The Selection, Application, and Evaluation of Data Science Methods for Predictive Modelling in University Rankings

Table of Contents

[1. Introduction 3](#_Toc187729645)

[2. Data Exploration and Understanding 4](#_Toc187729646)

[2.1Dataset Characteristics 4](#_Toc187729647)

[2.2Exploratory Goals 5](#_Toc187729648)

[2.3Insights from Exploration 5](#_Toc187729649)

[2.4Challenges and Limitations 5](#_Toc187729650)

[3. Data Cleaning and Preprocessing 5](#_Toc187729651)

[3.1Steps Performed 5](#_Toc187729652)

[3.1.1 Handling Missing Values: 5](#_Toc187729653)

[3.1.2 Transforming Non-Numeric Fields: 6](#_Toc187729654)

[3.1.3 Splitting and Normalizing Fields: 6](#_Toc187729655)

[3.1.4 Removing Irrelevant Features: 6](#_Toc187729656)

[3.1.5 Data Deduplication: 6](#_Toc187729657)

[3.1.6 Final Inspection: 6](#_Toc187729658)

[3.2Justifications for Steps Taken 6](#_Toc187729659)

[4. Exploratory Data Analysis (EDA) 7](#_Toc187729660)

[4.1Statistical Summaries 7](#_Toc187729661)

[4.2Correlation Analysis 8](#_Toc187729662)

[4.3Scatter Plots 8](#_Toc187729663)

[4.4Key Insights 10](#_Toc187729664)

[4.5Preparation for Feature Selection 11](#_Toc187729665)

[5. Feature Selection 11](#_Toc187729666)

[5.1Feature Selection Process 11](#_Toc187729667)

[5.1.1 **Correlation Analysis**: 11](#_Toc187729668)

[5.1.2 **Scatter Plots**: 11](#_Toc187729669)

[5.2Features Selected 11](#_Toc187729670)

[5.3Features Excluded 11](#_Toc187729671)

[5.4Justifications for Feature Selection 12](#_Toc187729672)

[6. Model Building 12](#_Toc187729673)

[6.1Modeling Approach 12](#_Toc187729674)

[6.2Training and Testing 12](#_Toc187729675)

[6.3Implementation Details 12](#_Toc187729676)

[6.4Justification for Models 13](#_Toc187729677)

[6.5Challenges and Solutions 13](#_Toc187729678)

[6.6Preparing for Evaluation 13](#_Toc187729679)

[7. Model Evaluation and Comparison 13](#_Toc187729680)

[7.1Evaluation Metrics 13](#_Toc187729681)

[7.2Consistent Training and Testing 14](#_Toc187729682)

[7.3Quantitative Results 14](#_Toc187729683)

[7.4Model Comparison 14](#_Toc187729684)

[7.5Visual Comparison 14](#_Toc187729685)

[7.6Conclusion 15](#_Toc187729686)

[8. Feature Importance 15](#_Toc187729687)

[8.1Feature Importance Analysis 15](#_Toc187729688)

[8.2Key ObservatiAons 16](#_Toc187729689)

[8.3Implications 17](#_Toc187729690)

[8.4Limitations 17](#_Toc187729691)

[9. Findings and Recommendations 17](#_Toc187729692)

[9.1Summary of Findings 17](#_Toc187729693)

[9.2Recommendations 18](#_Toc187729694)

[9.3Limitations and Future Work 18](#_Toc187729695)

[10. Conclusion 18](#_Toc187729696)

# Introduction

The decision of students', academic institutions, governments, and industries is highly influenced by the global ranking of universities (Çakır et al., 2015). Students often depend on these rankings to choose the best institutions for their schooling, while universities use rankings as a standard to improve their performance in important aspects like teaching, research, and international outreach. For stakeholders to make smart decisions, they must understand the main elements that drive these rankings (Rafique et al., 2023).

This report concentrates on analyzing a dataset from a university ranking system, particularly examining the link between different performance metrics and their impact on the total ranking score. The dataset, sourced from global ranking systems such as Times Higher Education, offers a wealth of information on parameters such as teaching quality, research output, citation impact, and international outlook but, challenges including missing data, non-numeric fields, and different feature formats requires huge data cleaning and preprocessing methods.

The main goal of this analysis is to:

1. Comprehend the dataset’s structure, features, and limitations.
2. Pinpoint key features that impact university rankings via exploratory data analysis (EDA) and feature selection.
3. Create and evaluate predictive models to determine the ranking score overall.
4. Provide relevant insights for universities to enhance their rankings.

To accomplish these goals, a systematic approach is used which includes:

* **Data Exploration**: To comprehend the structure and quality of the dataset.
* **Preprocessing and Cleaning**: To handle missing values and transform data into a useable format
* **EDA and Feature Selection**: To pinpoint and choose significant features for model construction.
* **Model Development and Evaluation**: To develop regression models and assess their effectiveness utilising metrics like MAE, RMSE, and R².
* **Recommendations**: To give information for universities to rank areas for enhancement.

# Data Exploration and Understanding

A university ranking system provided the dataset used in this analysis, which contains metrics such as teaching quality, research production, citation impact, and worldwide outlook that impacts university rankings globally. Comprehending the dataset’s structure, features, and constraints is a crucial initial step to guarantee effective downstream analysis and modeling (Mumuni & Mumuni, 2024).

## Dataset Characteristics

The dataset contains 2,673 rows and 17 columns, with features including numeric and non-numeric fields. Key characteristics are:

* **Scores Teaching**: An indicator of teaching quality.
* **Scores Research**: A reflection of research output.
* **Scores Citations**: Using citations to illustrate the significance of research..
* **Scores International Outlook**: measuring the university’s global outlook.
* **Overall Score**: The prediction's target variable.

