

# Predicting University Rankings Using Random Forest Regression on Institutional Metrics: A Data Mining Approach for Enhancing Higher Education Decision-Making

Min-Tsai Lai<sup>1</sup>, Taqwa Hariguna<sup>2,\*</sup> 

<sup>1</sup>Department of Business Administration, Southern Taiwan University of Science and Technology, Taiwan

<sup>2</sup>Department of Information System and Magister Computer Science, Universitas Amikom Purwokerto, Indonesia

## ABSTRACT

This study investigates the prediction of university rankings using Random Forest regression, leveraging institutional metrics as input features. The primary objective is to enhance the decision-making process in higher education by providing a data-driven model capable of forecasting rankings with greater transparency and accuracy. The research utilizes a comprehensive dataset containing institutional metrics such as research quality, teaching effectiveness, international outlook, and industry impact. Random Forest regression is chosen for its robustness, handling both linear and non-linear relationships between features and the target ranking variable. Feature selection techniques, including correlation analysis and dimensionality reduction, are applied to identify key metrics that influence rankings. Through rigorous model training and hyperparameter tuning, an optimal Random Forest model is developed indicating strong predictive accuracy. Evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and  $R^2$  are used to assess model performance. The feature importance analysis reveals that research quality and research environment have the highest impact on university rankings, followed by teaching and international outlook. These findings align with common assumptions in higher education rankings, while also revealing the potential of less-studied metrics, such as industry impact and international student population, to influence rankings. This study contributes to the field of open education by presenting a transparent and accessible method for predicting university rankings. It empowers students, administrators, and policymakers with a data-driven approach to assess institutional performance. The research also highlights the limitations of current ranking systems and suggests avenues for future studies, including the use of multi-year datasets and alternative machine learning models.

**Keywords** University Rankings, Random Forest, Feature Importance, Higher Education, Data Mining

## Introduction

University rankings have emerged as a crucial element in the higher education landscape, significantly influencing the decisions of students, faculty, and institutional administrators worldwide. These rankings shape perceptions of institutional quality and reputation, driving competition among universities to attain prestigious positions. As global competition intensifies, universities strive to enhance their appeal by focusing on metrics that impact their rankings, such as research output, teaching quality, and internationalization. Rankings hold a

Submitted 27 January 2025

Accepted 9 April 2025


Published 3 June 2025

\*Corresponding author

Taqwa Hariguna,  
taqwa@amikompurwokerto.ac.id

Additional Information and  
Declarations can be found on  
[page 132](#)

DOI: 10.63913/ail.v1i2.10

 Copyright  
2025 Lai and Hariguna

Distributed under  
Creative Commons CC-BY 4.0

unique power to affect funding opportunities, enrollment rates, and even strategic planning at universities, making them integral to the higher education ecosystem.

The reliance on rankings for decision-making is particularly evident among prospective students, who often perceive these rankings as proxies for institutional quality and potential career success [1]. Research highlights that this emphasis can lead to a "Matthew effect," where highly ranked institutions attract more resources, better faculty, and a stronger student body, thus perpetuating their status and widening disparities within the sector [2], [3]. This competitive dynamic pushes universities to adopt strategic measures, sometimes prioritizing short-term gains aligned with ranking criteria over comprehensive educational improvements [4]. Consequently, the global emphasis on rankings continues to shape institutional policies, educational quality, and the broader academic environment.

The increasing reliance on data-driven methods in evaluating educational institutions marks a transformative change in higher education assessment and management. This trend centers on the integration of diverse data sources, sophisticated algorithms, and analytical techniques to improve institutional transparency, performance, and decision-making processes. Traditional ranking methods that primarily emphasized reputational surveys and basic quantitative metrics have gradually given way to comprehensive data analytics approaches. Institutions and policymakers have recognized that solely focusing on static rankings does not capture the multifaceted nature of educational quality and institutional impact. Thus, the adoption of data-driven evaluations reflects a broader movement toward making higher education assessment more evidence-based and accountable [3].

One of the primary benefits of data-driven evaluation lies in its ability to synthesize complex information, offering a multidimensional view of institutional effectiveness. Approaches such as multi-criteria decision-making (MCDM) models, including the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), have been leveraged to analyze and rank institutions based on diverse indicators, such as research performance, student satisfaction, international outlook, and teaching quality [5]. These methodologies facilitate the combination of quantitative measures with qualitative insights, creating a comprehensive picture of institutional standing and performance. Furthermore, predictive analytics has become a cornerstone of data-driven approaches, enabling institutions to identify at-risk students and implement proactive strategies to improve retention and success rates [6]. The utilization of big data analytics allows for the identification of trends and behavioral patterns, guiding evidence-based interventions and enhancing educational outcomes [7].

Understanding the factors that influence university rankings presents a complex challenge due to the variability and multifaceted nature of the methodologies employed by ranking organizations. Traditional ranking systems typically emphasize a narrow set of quantitative metrics, such as research output and reputation surveys, often overlooking critical aspects of institutional performance and effectiveness. The emphasis on specific counting methods, such as the H-index, tends to disproportionately favor institutions with high research volumes in particular fields, notably those with a high number of citations in life sciences [8]. This creates a potential bias, misrepresenting the strengths of universities that prioritize teaching excellence or specialize in less-cited disciplines [8].

Moreover, a phenomenon known as clustering further complicates ranking dynamics; beyond certain thresholds, the influence of specific metrics can disproportionately shift rankings, leading to sudden changes that may not reflect genuine institutional performance improvements [9].

Another critical issue relates to the normalization of data and the weight assignments of various indicators within ranking methodologies. Small adjustments to the weighting of a particular metric can cause significant fluctuations in a university's rank, revealing the subjective nature of many ranking systems [3]. Additionally, exclusions of certain institutions can alter the relative standing of others, suggesting potential fragility and a lack of robustness in ranking schemes [10]. This raises valid concerns about the reliability and validity of rankings as measures of institutional quality, implying that rankings may, at times, be more reflective of the specific methodologies used than of the inherent qualities of the institutions themselves.

Traditional ranking approaches frequently fail to capture the unique contexts and diverse missions of different universities. By assessing institutions as whole entities without accounting for variations across different departments, many ranking systems overlook internal strengths and areas of excellence within each institution [8]. Such oversimplified evaluations may mislead stakeholders who depend on rankings for accurate representations of institutional performance [11]. The focus on quantitative metrics often eclipses important qualitative aspects, including student satisfaction, community engagement, and educational environments that contribute meaningfully to the student experience. Critics argue that these qualitative dimensions, which are essential for a holistic view of institutional success, are frequently disregarded in favor of more easily measurable quantitative criteria [12].

Additionally, the lack of transparency surrounding the methodologies used by ranking organizations complicates efforts to understand what truly drives a university's rank. Stakeholders often lack access to the underlying data or the rationale for choosing specific metrics, creating potential mistrust and ambiguity [13]. As the landscape of higher education becomes increasingly complex and competitive, these limitations underscore the need for more sophisticated, transparent, and inclusive approaches to evaluating university performance. Moving beyond purely quantitative measures and accounting for the diversity of institutional missions can offer a more accurate and meaningful reflection of a university's impact and standing in the global educational landscape.

The growing reliance on transparent, data-driven ranking predictions is of paramount importance for various open education stakeholders, including students, faculty, policymakers, and educational institutions. In an increasingly competitive global landscape, transparency and reliability in rankings are vital for informed decision-making. For students, transparent data-driven rankings provide essential insights into institutional performance, enabling them to select universities that align with their academic and career aspirations [14]. When students can evaluate universities based on clear criteria, such as research output, teaching quality, or student satisfaction, they are better positioned to make informed decisions that shape their educational experiences and long-term career paths [14]. As modern rankings integrate metrics that reflect sustainability and social impact, students are also afforded a way to assess how institutions align with global challenges and personal values [15].

