, T.; Armbrust, M.; Dave, A.; Meng, X.; Rosen, J.; Venkataraman, S.; Franklin, M.J.; et al. Apache Spark: A Unified Engine for Big Data Processing. *Commun. ACM* **2016**, *59*, 56–65. [CrossRef]
37. Aliaskarov, S.; Uskenbayeva, R.; Serbin, V.; Bekmurat, O.; Bazarbayeva, U.; Bakhtiyarova, Y.; Sansyzbay, K. Mathematical Approaches and Algorithms in Big Data Architecture and Hybrid System Efficiency. *HighTech Innov. J.* **2025**, *6*, 1013–1034. [CrossRef]
38. Dos Anjos, J.C.S.; Gross, J.L.G.; Matteussi, K.J.; González, G.V.; Leithardt, V.R.Q.; Geyer, C.F.R. Data processing model to perform big data analytics in hybrid infrastructures. *IEEE Access* **2020**, *8*, 170281–170294. [CrossRef]
39. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 2nd ed.; OTexts: Melbourne, Australia, 2018. Available online: https://otexts.com/fpp2/ (accessed on 15 April 2026).
40. Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *WIREs Data Min. Knowl. Discov.* **2020**, *10*, e1355. [CrossRef]
41. Latifis, K.; Sidiropoulos, A.; Evangelidis, G. International university rankings: Composite indicators and methodological approaches. *Front. Res. Metr. Anal.* **2026**, *11*, 1828850. [CrossRef]
42. Hijazi, S.; Moshref, M. A comparative benchmark of machine learning models for predicting university rankings: Insights for higher education analytics. *Int. J. Spec. Educ.* **2026**, *41*, 628–640.
43. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **1904**, *15*, 72–101. [CrossRef] [PubMed]
44. Kendall, M.G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93. [CrossRef]
45. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [CrossRef]
46. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [CrossRef]
47. Moskovkin, V.M.; Zhang, H.; Sadovski, M.V.; Serkina, O.V. Comprehensive quantitative analysis of Top 100s of ARWU, QS and THE World University Rankings for 2014–2018. *Educ. Inf.* **2022**, *38*, 151–171. [CrossRef]
48. Bergmeir, C.; Hyndman, R.J.; Koo, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput. Stat. Data Anal.* **2018**, *120*, 70–83. [CrossRef]
49. Fisher, A.; Rudin, C.; Dominici, F. All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* **2019**, *20*, 1–81.
50. Moed, H.F. A Critical Comparative Analysis of Five World University Rankings. *Scientometrics* **2017**, *110*, 967–990. [CrossRef]
51. Cerqueira, V.; Torgo, L.; Mozetič, I. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Mach. Learn.* **2020**, *109*, 1997–2028. [CrossRef]
52. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The M4 Competition: 100,000 Time Series and 61 Forecasting Methods. *Int. J. Forecast.* **2020**, *36*, 54–74. [CrossRef]
53. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics* **2010**, *26*, 1340–1347. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.