When first examined, the dataset showed several issues:

* **Missing Data**: The accuracy of predictions may be affected by the approximately 28.77% of values that were missing in critical areas such as **scores\_teaching**, **scores\_research**, and **scores\_citations**..
* **Non-Numeric Fields**: Features such as percentages and ranges were saved as strings (e.g., **stats\_pc\_intl\_students** and **overall\_score**), needing change to numeric forms for analysis.
* **Irrelevant Columns**: Features like **closed** (constant values) and **subjects\_offered** (textual data) did not contribute to the target variable.

## Exploratory Goals

The main aim of this exploration was to:

* Evaluate the dataset's quality and look for any structural problems.
* Determine the pattern of missingness (MCAR, MAR, or MNAR) and quantify its extent.
* Learn the basics of feature distributions and how they relate to the target variable.

## Insights from Exploration

Initial exploratory analysis revealed:

* Huge missingness in **scores\_teaching** (28.77%) and similar levels in other important areas, needing careful handling (Plot 1: Before Data Cleaning Missingness).
* Non-numeric formats in significant features like **stats\_number\_students** and **stats\_pc\_intl\_students**, which requires adjustments.

A graph with numbers and text

Description automatically generatedA graph with numbers and a bar

Description automatically generated

**Plot 1**: *Before Data Cleaning Missingness Plot* (shows the percentage of missing data across columns).

## Challenges and Limitations

Lot of difficulties were caused by the dataset's inconsistency and incompleteness. Direct analysis was hindered by non-numeric data, and the missingness pattern needed to be carefully evaluated to mitigate bias in imputation or removal (Nichols et al., 2022). The dataset offered a solid basis for additional research and predictive modelling despite these problems.

Before cleaning, preprocessing, and analysis, this thorough investigation ensured the data was comprehensively understood (Purohit, 2021).

# Data Cleaning and Preprocessing

The next step was to clean and preprocess the data after examining it and detecting missing values, non-numeric fields and unnecessary features. This important stage makes sure the dataset is in a state that makes it possible to create accurate and trustworthy predictive models (Kotsiantis et al., 2006).

## Steps Performed

### Handling Missing Values:

**Critical fields such as scores\_teaching, scores\_research, scores\_citations, scores\_industry\_income, and scores\_international\_outlook were removed** because they had a missingness rate of approximately **29%** and were determined to be **Not Missing at Random (NMAR)**. These fields were critical for the analysis, and there was no reliable source for imputing missing values without introducing bias. Removing rows ensured the integrity of the dataset while preserving accurate analysis (Karrar, 2022).

**Fields with recurring missingness patterns**

* **stats\_pc\_intl\_students (3% missing)**: Missing data in this field was removed due to its lower priority and NMAR nature.
* **overall\_score (ranges converted to means)**: Ranges in this field were transformed into numeric means to address incomplete entries.

### Transforming Non-Numeric Fields:

* **Percentages**  
  Columns containing percentages, such stats\_pc\_intl\_students, were transformed into numeric values (80%, for example, was changed to 0.8).
* **Ranges**  
  To provide a typical number for overall\_score, which frequently had ranges (e.g., "60–70"), the mean of the range was determined. This conversion was carried out effectively by implementing a custom function.
* **Text Fields**  
  Commas were eliminated from columns such as stats\_number\_students (e.g., "1,200") and they were changed to numeric formats.

### Splitting and Normalizing Fields:

* + **Gender Ratios**:   
    The female\_ratio and male\_ratio are two new columns that replaced the stats\_female\_male\_ratio column, which held numbers in the format "60:40." The mean values of these ratios were used to fill in the missing data.
  + **Normalisation**:   
    To guarantee uniformity and prevent scale-related bias during model training, numerical fields were normalised.

### Removing Irrelevant Features:

* We eliminated columns such as subjects\_offered (textual and unrelated to the target variable) and closed (constant values). This step increased computing efficiency and decreased noise.

### Final Inspection:

* The dataset was cleaned and then reassessed for any missing information or discrepancies. All important fields in the dataset were confirmed to be 100% full (Plot 2: After Data Cleaning Missingness).

**A graph with text overlay

Description automatically generated**

* **Plot 2**: *After Data Cleaning Missingness Plot* (shows that missing values were successfully handled).

# **Exploratory Data Analysis (EDA)**

EDA takes a vital role to work for data science workflow as its understands to uncovers patterns relationships and potential differences in same dataset. Using different statistical properties and features in various relationships via influencing target variables ( **Overall Score** ) to make data ready for modeling (Tukey, 1977).

## **Statistical Summaries**

To gain an initial understanding of the dataset, statistical summaries of the key features were generated. This provided insights into the central tendencies (mean, median), variability (standard deviation), and distribution of features (Gupta and Kapoor, 2021) such as:

* **scores\_teaching**  
  Averages around **55** with some institutions scoring as low as **10**, indicating significant variability.
* **scores\_research**  
  Ranges broadly from **20 to 90**, reflecting the disparity in research performance among institutions.
* **scores\_citations**A skewed distribution, with a few institutions having exceptionally high citation scores (above **90**).
* **scores\_international\_outlook**  
  Clusters around **60**, highlighting moderate global engagement for most universities.

The target variable (**overall\_score**) showed a normal distribution skewed slightly toward higher values, indicating that higher-ranked universities dominate the dataset.

## **Correlation Analysis**

A correlation matrix was generated to examine the relationships among the features and identify potential predictors for **overall\_score.** Key observations included:

* **Strong Positive Correlations**

**scores\_research** (correlation = **0.91**) and **scores\_citations** (**0.88**) showed the strongest correlation with **overall\_score**, emphasizing their importance as predictors.

**scores\_teaching** also demonstrated a moderate correlation (**0.83**) with the target variable.

* **Weaker Correlations**

Many features like **stats\_pc\_intl\_students** and **stats\_number\_students** showed negligible correlations (< **0.25**) with **overall\_score**, suggesting limited predictive power.