Faculty members also gain significant advantages from data-driven and

transparent ranking systems. Institutional rankings influence hiring practices, grant opportunities, and collaborative initiatives, shaping the professional landscape of academic staff. Transparent rankings allow faculty to better understand their institution's goals and performance metrics, fostering a culture of continuous improvement, accountability, and alignment with institutional strategies [16]. When faculty members are aware of the key metrics driving institutional performance, they are empowered to advocate for essential resources, improve teaching quality, and enhance their research output [17]. This data-driven alignment strengthens the faculty's ability to contribute to institutional reputation and prestige, ultimately benefiting the broader academic community [18].

Policymakers rely heavily on transparent ranking data to shape educational policies and guide resource allocation. A data-driven framework for evaluating institutional performance allows for evidence-based decision-making, which can target systemic disparities and promote educational equity [19]. By understanding the metrics that contribute to a university's standing, policymakers can craft targeted strategies to enhance institutional quality, align educational goals with societal needs, and prioritize initiatives that support broader goals, such as the United Nations Sustainable Development Goals (SDGs) [15]. Transparent data-driven methodologies thus act as a critical tool for promoting accountability and fostering informed policies in higher education [11].

For institutions themselves, data-driven ranking predictions offer valuable benchmarks for strategic planning and performance improvement. By analyzing and understanding ranking data, universities can identify key areas for growth, implement targeted initiatives to enhance their offerings, and strengthen their reputations both locally and globally [14]. The drive for competitive excellence can lead to innovations in curriculum, faculty development, and institutional collaborations, all of which ultimately benefit students and faculty. As institutions increasingly prioritize sustainability, community engagement, and global outreach, their public image and ability to attract diverse talent also improve [20].

Nevertheless, it is important to address the potential biases and limitations present in traditional ranking methodologies. Critics emphasize that an overemphasis on quantitative metrics, at the expense of qualitative measures, often leads to a narrow portrayal of institutional performance [21]. Advocating for a broader range of criteria, including factors like student engagement, community impact, and institutional diversity, is crucial for creating more comprehensive and meaningful evaluations [22].

The primary objective of this study is to develop a predictive model for university rankings based on institutional metrics using Random Forest Regression. As higher education continues to become more data-driven, institutions seek robust methodologies that can evaluate their standing accurately and holistically. Random Forest Regression, a widely recognized ensemble learning method, is particularly effective for handling complex, non-linear data and identifying the relative importance of various predictors. This study leverages institutional data, including metrics such as teaching quality, research output, student-to-staff ratios, and international outlook, to provide a comprehensive and accurate model of university rankings. This approach offers a more nuanced understanding of the factors influencing a university's position and aims to provide actionable insights that extend beyond traditional ranking criteria.

This study's contributions are significant for stakeholders across the education sector, including students, faculty, policymakers, and university administrators. Predictive models that accurately rank institutions based on data-driven metrics empower these stakeholders to make more informed decisions. For students, this model offers a transparent, quantifiable way to evaluate universities based on their priorities, such as academic excellence, research capabilities, or student diversity. Faculty and administrators can utilize these insights to align institutional goals, foster academic improvements, and optimize resource allocation. Additionally, policymakers benefit from an evidence-based perspective that highlights areas for potential investment and reform within the higher education landscape. Overall, this study exemplifies how data-driven methodologies, such as Random Forest Regression, can transform the way university performance is assessed and enhance strategic decision-making in the educational sector.

## Literature Review

### Overview of University Rankings and Institutional Metrics

University ranking systems have emerged as a cornerstone of the higher education sector, shaping the decisions of diverse stakeholders, including students, faculty, policymakers, and institutional administrators. These rankings rely on a range of institutional metrics to assess and compare universities, often focusing on dimensions such as teaching quality, research output, international outlook, and overall reputation. Teaching quality, as one of the most significant metrics, is frequently evaluated using indicators like student-to-staff ratios, which are assumed to reflect the level of personalized attention and instruction students receive [23]. Student satisfaction surveys are also utilized to gauge educational experience and learning outcomes [24]. However, scholars have raised concerns about the subjectivity of these metrics and their limited scope, arguing that they may not fully capture the multifaceted nature of teaching effectiveness within institutions [25].

Research quality is another pivotal component of university rankings, typically measured through quantitative indicators such as publication volume and citation counts, which aim to capture the impact and reach of academic research [26]. Additional indicators, such as prestigious academic awards, further contribute to a university's research reputation and, consequently, its position in various rankings [26]. Nevertheless, the emphasis on quantitative metrics has been criticized for overlooking qualitative dimensions, such as the societal relevance or practical impact of research initiatives [27]. The variability in methodologies across different ranking systems also introduces inconsistencies, as each system employs its own criteria and weightings to assess research output, leading to divergent conclusions regarding university performance [11].

International outlook has gained prominence as a measure of institutional engagement on a global scale. This metric often includes the proportion of international students and faculty, reflecting a university's commitment to diversity, collaboration, and cross-border academic exchange [26]. Collaborative research initiatives across international boundaries also contribute positively to an institution's standing in global rankings [19]. While this focus highlights the value of global engagement, it can place regional and locally focused institutions at a comparative disadvantage, as their contributions may



be undervalued in international ranking schemes [21].

Institutional reputation is another key element in university rankings, frequently assessed through surveys of academic peers and experts within the field [28]. Although reputation scores carry substantial weight in many ranking systems, they are often criticized for their susceptibility to biases and anchoring effects, which may distort perceptions of actual institutional quality [28]. This reliance on reputation can reinforce entrenched hierarchies within the higher education sector, disproportionately favoring institutions with established prestige regardless of current performance metrics [27].

The methodologies used in ranking systems play a central role in shaping the selection and weighting of metrics. Studies have demonstrated that even minor adjustments to indicator weightings can significantly alter university rankings [19]. Different ranking organizations, such as the Academic Ranking of World Universities (ARWU), Times Higher Education (THE), and Quacquarelli Symonds (QS), each apply distinct methodologies, leading to variations in institutional standings and complicating direct comparisons [11]. This variability underscores the necessity for greater transparency and consistency in ranking methodologies, allowing stakeholders to better understand and utilize ranking data for decision-making purposes [19].

In summary, university ranking systems incorporate a diverse range of metrics to assess institutional performance, including teaching quality, research output, international outlook, and reputation. While these metrics offer valuable insights, they also present challenges related to bias, subjectivity, and methodological variability. As higher education continues to evolve, a critical examination of ranking methodologies and the development of more nuanced and transparent approaches are essential for accurately reflecting the multifaceted missions and contributions of universities worldwide.

### **Data Mining and Machine Learning in Education**

The application of data mining and machine learning algorithms in education has transformed how institutions analyze and predict outcomes, including university rankings and other key metrics of institutional performance. Among these algorithms, Random Forest and Support Vector Machines (SVM) have proven particularly effective in handling complex and multidimensional datasets. By leveraging the predictive capabilities of these models, researchers have demonstrated their value in improving the accuracy and depth of educational analyses. For example, Udipi et al. applied regression-based machine learning techniques to assess teaching and learning parameters, successfully predicting the global ranking indices of universities [29]. This work underscores the potential of data-driven approaches to uncover patterns and relationships that traditional statistical analyses may overlook, thereby enhancing predictive accuracy and providing more nuanced insights into university performance.

