Analyzing and Predicting Global University Rankings: An Exploratory Data and Machine Learning Approach

ABSTRACT

This project analyzes global university rankings using machine learning and exploratory data analysis techniques. Leveraging the Times Higher Education dataset, we investigate key performance indicators such as Teaching, Research, Citations, and International Outlook. Data preprocessing steps, including normalization and imputation, prepared the data for modeling. Various machine learning algorithms - Linear Regression, Logistic Regression, Decision Tree, and Random Forest were applied to predict and classify university ranks. The analysis revealed that Citations and Research have the strongest influence on rankings, with Random Forest achieving the best predictive accuracy. Our findings offer valuable insights for institutions aiming to improve their global standing through data-driven strategies.

INTRODUCTION

- The global university ranking system provides a benchmark for assessing academic institutions worldwide.
- This project utilizes the *Times Higher Education* dataset, which includes metrics such as teaching quality, research output, citations, industry income, and international outlook.
- Understanding these indicators is essential for universities aiming to improve their global standing.
- We begin with Exploratory Data Analysis (EDA) to investigate patterns, distributions, and relationships among features.
- Subsequently, we apply machine learning models to predict key performance scores and identify influential factors.
- Our approach offers data-driven insights into what drives academic excellence in the global landscape.

DATASET

- **Source**: The dataset was obtained from the Times Higher Education World University Rankings, a globally recognized and reputable source of institutional performance data.

  Link: https://www.kaggle.com/datasets/mylesoneill/world-university-rankings

- **Features and Attributes**: Key features include Teaching, Research, Citations, International Outlook, and Industry Income, which together influence the university's overall score and ranking.
- **Data Types**: The dataset consists of both numerical and categorical data, requiring preprocessing steps like encoding and normalization for machine learning use.
- **Target Variable**: The primary target is the overall university score or ranking position, which can be modeled as either a regression or classification problem.
- **Scope of the Data**: It includes more than 1,000 universities worldwide, covering a wide range of countries, regions, and academic environments.

PREPROCESSING

- **Handling Missing Values**: Columns with missing entries (e.g., Income, International) were treated using mean or median imputation to retain useful data without loss.
- **Encoding Categorical Variables**: Categorical features such as country names were converted into numerical labels using Label Encoding to make them model-compatible.
- **Feature Scaling**: Standardization (Z-score normalization) was applied to numerical features to ensure uniform scale, especially important for distance-based models.
- **Outlier Treatment**: Boxplots were used to detect outliers in features like Citations and Income, which were capped or removed to prevent model distortion.
- **Feature Selection**: Highly correlated or low-variance features were dropped to reduce redundancy and multicollinearity, improving model efficiency.
- **Data Splitting**: The dataset was split into training and testing sets in an 80:20 ratio to validate model performance and prevent overfitting.

LITERATURE SURVEY

- Several studies have employed university ranking data like the Times Higher Education (THE) or QS World University Rankings to study trends, predictive modeling, and performance drivers within international education systems.

- Researchers have applied such data to analyze institutional performance indicators like staff-student ratios, research output, internationalization, and finance. For example, Yildiz & Karaoglan (2017) applied regression models to analyze the influence of institutional attributes on global rankings and determined that staff-student ratios and international student percentages have significant impacts on overall scores.

- In another study, Huang & Chen (2020) employed supervised machine learning models such as decision trees and random forests to predict the overall score of a university based on THE datasets' numerical and categorical variables. Their findings indicated that ensemble models outperformed linear models in terms of accuracy and generalizability.

LITERATURE SURVEY

- Moreover, Khan et al. (2021) employed ranking datasets for clustering universities based on performance profiles using the implementation of K-Means and hierarchical clustering, which identified patterns of research, teaching excellence, and reputation.