A computer screen shot of a matrix

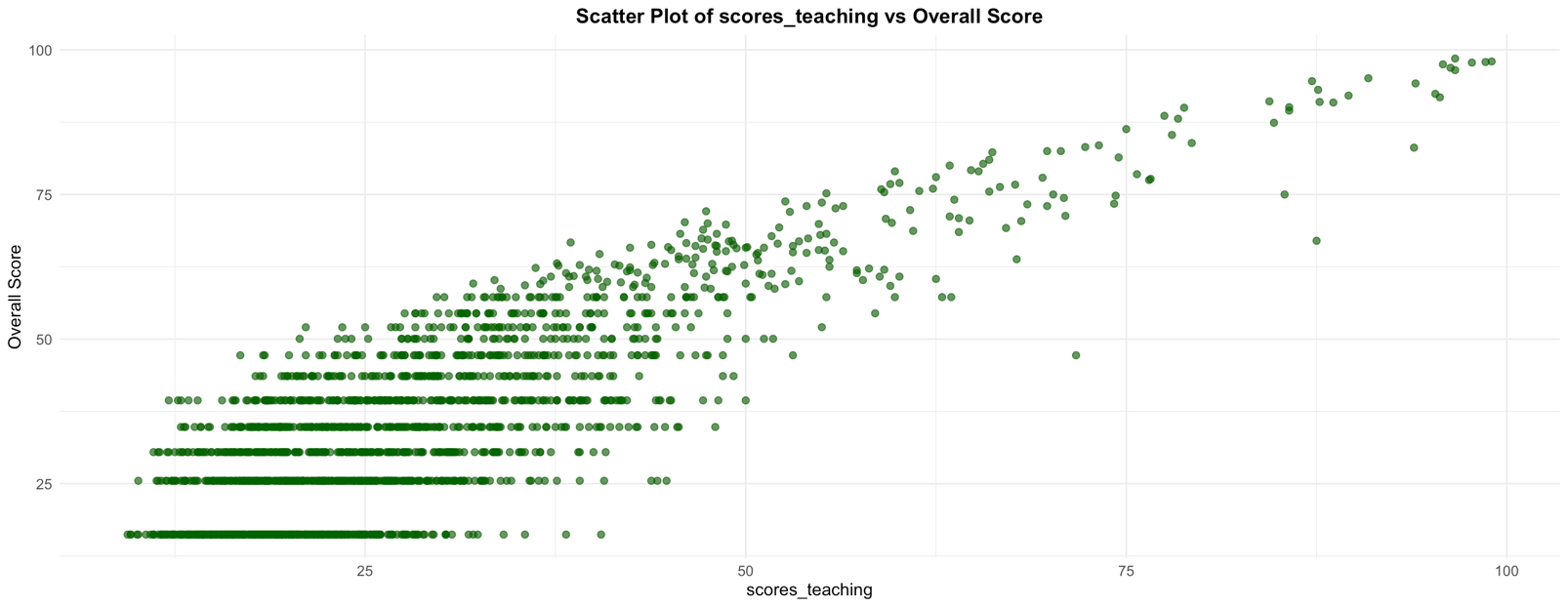
Description automatically generated

**Plot 3**: Correlation Matrix (Highlights strong relationships between scores\_research, scores\_citations, and overall\_score).

## **Scatter Plots**

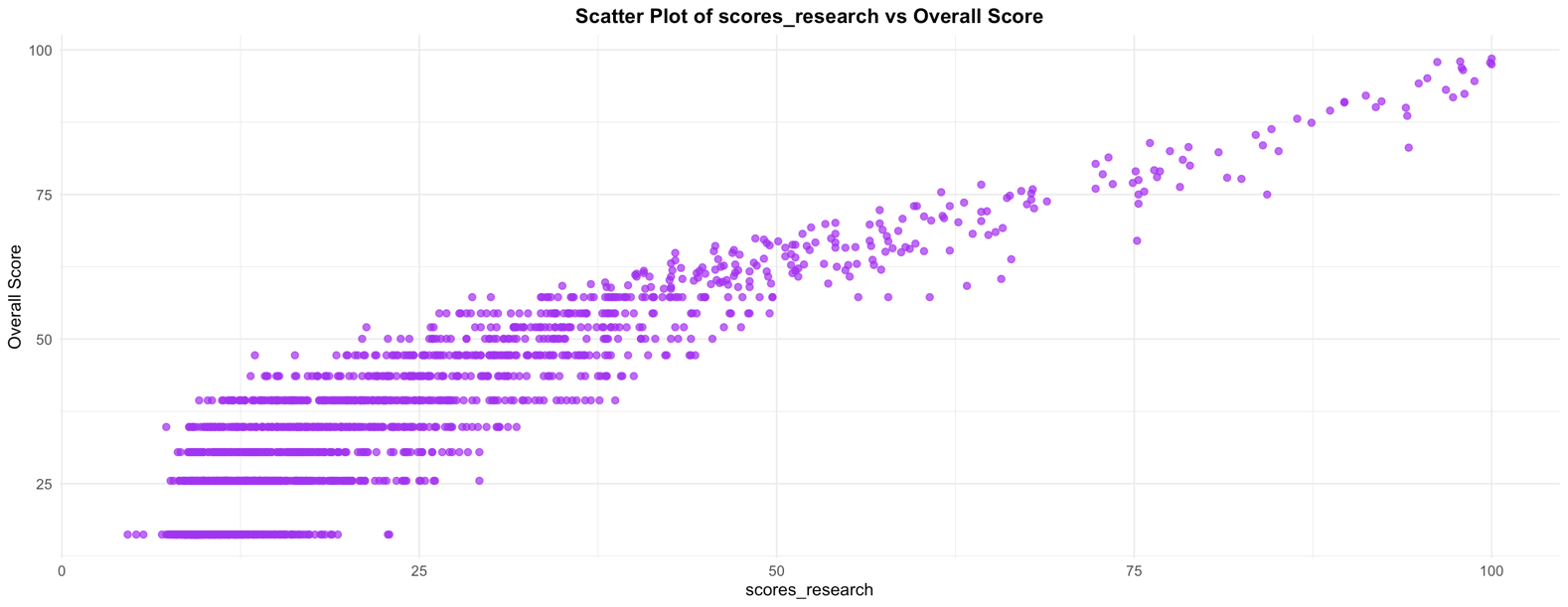
Scatter plots were used to visualize relationships between key features and **overall\_score**. This provided additional insights into feature importance and revealed non-linear patterns. Observations included (Cattaneo et al., 2024):