Similarly, data mining techniques have been employed to explore correlations between bibliometric indicators and university rankings. Szluka's study analyzed how different bibliometric metrics, such as research publications and citations, influence institutional standings within ranking systems [1]. This analysis highlights the complex, multifactorial nature of university rankings and underscores the importance of advanced data analysis techniques to capture these dynamics accurately. In addition to regression models, classification algorithms such as Random Forest and SVM have also been applied to

educational datasets to identify predictive factors of student outcomes and institutional performance. Sharma et al. demonstrated the use of these algorithms in predicting student career paths based on academic performance and other criteria, illustrating their utility in educational decision-making and strategic planning [30].

The utility of machine learning extends beyond ranking prediction to other areas of educational analysis. For example, Rawal and Lal developed a predictive model using the Naïve Bayes classifier to address uncertainties in university admissions processes, illustrating how data mining techniques can streamline operations and improve strategic decision-making in higher education [31]. Moreover, integrating machine learning techniques with traditional ranking methodologies has been explored to enhance the validity and comprehensiveness of ranking systems. Vernon et al. conducted a systematic review emphasizing the need for accurate measures of academic quality, showcasing how machine learning can refine these assessments [27].

In summary, the integration of data mining algorithms in predicting university rankings and other educational metrics represents a critical advancement in the analysis of higher education data. Machine learning methods provide the capability to uncover complex relationships, enhance the predictive accuracy of rankings, and inform data-driven decision-making processes within institutions. As the higher education landscape continues to evolve, these analytical tools are poised to play an increasingly vital role in shaping institutional strategies and improving educational outcomes.

### **Random Forest Regression**

The Random Forest algorithm, a prominent ensemble learning technique, has proven to be highly effective for both classification and regression tasks due to its ability to handle complex, high-dimensional data while mitigating overfitting. This algorithm builds numerous decision trees during the training process, and the output for regression tasks is obtained by averaging the predictions of these individual trees. One of its key advantages lies in the use of bootstrap aggregating, or bagging, which involves training each decision tree on a random subset of the dataset sampled with replacement. This approach ensures that each tree is exposed to a slightly different dataset, leading to a diverse set of models that, when combined, yield robust predictions. Additionally, at each node of the decision tree, a random subset of features is selected to determine the optimal split, which further reduces the correlation among trees and enhances predictive performance [32].

The criterion for node splitting in regression tasks often relies on minimizing the Mean Squared Error (MSE). The formula for MSE is expressed as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

This criterion ensures that the model selects splits that reduce the variance of predictions within the node, thereby enhancing the overall predictive accuracy. The ensemble nature of Random Forest mitigates the risk of overfitting, as averaging predictions across numerous decision trees smooths out noise and captures underlying patterns in the data more effectively [33]. This characteristic makes it particularly suitable for complex datasets where individual decision trees might otherwise overfit to random variations.

Random Forest also offers additional benefits, such as the ability to assess the importance of different features in making predictions. By calculating feature importance scores, this algorithm provides valuable insights into which variables most significantly influence the target outcome, facilitating a deeper understanding of the underlying data structure [34]. Moreover, Random Forest often requires minimal data preprocessing, making it accessible and practical for various real-world applications [35]. Its versatility extends across fields ranging from medical diagnosis to financial forecasting and environmental studies, further illustrating its broad applicability and strong predictive capabilities [36], [37]. Through the power of ensemble learning and robust decision tree construction, Random Forest continues to be a preferred choice for data scientists seeking accurate, reliable predictions.

### **Feature Importance in Machine Learning**

Feature importance within the Random Forest algorithm serves as a critical mechanism for identifying the most impactful metrics influencing predictive outcomes, such as university rankings. This capability is integral to understanding which factors contribute most significantly to institutional performance, enabling data-driven decision-making and resource allocation. Random Forest assesses feature importance through multiple methods, with the most common being Mean Decrease Impurity (MDI) and Mean Decrease Accuracy (MDA). The MDI approach evaluates the contribution of each feature by quantifying how much it reduces node impurity, such as Gini impurity or entropy, during the tree construction process. When a feature frequently splits the data and results in a significant reduction in impurity, it receives a higher importance score [38]. Meanwhile, MDA measures feature importance by evaluating the decline in model accuracy after permuting the values of a particular feature. This process reveals the extent to which the feature contributes to accurate predictions, offering insights into which variables are essential for model performance [38].

In the context of predicting university rankings, feature importance analysis provides a powerful tool for educational institutions to pinpoint the metrics that drive their competitive standing. For example, if metrics such as research output and international faculty representation are determined to be critical, universities may channel resources to enhance these areas, thus improving their overall rankings [39]. Additionally, feature importance can reveal less obvious yet influential factors, such as community engagement or student retention rates, prompting institutions to adopt targeted strategies for improvement. This approach not only guides strategic planning but also enables a deeper understanding of the complex relationships between various institutional metrics and ranking outcomes.

Feature importance in Random Forest also offers practical applications in higher education beyond ranking predictions. Studies have demonstrated that this method can identify key performance indicators for student success, such as attendance, socio-economic background, and prior academic achievement [40]. By focusing on these critical features, institutions can streamline data collection efforts, prioritize resource allocation, and enhance educational outcomes. Moreover, Random Forest's ability to handle non-linear relationships and complex feature interactions allows for a more comprehensive understanding of how diverse metrics impact institutional performance [41]. The robustness of the algorithm further ensures that feature importance scores remain reliable, even



when dealing with high-dimensional datasets, making it a preferred tool for evaluating complex educational data [42].

### **Gaps in the Literature**

Despite the extensive body of research surrounding university rankings, there remains a significant gap in studies that focus on leveraging open educational data and institutional metrics for predictive modeling. Much of the existing literature centers on traditional ranking methodologies, emphasizing metrics such as academic reputation, research output, and student satisfaction. However, these studies often neglect the potential of open educational data to enhance the transparency, inclusiveness, and predictive accuracy of university rankings. For instance, while various research efforts, such as those by Poza et al. [20], have explored sustainability indicators within rankings, empirical investigations into how open educational data could be systematically integrated into predictive models remain limited. This oversight restricts the potential to provide a more holistic view of institutional performance by tapping into freely accessible data sources.

The challenge of integrating open educational data into existing ranking frameworks has been underexplored in the literature. Research [43] discusses the motivations behind universities' engagement with ranking systems but does not delve deeply into how open data can be utilized to improve these frameworks. The lack of emphasis on incorporating open data restricts opportunities to create more comprehensive and inclusive ranking systems. Moreover, existing studies such as those by [44] and [45] recognize the importance of diverse institutional metrics but fall short of examining how open educational data can enhance the evaluation of these metrics, particularly in terms of capturing educational quality, social impact, and other nuanced factors.

The literature also reveals a need for innovative approaches that leverage open data within university ranking systems. While some researchers, such as [46], have analyzed sustainability-focused rankings, their studies do not address how open educational data can create more adaptive and responsive ranking methodologies. This gap highlights the untapped potential for dynamic models that better reflect real-time institutional performance. Additionally, the lack of focus on open educational data limits the potential for enhanced transparency in rankings. Discussions around ranking methodologies, as noted by [4], are critical, yet the inclusion of open data could further improve credibility and reliability by allowing broader access to underlying datasets.

The absence of research on open educational data's role in university rankings carries implications for both policy and practice. Institutions seeking to improve their rankings would benefit from understanding how to effectively utilize open data for strategic planning and evidence-based decision-making. This becomes increasingly relevant in the context of globalization, where universities compete for international recognition and resources [47]. Integrating open data into ranking methodologies offers the potential to bridge existing gaps in knowledge and foster a more equitable and transparent evaluation landscape, ultimately enhancing the overall quality and accessibility of higher education.

### **Method**

The research method for this study consists of several steps to ensure a comprehensive and accurate analysis. The flowchart in figure 1 outlines the detailed steps of the research method.