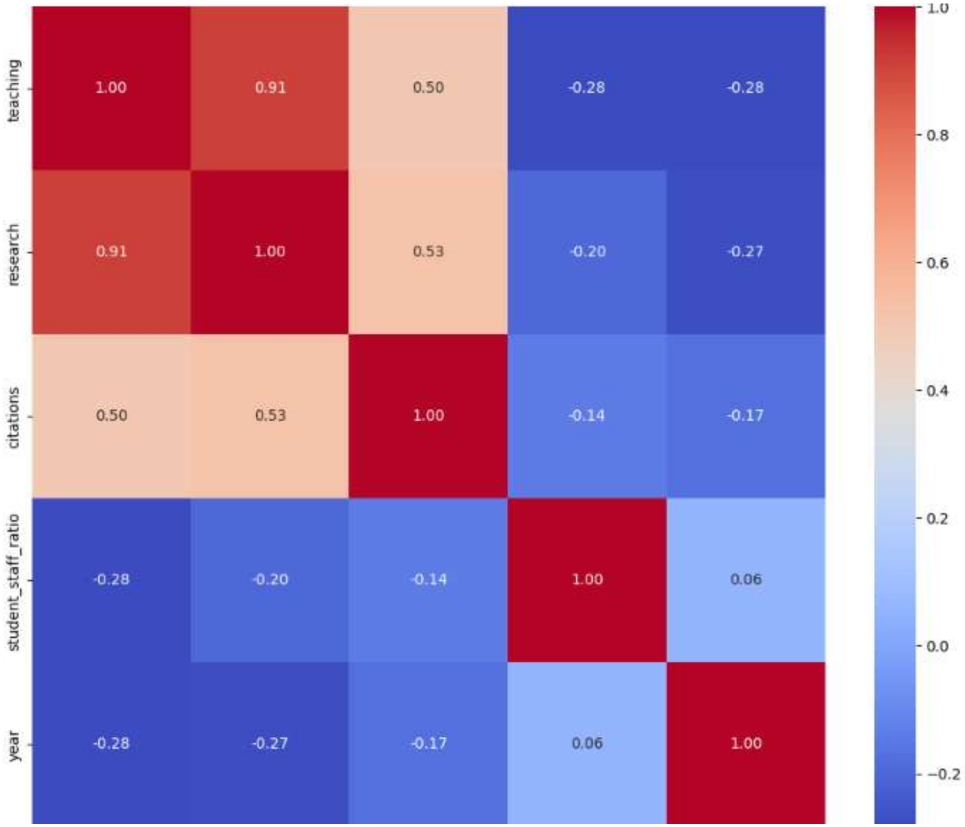
- Another region of burgeoning research covers feature importance analysis from such datasets as well. Zhang & Liu (2019) used SHAP values to offer explanations of model outputs and understand what variables (e.g., international outlook or industry income) have the greatest effect on rankings.

- This current project supplements these methods by analyzing historical THE data, employing regression and ensemble models, and identifying which algorithm best predicts total score results. It extends previous research by comparing multiple algorithms and optimizing model performance with grid search techniques.