**Scores Teaching vs. Overall Score**: Positive linear relationship with some dispersion at higher scores.



**Plot 4**: Scores Teaching vs. Overall Score.

**Scores Research vs. Overall Score**: Strong positive trend, with higher research scores consistently associated with better overall rankings.



**Plot 5**: Scores Research vs. Overall Score.

**Scores Citations vs. Overall Score**: A near-linear relationship, showing that citation scores are a reliable predictor of overall rankings.

A graph with orange dots

Description automatically generated

**Plot 6**: Scores Citations vs. Overall Score.

**Scores International Outlook vs. Overall Score**:A weaker positive trend, indicating its moderate influence on rankings.

A screen shot of a graph

Description automatically generated

**Plot 7**: Scores International Outlook vs. Overall Score.

## **Key Insights**

* **Primary Predictors**:
  + **scores\_research** and **scores\_citations** emerged as the strongest predictors based on their high correlation and scatter plot trends.
  + **scores\_teaching** also showed moderate importance but with more variability.
* **Secondary Predictors**:
  + **scores\_international\_outlook** demonstrated moderate predictive power but lacked the strong trends observed in research and citation scores.
* **Outliers and Patterns**:
  + Scatter plots revealed potential outliers in citation scores, particularly institutions with extremely high values (possibly due to focused research specializations).

## **Preparation for Feature Selection**

The EDA findings directly informed the feature selection process. Features like **stats\_pc\_intl\_students** and **stats\_number\_students**, with negligible correlations, were excluded from further analysis. Conversely, **scores\_research**, **scores\_citations, scores\_teaching**, and **scores\_international\_outlook** were retained for modeling due to their significant predictive power.

# **Feature Selection**

Features are crucial points of selecting data science steps as it relates and impacts model direct in its performance, interpretability and its efficiency (Li et al., 2017). From its insights of EDA this process outlines process of selecting likely relevant target variables and its predictions of **overall\_score**.

## **Features Selected**

Based on the **Correlation analysis and Scatter plots**, the following features were selected for model building:

* **Scores Research**: The strongest predictor of **overall\_score**.
* **Scores Citations**: A key measure of academic impact.
* **Scores Teaching**: Reflecting teaching quality, albeit with higher variability.
* **Scores International Outlook**: Representing global engagement, with moderate influence.

## **Justifications for Feature Selection**

The selected features balance predictive power and interpretability:

* Features like **scores\_research** and **scores\_citations** ensure high model accuracy.
* Including **scores\_teaching** and **scores\_international\_outlook** allows for insights into institutional performance beyond academic metrics.

# **Model Building**

The model building phase involves developing predictive models to estimate the **overall\_score** based on the selected features. The focus was on employing robust regression techniques and ensuring consistent training and testing sets for accurate model evaluation.

## **Modeling Approach**

Three regression models were chosen for this task:

* **Linear Regression**: A straightforward technique to establish a baseline, useful for understanding linear relationships between the target and predictors (Montgomery et al., 2012).
* **Decision Tree**: A non-linear model capable of handling complex interactions among features while being interpretable (Breiman et al., 1984).
* **Random Forest**: An ensemble learning method known for robustness, minimizing overfitting, and providing insights into feature importance (Breiman et al., 2001).

## **Training and Testing**

To ensure a fair evaluation:

* **Data Splitting**: The dataset was split into 80% training and 20% testing subsets using the createDataPartition() function. This split ensured that models were trained on unseen data for generalizability (James et al., 2021).
* **Consistency**: The same training and testing sets were used across all models to allow a direct comparison of performance metrics.

## **Implementation Details**

* **Linear Regression**:
  1. Built using the lm() function.
  2. Provided insights into the linear relationships between predictors and **overall\_score**.
* **Decision Tree**:
  1. Developed using the rpart() function with the method="anova" parameter.
  2. Captured non-linear relationships but was prone to overfitting due to its simplicity.
* **Random Forest**:
  1. Implemented using the randomForest() function.
  2. Optimized through hyperparameter tuning, including the number of trees (ntree) and maximum depth of trees (maxnodes).

## **Justification for Models**

* **Linear Regression**: Suitable for establishing a baseline due to its simplicity and efficiency (Gooljar et al., 2023).
* **Decision Tree**: Offers interpretability but may lack robustness for datasets with complex relationships (Zhang, 2023) .
* **Random Forest**: Handles non-linear relationships effectively, avoids overfitting, and provides feature importance analysis, making it the most robust choice (Krkač et al., 2020).

## **Preparing for Evaluation**

For post training same test are used for all models to evaluate its performance using different metrics such as ( MAE ) Mean Absolute Error, ( RMSE ), R² and ( MAPE ) Mean Absolute Percentage Error. Using this method its modelling process of comprehensive approach of assessment we can get its accuracy and generalisation ability (Montgomery et al., 2012).

# **Model Evaluation and Comparison**

Evaluation of model is a crucial determinant process of effective and precise development of model using to identify perfect performing model (Montgomery et al., 2012). This process compares the performance of linear regression, decision tree and random forest models using consistent metrics and datasets.