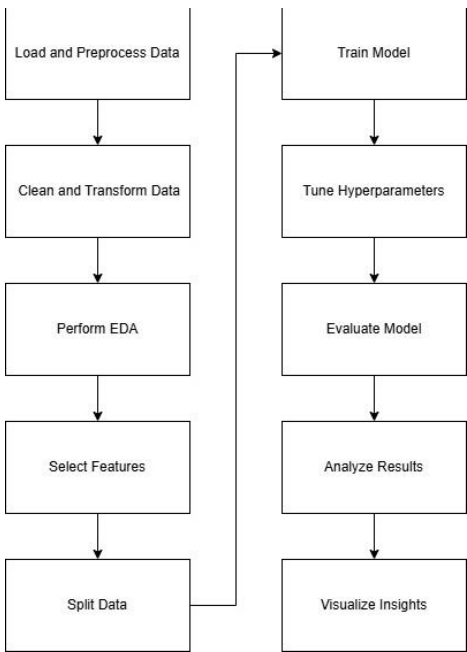


Figure 1 Research Method Flowchart

Dataset Description

The dataset used in this study contains institutional metrics relevant to university rankings, with columns capturing key attributes such as "Rank," "Teaching," "Research Environment," "Industry Impact," "International Students," and "Overall Score." These metrics serve as foundational indicators of institutional performance, with some, like "Rank," representing the target variable for predictive modeling. Descriptive statistics revealed that numerical columns, including "Student Population" and "Overall Score," displayed a wide range of values, highlighting the diversity among institutions in terms of size and performance. Additionally, non-numerical columns such as "International Students" required preprocessing to convert percentage-based values into a usable numerical format. Initial exploration also showed missing values in both numerical and categorical columns, necessitating further cleaning to ensure data integrity.

The dataset summary provided essential insights into the distribution of key variables. For example, the "Overall Score" metric exhibited a relatively broad range, with a mean value of 35.46 and a standard deviation of 16.76, indicating variability in institutional performance. Similarly, the "International Students" metric, expressed as a percentage, provided insights into the global engagement of universities, with some institutions showing minimal international student representation while others had a majority of international enrollees. These variations underscored the dataset's utility in capturing the complex dynamics of institutional rankings.

Data Cleaning

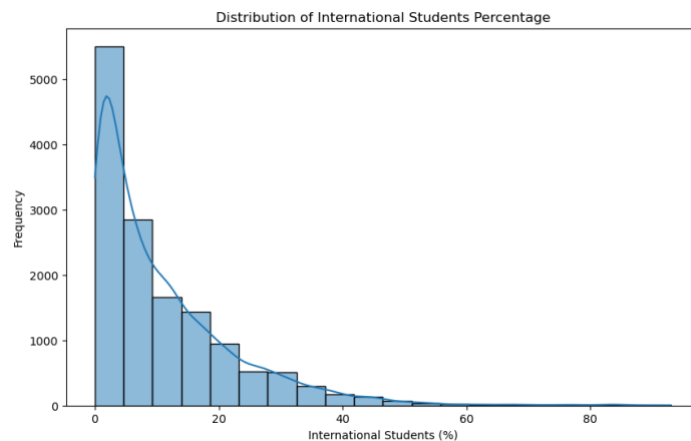
To prepare the dataset for analysis, missing values were addressed using appropriate imputation methods. Numerical columns were filled with their

respective mean values, while categorical columns, such as those containing text-based data, were imputed using their mode. For instance, missing percentages in the "International Students" column were replaced with the column mean after converting the values from text to a numerical format. This ensured that no data points were excluded due to missing values, maintaining the dataset's completeness.

Outliers were identified and handled using the interquartile range (IQR) method. For example, extreme values in the "Overall Score" column were examined and removed if they fell outside the acceptable range defined by 1.5 times the IQR from the first and third quartiles. Additionally, the "International Students" column required cleaning to remove percentage symbols and empty strings, followed by conversion to a numerical format. These steps ensured that the data was appropriately standardized and prepared for subsequent analysis and visualization.

## Visualization

Initial visualizations provided a deeper understanding of the distribution and variability within key metrics. A histogram of the "International Students" column (Figure 2) revealed a right-skewed distribution, indicating that while most institutions had a moderate percentage of international students, a few had exceptionally high values. This visualization highlighted the diversity in global engagement among universities and offered insights into the potential impact of this metric on rankings.



**Figure 2 Distribution of International Students**

A boxplot of the "Overall Score" column (Figure 3) further illustrated the spread of institutional performance scores. The visualization revealed several outliers on the upper and lower ends, underscoring the need for careful handling during data cleaning. The median score was notably lower than the maximum, indicating a significant disparity between top-performing universities and others. Together, these visualizations not only confirmed the dataset's variability but also underscored the importance of preprocessing to account for inconsistencies and outliers.

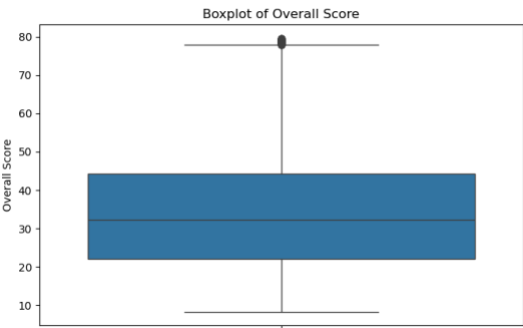


Figure 3 Boxplot of Overall Score

The exploratory data analysis provided a comprehensive overview of the dataset's structure, quality, and key characteristics. Through descriptive statistics, missing value handling, and visualization, this process revealed the underlying trends and patterns in institutional metrics that influence university rankings. The cleaning and transformation steps ensured a consistent and robust dataset, while visualizations highlighted critical insights into key variables. This thorough analysis established a solid foundation for the predictive modeling phase, aligning the data with the study's objective of forecasting university rankings using Random Forest Regression.

Feature Selection and Engineering

The process of feature selection plays a crucial role in determining which metrics most significantly influence university rankings. To achieve this, a combination of correlation analysis and statistical methods was employed to identify the most relevant features for predictive modeling. The correlation matrix of the numerical columns (Figure 4) provided an initial overview of relationships between variables, allowing for the identification of highly correlated features. Metrics such as "Teaching," "Research Quality," and "Research Environment" demonstrated strong correlations with the target variable, "Rank," underscoring their potential importance in the model. A heatmap visualization of the correlation matrix further clarified these relationships, highlighting which metrics might provide the most predictive power.

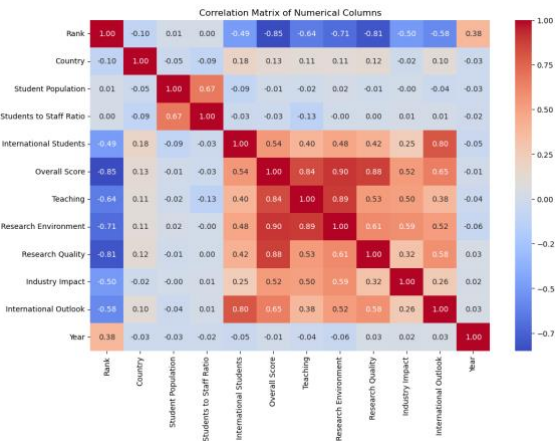
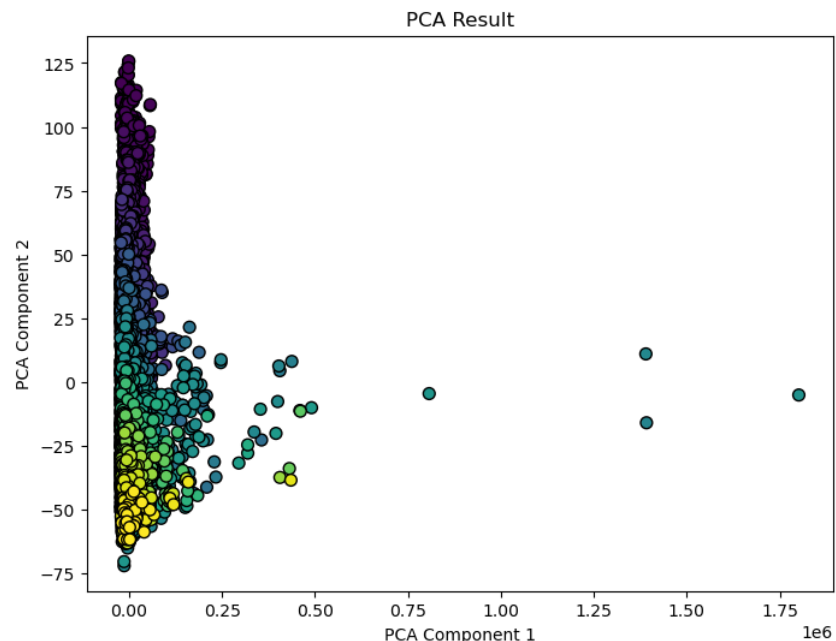