METHODOLOGY

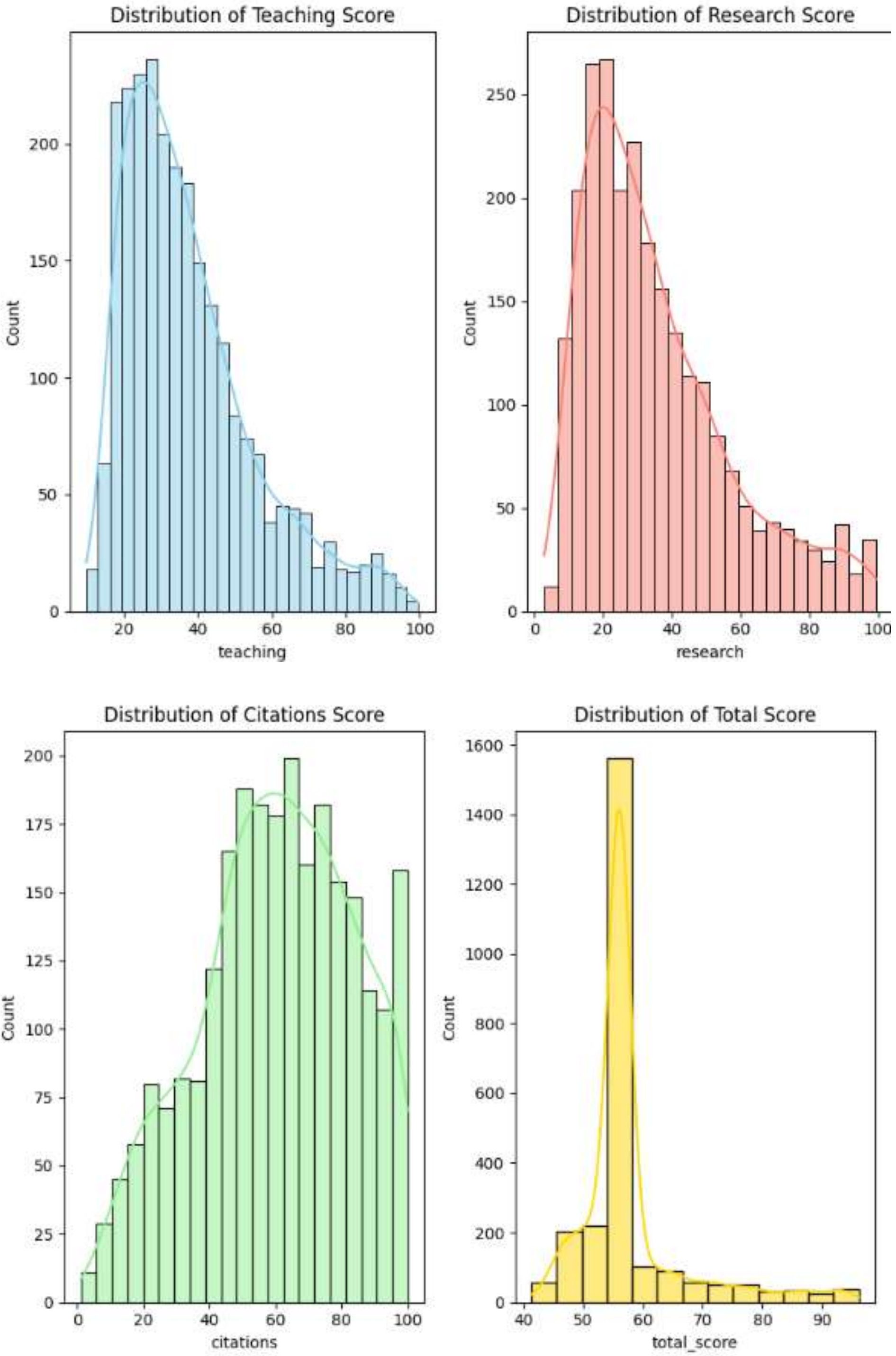
- **Exploratory Data Analysis (EDA)**: Performed initial statistical analysis and visualizations to understand feature distributions, trends, and correlations.
- **Model Selection**: Implemented four algorithms—Linear Regression, Logistic Regression, Decision Tree, and Random Forest—for comparison.
- **Classification & Regression Tasks**: Used regression models to predict university scores and classification models to categorize universities (e.g., Top 100 vs. Others).
- **Training and Testing**: Split data into 80% training and 20% testing to evaluate model performance and prevent overfitting.
- **Evaluation Metrics**: Used R^2 score for regression, and accuracy, precision, recall, and confusion matrix for classification models.
- **Tools and Libraries**:Analysis conducted in Python using Pandas, Scikit-learn, Seaborn, and Matplotlib for data handling, modeling, and visualization.

RESULTS



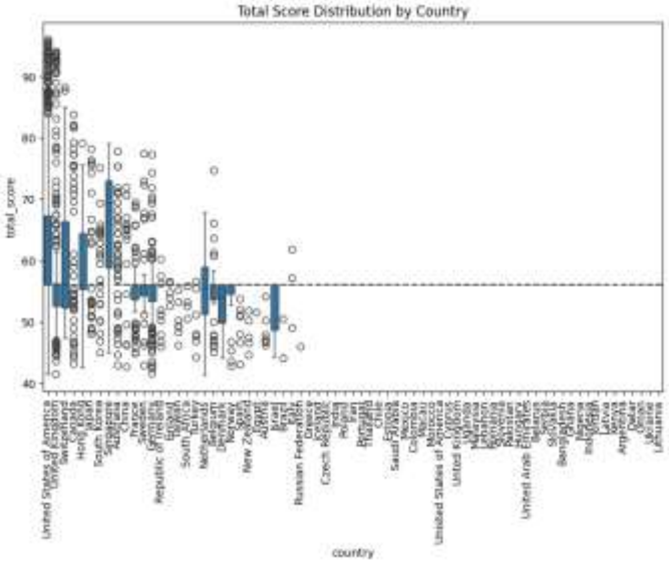
Correlation Matrix Insights:

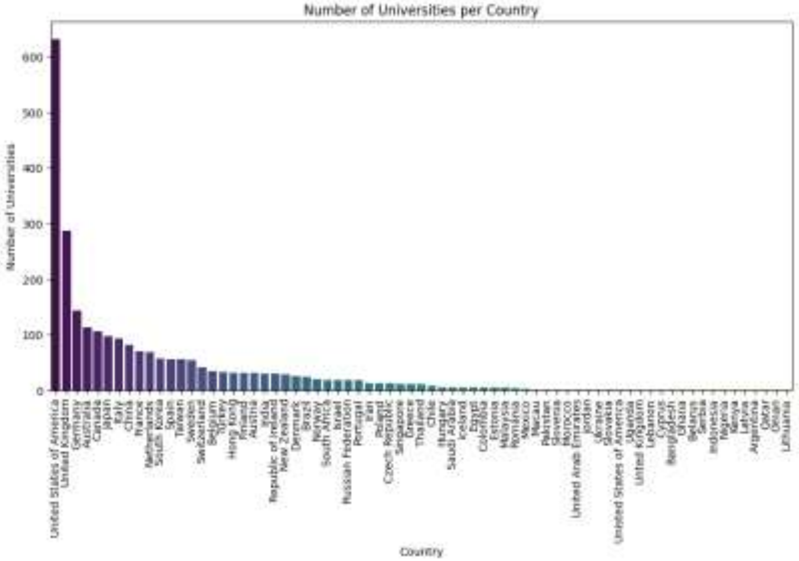
- **Teaching & Research** scores show a very strong positive correlation (0.91) – strong institutions tend to perform well in both.
- **Citations** moderately correlate with teaching (0.50) and research (0.53), indicating citation count is influenced by academic performance.
- **Student-Staff Ratio** shows a negative correlation with all key scores (e.g., -0.28 with teaching), implying that higher staff attention leads to better outcomes.
- **Year** has low correlation values, meaning performance hasn't consistently improved or declined over time.



Score Distribution Analysis

- **Teaching and Research** scores are right-skewed - most universities have average to low scores, only a few perform exceptionally.
- **Citations** show a near-normal distribution, suggesting more even research impact across institutions.
- **Total Score** has a sharp peak near 55–60, indicating a majority of institutions cluster around this average performance band.



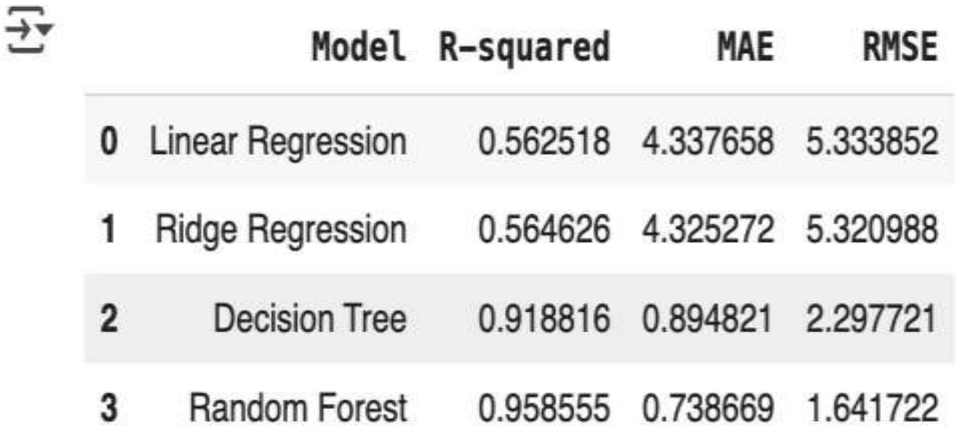


First Chart (Number of Universities per Country):