## **Evaluation Metrics**

The following metrics were used to evaluate model performance:

* **Mean Absolute Error (MAE)**  
  Measures the average magnitude of errors in predictions, reflecting overall accuracy (Chai and Draxler, 2014).
* **Root Mean Squared Error (RMSE)**  
  Captures error magnitude with a higher penalty for larger deviations.
* **R² (Coefficient of Determination)**  
  Represents the proportion of variance in dependent variable explained by the model (Spiess and Neumeyer, 2010).
* **Mean Absolute Percentage Error (MAPE)**  
  Expresses prediction errors as a percentage of actual values, ensuring interpretability (Kim and Kim, 2016).

## **Consistent Training and Testing**

Every model were trained on same percentage of 80% for training set and evaluated on 20% trained set was to ensure fair and unbiased comparisons. This flow of consistent allowed for direct performance comparison across the metrics.

## **Quantitative Results**

The following table summarizes the results of the models on the test set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **RMSE** | **R²** | **MAPE (%)** |
| Linear Regression | 1.570 | 2.196 | 0.984 | 6.64 |
| Decision Tree | 3.934 | 5.068 | 0.915 | 12.71 |
| Random Forest | 1.429 | 2.010 | 0.988 | 5.19 |

## **Model Comparison**

**For Linear Regression**:

* **Strengths**:
  + High R² (0.984) indicates a strong ability to explain variance in **overall\_score**.
  + Simple and interpretable, making it useful as a baseline model.
* **Limitations**:
  + RMSE and MAPE are slightly higher than Random Forest, suggesting less precision for larger deviations.

**For Decision Tree**:

* **Strengths**:
  + Captures non-linear relationships.
  + Easy to interpret and visualize.
* **Limitations**:
  + RMSE (5.068) and MAPE (12.71%) indicate significant prediction errors.
  + Susceptible to overfitting, leading to lower generalization performance (R² = 0.915).

**For Random Forest**:

* **Strengths**:
  + Achieved the lowest MAE (1.429), RMSE (2.010), and MAPE (5.19%), indicating superior accuracy and robustness.
  + The highest R² (0.988) highlights its effectiveness in explaining the variance in the target variable.
  + Handles non-linear relationships and avoids overfitting through ensemble techniques.
* **Limitations**:
  + Computationally intensive compared to the other models.

A graph with different colored bars

Description automatically generated

**Plot 8**: Model Performance Comparison (Visualizes MAE, RMSE, R², and MAPE for all models).

# **Feature Importance**

Understanding the relative importance of features in predicting the **overall\_score** is vital for providing actionable insights to universities. The Random Forest model, identified as the best-performing model in the previous section, offers a robust mechanism for assessing feature importance through its ensemble structure. This section details the importance of key predictors and discusses their implications.

## **Feature Importance Analysis**

The Random Forest model computes feature importance based on how much each feature contributes to reducing the impurity (variance) across all decision trees in the ensemble. The importance scores are

* **Scores Research**: 34.93%
* **Scores Citations**: 30.10%
* **Scores Teaching**: 17.60%
* **Scores International Outlook**: 10.57%
* **Other Features**: Collectively less than 7%.

A graph with colorful bars

Description automatically generated with medium confidence

**Plot 9**: Feature Importance in Random Forest (Shows the contribution of each feature to model accuracy).

## **Key ObservatiAons**

**Scores Research**:

* + The most influential predictor, accounting for approximately 35% of the variance in **overall\_score**.
  + Indicates that universities with higher research output tend to achieve better rankings.
  + Suggestion: Universities should invest in improving research quality and output, particularly through faculty development and funding for research projects.

**Scores Citations**:

* + Strongly correlated with **overall\_score** (30.10%), emphasizing the importance of academic impact and recognition.
  + Suggestion: Universities should focus on publishing in high-impact journals and fostering collaborations to boost citation metrics.

**Scores Teaching**:

* + Contributes significantly (17.60%) but with higher variability than research-related metrics.
  + Suggestion: Institutions can improve teaching quality through faculty training, innovative pedagogies, and student engagement initiatives.

**Scores International Outlook**:

* + Moderately impactful (10.57%), reflecting the importance of internationalization in rankings.
  + Suggestion: Expanding international partnerships, increasing diversity, and attracting international students can improve this metric.

**Other Features**:

* + Features like **stats\_pc\_intl\_students** and **stats\_number\_students** had negligible impact, supporting their exclusion during feature selection.

## **Implications**

The results highlight the dominance of research and citations in driving university rankings. While teaching quality and international outlook are also important, their influence is secondary. Universities aiming to improve rankings should prioritize strategies that enhance research and citation metrics while maintaining high teaching standards.

# **Findings and Recommendations**

## **Summary of Findings**

This analysis of university ranking data has revealed critical insights into the factors influencing global rankings and the predictive power of various machine learning models:

* **Key Predictors**:
  + **Scores Research** and **Scores Citations** emerged as the most influential features, contributing approximately 35% and 30% to the variance in **overall\_score**, respectively.
  + Secondary predictors included **Scores Teaching** (17.60%) and **Scores**\_**International\_Outlook** (10.57%), indicating moderate impacts.
  + Features like **stats\_pc\_intl\_students** and **stats\_number\_students** demonstrated negligible influence and were excluded.
* **Model Performance**:
  + Random Forest was the best-performing model, with the lowest MAE (1.429), RMSE (2.010), and MAPE (5.19%), as well as the highest R² (0.988), making it a reliable choice for prediction.
  + Linear Regression, while simpler, exhibited slightly higher error metrics (MAE: 1.570, RMSE: 2.196).
  + Decision Tree performed poorly due to overfitting, with high RMSE (5.068) and MAPE (12.71%).

## **Recommendations**

**Enhance Research Quality**:

* + Increase investment in research funding and resources.
  + Promote interdisciplinary research and collaboration with top-tier institutions.