Figure 4 Correlation Matrix

To refine the feature selection process, the SelectKBest method with the `'f_regression'` function was applied. This statistical approach evaluated the linear relationship between each feature and the target variable, assigning scores based on their predictive relevance. Metrics such as "Research Quality," "Research Environment," and "Teaching" received the highest scores, confirming their significance in determining university rankings. This process not only prioritized the most impactful features but also provided a quantitative basis for their inclusion in the final model. Lower-scoring features, such as "Students to Staff Ratio," exhibited weaker associations and were deprioritized for further analysis.

In addition to feature selection, dimensionality reduction techniques were explored to optimize the dataset for predictive modeling. Principal Component Analysis (PCA) was employed to transform the selected features into a smaller set of uncorrelated components while retaining the majority of the data's variance. This approach aimed to reduce redundancy among highly correlated variables and simplify the dataset for model training. PCA revealed that two principal components explained the majority of the variance in the data, indicating that the dimensionality of the dataset could be effectively reduced without significant loss of information.

The results of PCA were visualized through a scatter plot (Figure 5) of the two principal components, providing insights into the structure of the data. Institutions with similar rankings clustered together, reflecting shared characteristics among their metrics. This step demonstrated the utility of dimensionality reduction in capturing the underlying patterns within the dataset while enhancing computational efficiency. Although PCA was not directly integrated into the final model, its application confirmed the robustness of the selected features and supported the overall validity of the dataset for regression analysis.



**Figure 5 Scatter Plot of PCA**



The combined use of correlation analysis, SelectKBest, and PCA ensured that the feature selection and engineering process was both comprehensive and data-driven. These methods prioritized metrics that exhibited strong relationships with the target variable while addressing potential issues of multicollinearity. By leveraging statistical techniques and dimensionality reduction, the dataset was prepared to effectively train the Random Forest Regression model, optimizing its predictive capabilities and aligning the analysis with the study's objective of enhancing decision-making in higher education.

### Random Forest Regression Model

The Random Forest Regression algorithm, a widely used ensemble learning method, was employed to predict university rankings based on institutional metrics. This algorithm operates by constructing multiple decision trees during the training phase, each built using a random subset of the data. For regression tasks, the final prediction is determined by averaging the outputs of all the trees. This approach, known as bagging (bootstrap aggregating), enhances model stability and accuracy by reducing the risk of overfitting. The Random Forest algorithm is particularly suitable for predicting university rankings due to its ability to handle complex, high-dimensional datasets and capture non-linear relationships among features.

To optimize model performance, hyperparameter tuning was conducted using a grid search with cross-validation. The parameters adjusted included the number of decision trees (`n_estimators`), the maximum depth of each tree (`max_depth`), the minimum number of samples required to split a node (`min_samples_split`), and the minimum number of samples per leaf (`min_samples_leaf`). The grid search identified the optimal combination of these parameters to minimize prediction error. The final model, tuned for maximum accuracy, demonstrated its ability to generalize effectively across unseen data, ensuring reliable predictions of university rankings.

### Evaluation Metrics

The performance of the Random Forest model was evaluated using standard regression metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination ( $R^2$ ). MAE measured the average magnitude of errors between predicted and actual rankings, providing an interpretable metric for assessing the accuracy of the model's predictions. MSE, calculated as the mean of squared differences between predicted and actual values, was used to emphasize larger errors in the evaluation. RMSE, derived as the square root of MSE, offered a scaled measure of error, aligning more closely with the original ranking values.

The  $R^2$  score quantified the proportion of variance in the target variable explained by the model, offering an indicator of its predictive strength. The model achieved competitive results across all evaluation metrics, with a high  $R^2$  score indicating robust predictive performance and low MAE, MSE, and RMSE values signifying minimal deviation from actual rankings. A scatter plot comparing actual versus predicted rankings further illustrated the model's effectiveness, with most predictions closely aligning with the diagonal, representing perfect accuracy.

The Random Forest Regression model, equipped with optimized hyperparameters, demonstrated its suitability for predicting university rankings by effectively leveraging the diverse set of institutional metrics. The evaluation metrics provided clear evidence of the model's accuracy and reliability, making it a powerful tool for decision-making in the higher education landscape. This approach underscores the potential of ensemble learning methods to address complex predictive challenges in academic analytics.

## Result and Discussion

### Model Performance

The Random Forest Regression model was evaluated using standard regression metrics, which provided a comprehensive understanding of its predictive accuracy and reliability. The best-performing model parameters, determined through grid search hyperparameter tuning, included a maximum tree depth of 20, a minimum of one sample per leaf, a minimum of two samples per split, and 200 decision trees ('n\_estimators'). These optimal parameters balanced model complexity and accuracy, ensuring robust predictions while minimizing overfitting. The evaluation metrics for the model highlighted its strong performance. The Mean Absolute Error (MAE) of 113.46 indicated that, on average, the predicted university rankings deviated by approximately 113.46 points from the actual rankings. The Mean Squared Error (MSE), calculated as 28,454.29, emphasized the influence of larger errors, which were relatively infrequent. The Root Mean Squared Error (RMSE) of 168.68, a scaled measure of prediction error, further supported the model's accuracy, reflecting its ability to align closely with the actual ranking values. Moreover, the coefficient of determination ( $R^2$ ) score of 0.883 demonstrated that the model explained 88.3% of the variance in the target variable, underscoring its predictive strength.

The performance metrics achieved by the Random Forest Regression model indicated a high level of predictive accuracy, validating the suitability of the algorithm for this application. The low MAE and RMSE values signified that the model effectively captured the underlying patterns in the dataset, enabling accurate predictions of university rankings. The high ( $R^2$ ) score further demonstrated the model's capability to account for the variability in the target variable using the selected institutional metrics, reinforcing the importance of features such as "Research Quality" and "Teaching" in determining rankings. These results were consistent with findings from prior studies, which highlighted the effectiveness of Random Forest in handling complex, non-linear relationships and high-dimensional datasets. The relatively low MAE and RMSE values also suggested that the model successfully minimized prediction errors across a wide range of rankings, making it a reliable tool for decision-making in the context of higher education. The ability of the model to generalize well across unseen data, as indicated by the evaluation metrics, further validated its application in real-world scenarios where accurate and transparent ranking predictions are critical.