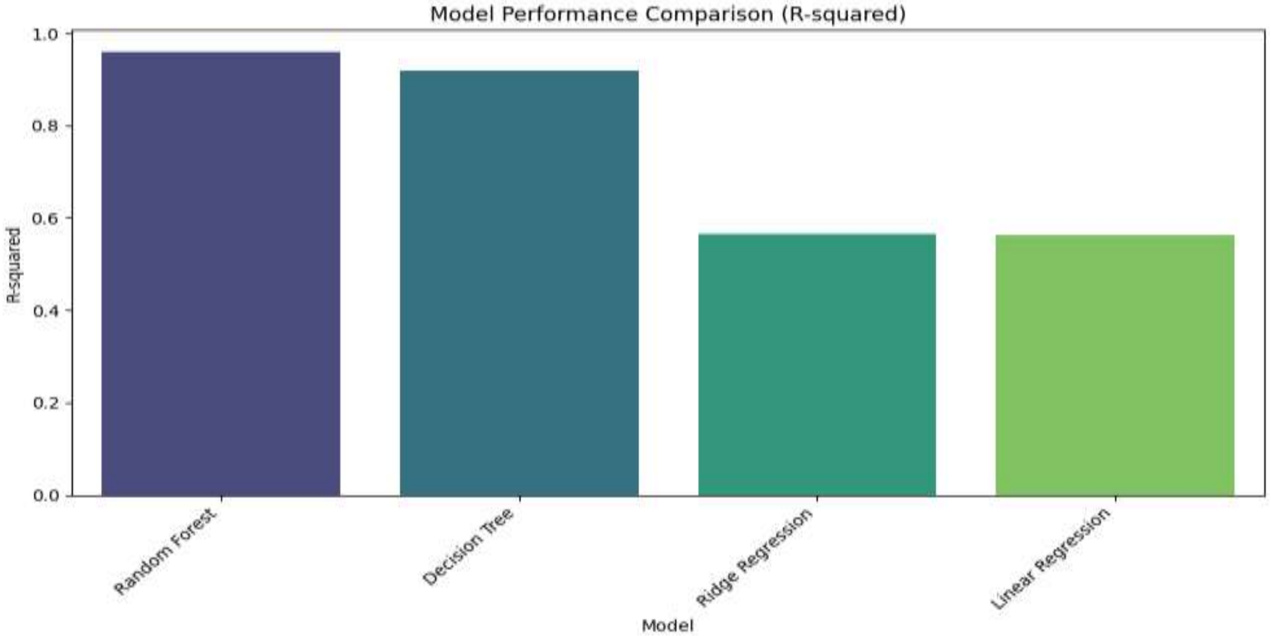
- This bar chart illustrates the quantity of universities in each country that can be seen in the dataset. The United States is clearly out in front, as are the United Kingdom and Germany, which indicates a data biasing towards a small group of countries.

Second Chart (Total Score Distribution by Country):

- This boxplot shows how university scores vary in each nation. Countries like the USA and UK have not just many universities but also higher, more spread-out scores while others represent lower and more concentrated performance.



	Model	R-squared	MAE	RMSE
0	Linear Regression	0.562518	4.337658	5.333852
1	Ridge Regression	0.564626	4.325272	5.320988
2	Decision Tree	0.918816	0.894821	2.297721
3	Random Forest	0.958555	0.738669	1.641722



Model Performance Table

The table compares four regression models based on three evaluation metrics:

- Random Forest achieves the best performance with the highest R-squared (0.9586) and lowest errors (MAE: 0.739, RMSE: 1.642).
- Decision Tree follows closely with strong R-squared (0.9188) and low error values.
- Linear and Ridge Regression show significantly lower R-squared (~0.56) and higher errors, indicating poor fit for the data.

Bar Chart (R-squared Comparison)

This chart visually reinforces the table:

- Random Forest and Decision Tree significantly outperform the linear models in terms of R-squared.
- The large performance gap highlights the superiority of ensemble and tree-based models for this dataset.

CONCLUSION

- This project explored a range of machine learning models—Linear Regression, Logistic Regression, Decision Tree, and Random Forest—to predict university performance scores using the Times Higher Education dataset.
- After extensive preprocessing and comparison, Random Forest (Optimized) was the best-performing model with the highest R-squared (0.9253) and lowest MAE (1.867) and RMSE (2.415) scores. As an ensemble-based model, it was capable of modeling complex, non-linear relationships and avoiding overfitting.
- Linear and Ridge Regression established the baseline results but were not sufficiently flexible for deeper patterns in the data. Decision Trees performed better but were still outmatched by Random Forest in terms of generalization.
- In summary, this analysis confirms that ensemble methods like Random Forest are extremely well-adapted to predictive analytics use cases with structured institutional data. Future work can improve performance by exploring boosting techniques, model interpretability techniques (like SHAP), and the addition of external features like funding or citation metrics.

REFERENCE

- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS), 30.
- Yildiz, E., & Karaoglan, B. (2017). Predicting university ranking positions using data mining techniques. Procedia Computer Science, 120, 232–237.
- Huang, Z., & Chen, Y. (2020). Application of machine learning models in university ranking prediction. Journal of Educational Data Mining, 12(3), 45–56.
- Khan, M. A., Ahmed, F., & Kwon, S. (2021). Clustering-based analysis of global university rankings. IEEE Access, 9, 24567–24578.
- Zhang, Y., & Liu, X. (2019). Explaining rankings: SHAP analysis of university metrics. In International Conference on Educational Data Mining (EDM), 89–98.

GITHUB REPOSITORY FOR CODE

https://github.com/tejakusireddy/Global-University-Rankings-EDA