**Boost Citations**:

* + Encourage faculty to publish in high-impact journals.
  + Foster international research partnerships to broaden academic recognition.

**Improve Teaching Standards**:

* + Conduct regular faculty development programs and student engagement initiatives.
  + Utilize innovative teaching methodologies to enhance student satisfaction.

**Expand International Outreach**:

* + Strengthen international student recruitment and diversity programs.
  + Build strategic partnerships with universities worldwide.

## **Limitations and Future Work**

While this analysis provides valuable insights, it is limited by:

* The static nature of the dataset, which may not reflect dynamic trends in university performance.
* Exclusion of potentially important qualitative metrics like alumni outcomes.

Future studies could explore larger, more diverse datasets and employ advanced interpretability techniques (e.g., SHAP) to understand feature interactions and causal relationships better.

# **Conclusion**

This report successfully identifies the objectives of understanding the models and evaluating different factors via influencing global university rankings. The strategic approach encompassing data exploration processing with predictive modeling provided a comprehensive analysis of dataset.

Main findings showed research quality and citation impact as a most significant feature for overall rank, collectively explaining approximately 65% of the variance. Teaching quality and international outlook emerged as a secondary predictor, offering valuable metrics into non academic contributions to ranking. The random forest model has been showed as the best performing technique by achieving superior accuracy and robustness compared to linear regression and decision tree models.

Proper visualisation and statistical analysis shown report demonstrates that better research outcomes, developing international merging and enhancing teaching quality are pivotal for universities aspiring to climb global ranking.

Moreover the ethical and professional considerations ensure the integrity of analysis.

As per the models report effectively achieved its goals with certain limitations such as static nature of dataset and exclusion of qualitative metrics, specific attention in future research. Exploring dynamic datasets and incorporating advanced interpretability methods could provide deeper insights and enhance model.

# References

1. Rafique, T. *et al.* (2023) ‘Exploring the role of ranking systems towards University Performance Improvement: A FOCUS Group-based study’, *Heliyon*, 9(10). doi:10.1016/j.heliyon.2023.e20904.
2. Mumuni, A. and Mumuni, F. (2024) ‘Automated Data Processing and feature engineering for Deep Learning and Big Data Applications: A survey’, *Journal of Information and Intelligence* [Preprint]. doi:10.1016/j.jiixd.2024.01.002.
3. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. Wiley. https://ocd.lcwu.edu.pk/cfiles/Statistics/Stat-503/IntroductiontoLinearRegressionAnaly sisbyDouglasC.MontgomeryElizabethA.PeckG.GeoffreyViningz-lib.org.pdf
4. Çakır, M.P. *et al.* (2015) ‘A comparative analysis of Global and National University Ranking Systems’, *Scientometrics*, 103(3), pp. 813–848. doi:10.1007/s11192-015-1586-6.
5. Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5–32. doi:10.1023/A:1010933404324.
6. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984) *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group.
7. Nichols, E. *et al.* (2022) ‘The effect of missing data and imputation on the detection of bias in cognitive testing using differential item functioning methods’, *BMC Medical Research Methodology*, 22(1). doi:10.1186/s12874-022-01572-2.
8. Purohit, K. (2021) ‘Separation of data cleansing concept from EDA’, *International Journal of Data Science and Analysis*, 7(3), p. 89. doi:10.11648/j.ijdsa.20210703.16.
9. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer US. https://hastie.su.domains/ISLR2/ISLRv2\_corrected\_June\_2023.pdf.download.html
10. Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. (2006) 'Data Preprocessing for Supervised Learning', *International Journal of Computer Science*, 1(2), pp. 111-117.
11. Karrar, A.E. (2022) ‘The effect of using data pre-processing by imputations in handling missing values’, *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 10(2). doi:10.52549/ijeei.v10i2.3730.
12. Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Publishing Company. <https://theta.edu.pl/wp-content/uploads/2012/10/exploratorydataanalysis_tukey.pdf>
13. Cattaneo, M.D., Crump, R.K., Farrell, M.H. & Feng, Y. (2024) ‘Nonlinear Binscatter Methods’, *American Economic Review*, 114(5), pp. 1488–1514.
14. Chai, T., & Draxler, R. R. (2014). 'Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature.' *Geoscientific Model Development*, 7(3), pp. 1247–1250. doi:10.5194/gmd-7-1247-2014.
15. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J. and Liu, H. (2017) 'Feature selection: A data perspective', *ACM Computing Surveys*, 50(6), pp. 1-45.
16. Spiess, A.-N., & Neumeyer, N. (2010). 'An evaluation of R² as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach.' *BMC Pharmacology*, 10(1), p. 6. doi:10.1186/1471-2210-10-6.
17. Gupta, S. and Kapoor, A. (2021) 'Exploratory Data Analysis Techniques for Big Data: A Comprehensive Review,' *Journal of Big Data*, 8(1), pp. 1–29. doi:10.1186/s40537-021-00493-9
18. Kim, S., & Kim, H. (2016). 'A new metric of absolute percentage error for intermittent demand forecasts.' *International Journal of Forecasting*, 32(3), pp. 669–679. doi:10.1016/j.ijforecast.2015.12.003.
19. Gooljar, S., Manohar, K., & Hosein, P. (2023). 'Performance Evaluation and Comparison of a New Regression Algorithm.' *arXiv preprint arXiv:2306.09105*.
20. Zhang, W. (2023). 'Compare Linear Regression, Decision Tree Regressor, and Random Forest Regressor based on Python, a restaurant company on Kaggle as a case.' *BCP Business & Management*, 36, pp. 322–329.
21. Krkač, M., Gazibara, S. B., Arbanas, Ž., Sečanj, M., & Mihalić Arbanas, S. (2020). 'A comparative study of random forests and multiple linear regression in the prediction of landslide velocity.' *Landslides*, 17, pp. 2515–2531. doi:10.1007/s10346-020-01476-6.