### Feature Importance Analysis

The Random Forest Regression model provided feature importance scores that ranked institutional metrics based on their impact on university rankings. The

most influential metric was "Research Quality," with a feature importance score of 27,123.37, significantly surpassing other features. This finding highlighted the critical role of research output and its quality in determining an institution's ranking. Following "Research Quality," "Research Environment" (14,810.78) and "Teaching" (10,237.16) emerged as the second and third most important metrics, respectively. These metrics underscored the importance of robust research ecosystems and high teaching standards as key drivers of institutional success. Metrics such as "International Outlook" (7,343.37) and "Industry Impact" (4,805.06) were moderately influential, reflecting their role in fostering global engagement and practical relevance. In contrast, "Student Population" (0.51) and "Students to Staff Ratio" (0.30) demonstrated negligible impact, suggesting their limited relevance in the predictive model.

### **Insights from Feature Importance Results**

The feature importance analysis revealed critical insights into the underlying factors influencing university rankings. The prominence of "Research Quality" and "Research Environment" aligns with the methodologies of widely recognized ranking systems, which often prioritize research excellence as a key indicator of institutional performance. This finding emphasizes the need for universities to invest in research infrastructure, faculty development, and publication quality to enhance their rankings. Furthermore, the high importance of "Teaching" highlights the dual emphasis on academic excellence and instructional quality, suggesting that balanced efforts in research and teaching contribute significantly to institutional success.

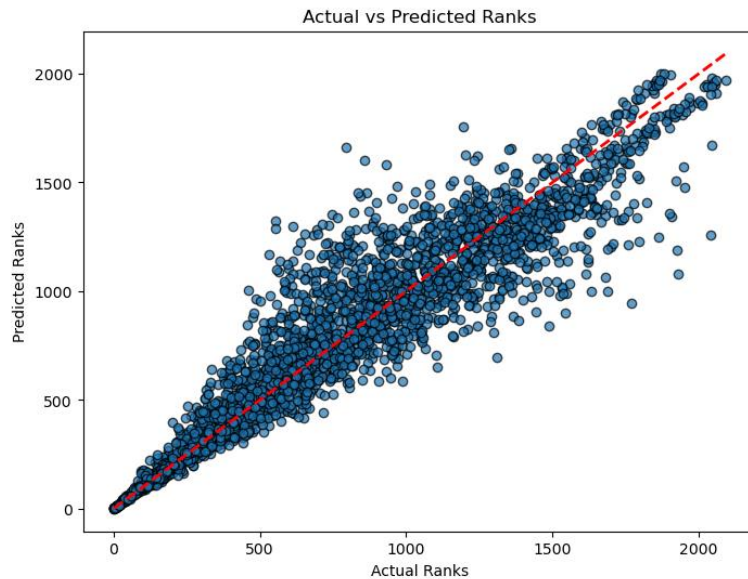
The moderate importance of "International Outlook" and "Industry Impact" suggests that while these metrics are not as dominant as research-related factors, they play a supporting role in establishing institutional reputation and relevance. A strong international presence and active industry collaborations reflect the global engagement and practical applicability of an institution's offerings, factors that increasingly influence stakeholder perceptions. The low impact of metrics such as "Student Population" and "Students to Staff Ratio" provides insight into the limited role of quantitative size indicators in ranking methodologies, which tend to focus more on qualitative performance metrics.

Overall, the feature importance analysis underscores the multifaceted nature of university rankings, where research, teaching, and global engagement collectively determine institutional standing. These insights provide actionable guidance for universities aiming to improve their rankings by prioritizing efforts in areas with the most significant impact. The analysis also highlights the utility of machine learning techniques like Random Forest in uncovering nuanced relationships within institutional data, enabling data-driven decision-making in the education sector.

### **Predicted vs. Actual Ranks**

Figure 6 comparing the predicted and actual university rankings illustrates the effectiveness of the Random Forest Regression model in capturing the underlying relationships between institutional metrics and their corresponding ranks. The red dashed line in the plot represents the ideal scenario where predicted ranks perfectly match the actual ranks. Most data points cluster

around this line, indicating a high degree of alignment between the predicted and actual values. This pattern underscores the model's ability to provide accurate predictions for the majority of universities, thereby validating its suitability for this application.



**Figure 6 Actual vs Predicted Ranks**

Despite the overall accuracy, the scatter plot also highlights areas where the model's predictions deviate from the actual ranks. These deviations are more pronounced in the upper rank ranges, where larger prediction errors can be observed. This discrepancy suggests that the model faces challenges in capturing the nuances of universities with either extremely high or low ranks. Such variations may stem from the inherent complexity of these institutions, as they often exhibit unique characteristics that are difficult to generalize or predict solely based on institutional metrics.

The observed deviations in predictions also point to potential areas for model improvement. One approach could involve incorporating additional metrics that better capture the unique attributes of outlier universities. For example, factors such as regional influences, niche academic programs, or historical reputation might contribute to ranking variability but are not adequately reflected in the current dataset. Addressing these gaps through feature engineering or dataset expansion could enhance the model's predictive accuracy for outlier cases.

Overall, the scatter plot serves as a valuable diagnostic tool for assessing model performance and identifying areas for refinement. The strong alignment of most points with the ideal prediction line reinforces the reliability of the Random Forest Regression model for ranking prediction. At the same time, the discrepancies observed for certain universities provide actionable insights for future research, emphasizing the need for more granular and diverse data inputs to capture the full spectrum of factors influencing university rankings.

## Discussion of Findings

The findings from the Random Forest Regression model highlighted the critical role of specific metrics in determining university rankings. Metrics such as "Research Quality," "Research Environment," and "Teaching" emerged as the most influential features, reflecting their alignment with the methodologies used by major ranking systems. These results reaffirm the widely held assumption that institutions excelling in research output and quality are more likely to secure higher rankings. The prominence of "Research Environment" also underscores the importance of fostering a supportive ecosystem for academic inquiry, which includes funding, infrastructure, and collaboration opportunities.

Conversely, metrics such as "Students to Staff Ratio" and "Student Population" showed minimal influence on rankings, challenging the traditional perception that quantitative indicators of institutional size or student-faculty interaction play a substantial role. While these metrics may contribute to operational efficiency, their negligible impact in this context suggests that rankings prioritize performance-oriented factors like research and teaching excellence over structural attributes. This finding invites stakeholders to reconsider how resources are allocated and evaluated within institutions, focusing more on enhancing academic and research quality.

The model's findings have significant implications for open education and its stakeholders, including students, faculty, and administrators. For students, the results provide valuable insights into the key drivers of university rankings, enabling more informed decision-making when selecting institutions. Students seeking institutions with robust research programs or high teaching quality can leverage these findings to align their academic and career goals with institutional strengths. Moreover, the moderate importance of metrics like "International Outlook" highlights the growing relevance of global engagement, encouraging students to consider universities that foster international collaboration and diversity.

For administrators, the findings offer actionable guidance for strategic planning and policy development. Universities aiming to improve their rankings can focus on enhancing research output, faculty expertise, and teaching methodologies. The results also emphasize the need for institutional transparency in showcasing strengths that align with high-impact metrics. In the context of open education, administrators can use this data to better communicate their institution's value proposition to prospective students and funding bodies, ensuring alignment with stakeholder expectations.

The emphasis on performance-driven metrics also underscores the potential for innovation in ranking methodologies. Incorporating more nuanced and diverse data points, such as sustainability initiatives or societal impact, could provide a broader understanding of institutional effectiveness. These insights are particularly relevant for open education initiatives, which often prioritize accessibility, equity, and social responsibility alongside traditional academic metrics. Leveraging such findings can drive a more inclusive and comprehensive approach to evaluating and enhancing educational institutions.