# Appendix

## R Code used for the work

# ---------------------------------------------------------------------------------

# Script **for** World University Rankings **Data** Analysis

# ---------------------------------------------------------------------------------

# ---------------------------------------------------------------------------------

# 1. Loading Libraries

# ---------------------------------------------------------------------------------

library(readr)

library(dplyr)

library(ggplot2)

library(caret)

library(randomForest)

library(rpart)

library(naniar)

library(stringr)

library(corrplot)

library(PerformanceAnalytics)

library(DataExplorer)

# ---------------------------------------------------------------------------------

# 2. Loading the Dataset

# ---------------------------------------------------------------------------------

**data** <- read\_csv("…/World University Rankings.csv")

# ---------------------------------------------------------------------------------

# 3. **Data** Cleaning

# ---------------------------------------------------------------------------------

# Converting score columns **to** **character** **type** **and** replacing "n/a" **with** NA

**data** <- **data** %>%

mutate(scores\_teaching = **as**.**character**(scores\_teaching),

scores\_research = **as**.**character**(scores\_research),

scores\_citations = **as**.**character**(scores\_citations),

scores\_industry\_income = **as**.**character**(scores\_industry\_income),

scores\_international\_outlook = **as**.**character**(scores\_international\_outlook),

overall\_score = **as**.**character**(overall\_score)) %>%

mutate(across(c(scores\_teaching, scores\_research, scores\_citations,

scores\_industry\_income, scores\_international\_outlook,overall\_score),

~na\_if(., "n/a")))

# Checking **for** missing **values** **in** 'scores\_teaching'

missing <- **is**.na(data$scores\_teaching)

missing\_sum <- **sum**(missing)

missingness\_percentage <- (missing\_sum / nrow(**data**)) \* 100

print(paste("Missingness in 'scores\_teaching':", missingness\_percentage, "%"))

# Identifying **and** removing constant columns

constant\_columns <- sapply(**data**, **function**(x) **length**(**unique**(x)) == 1)

print("Constant columns:")

print(names(**data**)[constant\_columns])

**data** <- **data** %>% **select**(-one\_of(names(**data**)[constant\_columns]))

# Visualizing missing **data** **and** testing **for** MCAR

library(naniar)

vis\_miss(**data**) +

labs(

title = "Before Data Cleaning" # **Add** a heading

)

# Visualize missing **data** percentages

plot\_missing(

**data**,

title = "Percentage of Missing Data", # **Add** a meaningful title

ggtheme = theme\_minimal() # Use a clean theme

)

# Statistics **of** missing **data**

mcar\_test(**data**)

# Removing **rows** **with** NA **in** 'scores\_teaching'

**data** <- **data** %>% **filter**(!**is**.na(scores\_teaching))

# Handling 'overall\_score' ranges **by** calculating averages

calculate\_average <- **function**(x) {

**if** (str\_detect(x, "–")) {

nums <- **as**.**numeric**(str\_split(x, "–", simplify = **TRUE**))

**return**(mean(nums))

} **else** {

**return**(**as**.**numeric**(x))

}

}

**data** <- **data** %>%

mutate(overall\_score = sapply(overall\_score, calculate\_average))

# Removing duplicates **and** saving a backup

dataset\_backup <- **data**

**data** <- **data** %>% **distinct**()

# Converting **numeric**-**like** columns **and** percentages

data$stats\_number\_students <- **as**.**numeric**(gsub(",", "", data$stats\_number\_students))

data$stats\_pc\_intl\_students <- **as**.**numeric**(sub("%", "", data$stats\_pc\_intl\_students)) / 100

# Splitting **and** handling 'stats\_female\_male\_ratio'

**data** <- **data** %>%

mutate(

female\_ratio = **as**.**numeric**(gsub(":.\*", "", stats\_female\_male\_ratio)),

male\_ratio = **as**.**numeric**(gsub(".\*:", "", stats\_female\_male\_ratio))

) %>%

**select**(-stats\_female\_male\_ratio)

# Filling NA **in** gender ratios **with** mean **values**

data$female\_ratio[**is**.na(data$female\_ratio)] <- mean(data$female\_ratio, na.rm = **TRUE**)

data$male\_ratio[**is**.na(data$male\_ratio)] <- mean(data$male\_ratio, na.rm = **TRUE**)

# **Final** visualization **of** missing **data**

vis\_miss(**data**) +

labs(

title = "After Data Cleaning" # **Add** a heading

)

# ---------------------------------------------------------------------------------

# 4. Exploratory **Data** Analysis (EDA)

# ---------------------------------------------------------------------------------

# Converting score columns **to** **numeric**

**data** <- **data** %>%

mutate(across(

c(scores\_teaching, scores\_research, scores\_citations, scores\_international\_outlook, overall\_score),

**as**.**numeric**

))

# Removing **rows** **with** NA **in** critical **numeric** columns

**data** <- na.omit(**data**)

# Compute correlation matrix

corr\_matrix <- cor(**data**[, sapply(**data**, **is**.**numeric**)], use = "complete.obs") # Use complete observations **only**

# Enhanced correlation plot

corrplot(

corr\_matrix,

**method** = "shade", # Use shaded squares **for** a polished look

shade.col = NA, # Disable additional shading color

tl.col = "black", # **Set** **text** label color **to** black

tl.srt = 45, # Rotate **text** labels **for** better readability

col = colorRampPalette(c("red", "white", "blue"))(200), # Red-white-blue gradient

addCoef.col = "black", # **Add** correlation coefficients **in** black

**number**.cex = 0.7, # Adjust **size** **of** correlation coefficients

cl.pos = "b", # Place the color legend **at** the bottom

title = "Correlation Matrix", # **Add** a title

mar = c(0, 0, 2, 0) # Adjust margins **to** fit the title

)