## Conclusion

This study demonstrated the effectiveness of the Random Forest Regression model in predicting university rankings based on institutional metrics. The model achieved strong predictive performance, as evidenced by an  $R^2$  score of 0.883 and low error values, including a Mean Absolute Error (MAE) of 113.46 and a Root Mean Squared Error (RMSE) of 168.68. These results highlighted the model's ability to explain the majority of the variability in university rankings using features such as "Research Quality," "Research Environment," and "Teaching," which emerged as the most influential metrics in determining institutional performance. Additionally, the feature importance analysis provided actionable insights, confirming the dominance of research-focused metrics while revealing the limited predictive power of structural variables like "Students to Staff Ratio." The study contributes to the field of open education by providing a data-driven framework for university ranking prediction. This approach enhances transparency in the evaluation of institutional performance and equips stakeholders—students, faculty, and administrators—with evidence-based tools for informed decision-making. For students, the findings offer clarity on which institutional attributes are most critical, aiding in their selection of universities that align with their academic and career aspirations. Administrators, on the other hand, can leverage the insights to prioritize strategic investments in high-impact areas such as research infrastructure and teaching excellence, thereby improving their institutional standing. This emphasis on transparency and informed choice aligns with the broader objectives of open education, which seeks to democratize access to educational opportunities and resources.

While the study provides valuable insights, it is not without limitations. The dataset used for analysis focused on a specific set of institutional metrics, which may not fully capture the diverse factors influencing university rankings. Additionally, the dataset represented rankings for a single year, limiting the study's ability to account for temporal trends or shifts in institutional performance over time. These constraints underscore the need for more comprehensive datasets that include multi-year rankings, qualitative metrics, and additional contextual factors such as regional or cultural influences on institutional success. Future research could address these limitations by incorporating broader datasets with diverse and dynamic metrics, including those related to sustainability, social impact, and student satisfaction. Exploring alternative algorithms, such as Gradient Boosting Machines or Neural Networks, could also provide comparative insights into model performance. Furthermore, a longitudinal analysis using multi-year datasets could uncover trends and changes in ranking determinants, offering a deeper understanding of how institutional priorities evolve over time. These avenues for future exploration would enhance the robustness and applicability of data-driven ranking methodologies in higher education.

## Declarations

### Author Contributions

Conceptualization: M.T.L.; Methodology: T.H.; Software: T.H.; Validation: M.T.L.; Formal Analysis: M.T.L.; Investigation: T.H.; Resources: M.T.L.; Data Curation: T.H.; Writing Original Draft Preparation: M.T.L.; Writing Review and

Editing: T.H.; Visualization: M.T.L.; All authors have read and agreed to the published version of the manuscript.

### Data Availability Statement

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

### Funding

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

### Institutional Review Board Statement

Not applicable.

### Informed Consent Statement

Not applicable.

### Declaration of Competing Interest

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

## References

- [1] P. Szluka, "Relationship Between Bibliometric Indicators and University Ranking Positions," *Sci. Rep.*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-35306-1.
- [2] M. N. Bastedo and N. A. Bowman, "U.S. News & World Report College Rankings: Modeling Institutional Effects on Organizational Reputation," *Am. J. Educ.*, vol. 116, no. 2, pp. 163–183, 2010, doi: 10.1086/649437.
- [3] F. Selten, C. Neylon, C. Huang, and P. Groth, "A Longitudinal Analysis of University Rankings," *Quant. Sci. Stud.*, vol. 1, no. 3, pp. 1109–1135, 2020, doi: 10.1162/qss\_a\_00052.
- [4] J. Horstschräer, "University Rankings in Action? The Importance of Rankings and an Excellence Competition for University Choice of High-Ability Students," *Econ. Educ. Rev.*, vol. 31, no. 6, pp. 1162–1176, 2012, doi: 10.1016/j.econedurev.2012.07.018.
- [5] K. A. Özcan, "Sustainability Ranking of Turkish Universities With Different Weighting Approaches and the TOPSIS Method," *Sustainability*, vol. 15, no. 16, p. 12234, 2023, doi: 10.3390/su151612234.
- [6] I. Mustapha, "Data-Driven Insights in Higher Education: Exploring the Synergy of Big Data Analytics and Mobile Applications," *Int. J. Interact. Mob. Technol. Ijtim*, vol. 17, no. 20, pp. 21–37, 2023, doi: 10.3991/ijim.v17i20.45037.
- [7] L.-M. Ang, F. Ge, and K. P. Seng, "Big Educational Data & Analytics: Survey, Architecture and Challenges," *Ieee Access*, vol. 8, pp. 116392–116414, 2020, doi: 10.1109/access.2020.2994561.
- [8] L. Bornmann, R. Mutz, and H. Daniel, "Multilevel-statistical Reformulation of Citation-based University Rankings: The Leiden Ranking 2011/2012," *J. Am. Soc. Inf. Sci. Technol.*, vol. 64, no. 8, pp. 1649–1658, 2013, doi: 10.1002/asi.22857.
- [9] M.-H. Huang and C.-S. Lin, "Counting Methods & University Ranking by H-Index," *Proc. Am. Soc. Inf. Sci. Technol.*, vol. 48, no. 1, pp. 1–6, 2011, doi: 10.1002/meet.2011.14504801191.
- [10] C. Tofallis, "A Different Approach to University Rankings," *High. Educ.*, vol. 63, no. 1, pp. 1–18, 2011, doi: 10.1007/s10734-011-9417-z.
- [11] H. F. Moed, "A Critical Comparative Analysis of Five World University Rankings," *Scientometrics*, vol. 110, no. 2, pp. 967–990, 2016, doi: 10.1007/s11192-016-

- 2212-y.
- [12] N. Erdoğan and M. Esen, "Classifying Universities in Turkey by Hierarchical Cluster Analysis," *Ted Eğitim Ve Bilim*, vol. 41, no. 184, 2016, doi: 10.15390/eb.2016.6232.
  - [13] M. Jarocka, "Transparency of University Rankings in the Effective Management of University," *Bus. Manag. Educ.*, vol. 13, no. 1, pp. 64–75, 2015, doi: 10.3846/bme.2015.260.
  - [14] A. K. Nassa and J. Arora, "Revisiting Ranking of Academic Institutions," *Desidoc J. Libr. Inf. Technol.*, vol. 41, no. 1, pp. 5–19, 2021, doi: 10.14429/djlit.41.1.16673.
  - [15] TorabianJuliette, "Revisiting Global University Rankings and Their Indicators in the Age of Sustainable Development," *Sustain. J. Rec.*, vol. 12, no. 3, pp. 167–172, 2019, doi: 10.1089/sus.2018.0037.
  - [16] C. Burmann, F. García, F. Guijarro, and J. Oliver, "Ranking the Performance of Universities: The Role of Sustainability," *Sustainability*, vol. 13, no. 23, p. 13286, 2021, doi: 10.3390/su132313286.
  - [17] Y. G. Shan, J. Zhang, M. Alam, and P. Hancock, "Does Sustainability Reporting Promote University Ranking? Australian and New Zealand Evidence," *Meditari Account. Res.*, vol. 30, no. 6, pp. 1393–1418, 2021, doi: 10.1108/medar-11-2020-1060.
  - [18] J. V. García and C. Ferreira, "Universities Under Pressure: The Impact of International University Rankings," *J. New Approaches Educ. Res.*, vol. 9, no. 2, pp. 181–193, 2020, doi: 10.7821/naer.2020.7.475.
  - [19] F. Noreen and B. Hussain, "HEC Ranking Criteria in the Perspective of Global University Ranking Systems," *Glob. Soc. Sci. Rev.*, vol. IV, no. II, pp. 43–50, 2019, doi: 10.31703/gssr.2019(iv-ii).06.
  - [20] E. d. I. Poza, P. Merello, A. Barberá, and A. Celani, "Universities' Reporting on SDGs: Using THE Impact Rankings to Model and Measure Their Contribution to Sustainability," *Sustainability*, vol. 13, no. 4, p. 2038, 2021, doi: 10.3390/su13042038.
  - [21] O. H. Sayed, "Critical Treatise on University Ranking Systems," *Open J. Soc. Sci.*, vol. 07, no. 12, pp. 39–51, 2019, doi: 10.4236/jss.2019.712004.
  - [22] L. Saraite-Sariene, M. d. M. Gálvez-Rodríguez, A. Haro-de-Rosario, and C. Caba-Pérez, "Unpackaging Stakeholders' Motivation for Participating in the Social Media of the Higher Education Sector," *Online Inf. Rev.*, vol. 43, no. 7, pp. 1151–1168, 2019, doi: 10.1108/oir-09-2018-0273.
  - [23] A. Meseguer-Martinez, A. Ros-Galvez, A. Rosa-García, and J. A. Catalan-Alarcon, "Online Video Impact of World Class Universities," *Electron. Mark.*, vol. 29, no. 3, pp. 519–532, 2018, doi: 10.1007/s12525-018-0315-4.
  - [24] G. A. Olcay and M. Bulu, "Is Measuring the Knowledge Creation of Universities Possible?: A Review of University Rankings," *Technol. Forecast. Soc. Change*, vol. 123, pp. 153–160, 2017, doi: 10.1016/j.techfore.2016.03.029.
  - [25] R. K. Lukman, D. Krajnc, and P. Glavič, "University Ranking Using Research, Educational and Environmental Indicators," *J. Clean. Prod.*, vol. 18, no. 7, pp. 619–628, 2010, doi: 10.1016/j.jclepro.2009.09.015.
  - [26] L. I. Meho, "Highly Prestigious International Academic Awards and Their Impact on University Rankings," *Quant. Sci. Stud.*, pp. 1–25, 2020, doi: 10.1162/qss\_a\_00045.
  - [27] M. M. Vernon, E. A. Balas, and S. Momani, "Are University Rankings Useful to Improve Research? A Systematic Review," *Plos One*, vol. 13, no. 3, p. e0193762, 2018, doi: 10.1371/journal.pone.0193762.
  - [28] N. A. Bowman and M. N. Bastedo, "Anchoring Effects in World University Rankings: Exploring Biases in Reputation Scores," *High. Educ.*, vol. 61, no. 4, pp. 431–444, 2010, doi: 10.1007/s10734-010-9339-1.
  - [29] P. K. Udupi, V. Dattana, P. S. Netravathi, and J. Pandey, "Predicting Global Ranking of Universities Across the World Using Machine Learning Regression Technique," *SHS Web Conf.*, vol. 156, p. 04001, 2023, doi: 10.1051/shsconf/202315604001.

- [30] S. Sharma\*, S. Pandey, and Prof. K. Garg, "Machine Learning for Predictions in Academics," *Int. J. Recent Technol. Eng.*, vol. 8, no. 5, pp. 4624–4627, 2020, doi: 10.35940/ijrte.e6965.018520.
- [31] A. T. Rawal and B. Lal, "Predictive Model for Admission Uncertainty in High Education Using Naïve Bayes Classifier," *J. Indian Bus. Res.*, vol. 15, no. 2, pp. 262–277, 2023, doi: 10.1108/jibr-08-2022-0209.
- [32] S. Kitano *et al.*, "Development of a Machine Learning Model to Predict Cardiac Arrest During Transport of Trauma Patients," *J. Nippon Med. Sch.*, vol. 90, no. 2, pp. 186–193, 2023, doi: 10.1272/jnms.jnms.2023\_90-206.
- [33] H. Byeon, "Is the Random Forest Algorithm Suitable for Predicting Parkinson's Disease With Mild Cognitive Impairment Out of Parkinson's Disease With Normal Cognition?," *Int. J. Environ. Res. Public. Health*, vol. 17, no. 7, p. 2594, 2020, doi: 10.3390/ijerph17072594.
- [34] R. Rismayati, I. Ismarmiaty, and S. Hidayat, "Ensemble Implementation for Predicting Student Graduation With Classification Algorithm," *Int. J. Eng. Comput. Sci. Appl. Ijecs*, vol. 1, no. 1, pp. 35–42, 2022, doi: 10.30812/ijecs.v1i1.1805.
- [35] B. Imran, H. Hambali, A. Subki, Z. Zaeniah, A. Yani, and M. R. Alfian, "Data Mining Using Random Forest, Naïve Bayes, and Adaboost Models for Prediction and Classification of Benign and Malignant Breast Cancer," *J. Pilar Nusa Mandiri*, vol. 18, no. 1, pp. 37–46, 2022, doi: 10.33480/pilar.v18i1.2912.
- [36] H. A. Khoirunissa, A. R. Widyaningrum, and A. P. A. Maharani, "Comparison of Random Forest, Logistic Regression, and MultilayerPerceptron Methods on Classification of Bank Customer Account Closure," *Indones. J. Appl. Stat.*, vol. 4, no. 1, p. 14, 2021, doi: 10.13057/ijas.v4i1.41461.
- [37] D. P. Hapsari, "Hospital Length of Stay Prediction With Ensemble Learning Methode," *J. Appl. Sci. Manag. Eng. Technol.*, vol. 4, no. 1, pp. 29–36, 2023, doi: 10.31284/j.jasmet.2023.v4i1.4437.
- [38] X. Man and E. P. Chan, "The Best Way to Select Features? Comparing MDA, LIME, and SHAP," *J. Financ. Data Sci.*, vol. 3, no. 1, pp. 127–139, 2020, doi: 10.3905/jfds.2020.1.047.
- [39] Y. Liu and H. Zhao, "Variable Importance-weighted Random Forests," *Quant. Biol.*, vol. 5, no. 4, pp. 338–351, 2017, doi: 10.1007/s40484-017-0121-6.
- [40] Y. Sanit-in and K. R. Saikaew, "Prediction of Waiting Time in One-Stop Service," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 3, pp. 322–327, 2019, doi: 10.18178/ijmlc.2019.9.3.805.
- [41] G. Riddick *et al.*, "Predicting in Vitro Drug Sensitivity Using Random Forests," *Bioinformatics*, vol. 27, no. 2, pp. 220–224, 2010, doi: 10.1093/bioinformatics/btq628.
- [42] L. Guo, C. Wang, D. Zhang, and G. M. Yang, "An Improved Feature Selection Method Based on Random Forest Algorithm for Wind Turbine Condition Monitoring," *Sensors*, vol. 21, no. 16, p. 5654, 2021, doi: 10.3390/s21165654.
- [43] M. S. Alholiby, "A Qualitative Exploration of Motivations and Challenges for Universities Seeking to Join the University Rankings Race," *J. Educ. Res.*, vol. 46, no. 4, pp. 13–40, 2022, doi: 10.21608/jfees.2022.279479.
- [44] A.-A. Haghdoost, N. Momtazmanesh, F. S. S. Aria, and H. Ranjbar, "Educational Ranking of Medical Universities in Iran (ERMU)," *Med. J. Islam. Repub. Iran*, pp. 736–742, 2018, doi: 10.14196/mjiri.32.126.
- [45] A. Puzatykh, "Russian Institutions of Higher Education in International Rankings: The Problem Social and Environmental Sustainability," *E3s Web Conf.*, vol. 458, p. 06003, 2023, doi: 10.1051/e3sconf/202345806003.
- [46] B. Galleli, N. E. B. Teles, Joyce Aparecida Ramos dos Santos, M. S. Freitas-Martins, and F. Hourneaux, "Sustainability University Rankings: A Comparative Analysis of UI Green Metric and the Times Higher Education World University Rankings," *Int. J. Sustain. High. Educ.*, vol. 23, no. 2, pp. 404–425, 2021, doi: 10.1108/ijshe-12-2020-0475.
- [47] A. Ahmadi, M. Taghavinia, S. K. S. Arabshahi, and M. Ghasemi, "The Interacting