# Scatter plots **of** features against 'overall\_score'

feature\_names <- c('scores\_teaching', 'scores\_research', 'scores\_citations', 'scores\_international\_outlook')

**for** (feature **in** feature\_names) {

ggplot(**data**, aes(x = !!sym(feature), y = overall\_score)) +

geom\_point() +

ggtitle(paste("Scatter Plot of", feature, "vs Overall Score"))

}

# Saving scatter plots **to** a PDF

pdf("scatter\_plots.pdf")

**for** (feature **in** feature\_names) {

p <- ggplot(**data**, aes\_string(x = feature, y = "overall\_score")) +

geom\_point() +

ggtitle(paste("Scatter Plot of", feature, "vs Overall Score")) +

xlab(feature) +

ylab("Overall Score")

print(p)

}

dev.off()

# Define the feature names **and** their **corresponding** colors

feature\_names <- c('scores\_teaching', 'scores\_research', 'scores\_citations', 'scores\_international\_outlook')

plot\_colors <- c("darkgreen", "purple", "orange", "blue") # **Unique** colors **for** **each** plot

# Generate individual scatter plots **with** **unique** colors

**for** (i **in** seq\_along(feature\_names)) {

feature <- feature\_names[i]

color <- plot\_colors[i]

# Store the ggplot **object**

p <- ggplot(**data**, aes\_string(x = feature, y = "overall\_score")) +

geom\_point(color = color, alpha = 0.7, **size** = 2) + # **Add** transparency **and** **size**

ggtitle(paste("Scatter Plot of", feature, "vs Overall Score")) +

xlab(feature) +

ylab("Overall Score") +

theme\_minimal() + # Clean theme **for** better visibility

theme(

plot.title = element\_text(hjust = 0.5, **size** = 14, face = "bold"),

axis.title = element\_text(**size** = 12),

axis.**text** = element\_text(**size** = 10)

)

# Print the plot

print(p)

}

#---------------------------

keep\_numerical <- **function**(**data**) {

library(dplyr)

**data** %>%

select\_if(**is**.**numeric**)

}

new\_data <- keep\_numerical(**data**)

print(new\_data)

chart.Correlation(new\_data, histogram = **TRUE**)

# ---------------------------------------------------------------------------------

# 5. Model Building **and** Evaluation

# ---------------------------------------------------------------------------------

# Splitting **data** **into** training **and** testing **sets**

**set**.seed(42) # **for** reproducibility

trainIndex <- createDataPartition(new\_data$overall\_score, p = .8,

list = **FALSE**,

times = 1)

data\_train <- new\_data[trainIndex, ]

data\_test <- new\_data[-trainIndex, ]

# Defining **and** evaluating models

# Defining **and** evaluating models

model\_list <- list(

linear\_reg = lm(overall\_score ~ ., **data** = data\_train),

decision\_tree = rpart(overall\_score ~ ., **data** = data\_train, **method** = "anova"),

random\_forest = randomForest(overall\_score ~ ., **data** = data\_train)

)

results <- list()

**for** (model\_name **in** names(model\_list)) {

model <- model\_list[[model\_name]]

predictions <- predict(model, newdata = data\_test)

# Important Evaluation Metrics

mae <- mean(**abs**(predictions - data\_test$overall\_score))

rmse <- **sqrt**(mean((predictions - data\_test$overall\_score)^2))

r2 <- cor(predictions, data\_test$overall\_score)^2

mape <- mean(**abs**((predictions - data\_test$overall\_score) / data\_test$overall\_score)) \* 100

# Store results

results[[model\_name]] <- list(

MAE = mae,

RMSE = rmse,

R2 = r2,

MAPE = mape

)

}

# Print results

print(results)

# Random Forest Feature Importance

# Random Forest Feature Importance **with** Colorful Bars

importance\_df <- **as**.**data**.frame(randomForest(overall\_score ~ ., **data** = data\_train)$importance)

ggplot(importance\_df, aes(x = reorder(**row**.names(importance\_df), IncNodePurity), y = IncNodePurity, fill = IncNodePurity)) +

geom\_bar(stat = "identity") +

coord\_flip() +

scale\_fill\_gradient(low = "blue", high = "red") + # Gradient **from** blue **to** red

xlab("Features") +

ylab("Importance") +

ggtitle("Feature Importances in Random Forest Model") +

theme\_minimal() + # Use a clean theme

theme(

plot.title = element\_text(hjust = 0.5, **size** = 16, face = "bold"), # Centered title

axis.title = element\_text(**size** = 12),

axis.**text** = element\_text(**size** = 10)

)

# Plotting the **result** comparison

# Combine results **into** a **data** frame **for** visualization

results\_df <- do.**call**(rbind, lapply(results, **as**.**data**.frame))

results\_df$model <- rownames(results\_df)

rownames(results\_df) <- **NULL**

# Reshape the **data** **for** ggplot

results\_long <- pivot\_longer(

results\_df,

cols = -model, # Keep the "model" **column** intact

names\_to = "Metric",

values\_to = "Value"

)

# Plot the comparison **of** metrics across models

ggplot(results\_long, aes(x = Metric, y = Value, fill = model)) +

geom\_bar(stat = "identity", **position** = "dodge") +

labs(

title = "Model Performance Comparison",

x = "Metrics",

y = "Values",

fill = "Model"

) +

theme\_minimal() +

theme(

plot.title = element\_text(hjust = 0.5, **size** = 16, face = "bold"),

axis.title = element\_text(**size** = 12),

axis.**text** = element\_text(**size** = 10),

legend.title = element\_text(**size** = 12),

legend.**text** = element\_text(**size** = 10)

)

## Additional Visualization

Advanced correlation plot for better understanding

A graph of a graph

Description automatically generated with medium confidence