Forecasting Global University Rankings: A Data-Driven Predictive Model for Evaluating Future Performance

Final Report Submission

Course Title & Code: Data Science (CS 8339)

Group Members:

1. **Engr. Priha Bhatti (FA23-PHCS-0001) GL**
2. **Erum Mumtaz (SP22-PHCS-0003)**
3. **Mirza Shaharyar Ali Baig (FA22-MSCS-0037)**

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

Forecasting Global University Rankings: A Data-Driven Predictive Model for Evaluating Future Performance

Abstract:

In this research, we explore the significant impact that worldwide rankings have on higher education institutions (HEIs) as well as the complex difficulties associated with improving their ranks. Ranking systems have their detractors, but HEIs keep an eye out for improvements. We present a comprehensive methodology for ranking data prediction that combines advanced regression-based algorithms—such as Linear Regression—with exploratory data analysis (EDA). We present a novel approach that modifies probability estimations to evaluate choice path confidence levels. Surprisingly, we identify gradient-boosting regression as the best model, guaranteeing accuracy in rankings. This approach provides HEIs with the resources necessary to create a roadmap and

Introduction

Choosing which university to believe might be challenging because there is so much information out there. The World University Rankings compiles information on university rankings from a variety of sources, each with its own distinct hierarchy. In addition to these rating methods, there is information available about the universities themselves, their respective countries, success rates, and the cost of education.

University rankings are determined by a convoluted, contentious, and politically charged procedure. There are numerous national and international ranking systems, and their outcomes are frequently contradictory. Three unique global university rankings are included in this dataset, and they are very different from one another.

During the turn of the century, the idea of elite universities surfaced. These universities are known for their outstanding alumni, innovative research, and capacity to spread knowledge across international borders and technologies. Higher education institutions compete and cooperate to attract the most talented students, best scholars, and financial assistance from around the world.[1].

In this research, we explore the significant impact that worldwide rankings have on higher education institutions (HEIs) as well as the complex difficulties associated with improving their ranks. Ranking systems have their detractors, but HEIs keep an eye out for improvements. The main problem is that the ranking systems are so complex, with each one using a different set of

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

measurements and characteristics, that it is difficult for institutions to base their conclusions only on traditional exploratory data analysis (EDA).

To solve this problem, we have developed a comprehensive technique that combines EDA with advanced regression-based algorithms, like the Gradient Boosting Regressor and Random Forest Regressor, to predict ranking data precisely. The main focus of the research problem is the need for a practical and efficient strategy that addresses the drawbacks of current approaches and gives HEIs the ability to strategically plan and quantitatively measure their progress in the fiercely competitive world of higher education rankings. This strategy should combine complex modeling techniques with interpretable decision paths.

This study aims to offer a creative and workable answer to the difficulties faced by HEIs in their quest for higher positions, while still acknowledging the importance of worldwide rankings. We seek to create a solid foundation for strategic prediction and ongoing development by presenting a novel approach that modifies probability estimates to evaluate choice route confidence levels.

Scope of the Study

This research aims to give a detailed analysis of how global rankings affect higher education institutions (HEIs) and to offer a workable approach for raising HEIs' ranks. This includes the difficulties brought on by the complexity of ranking systems, the requirement for advanced modeling methods, and the necessity for useful information to direct strategic decision-making. The paper offers workable alternatives that make use of data science techniques, expanding its scope beyond the critiques of ranking systems.

This project has many objectives, with the main one being to offer a thorough and workable answer to the problems higher education institutions (HEIs) confront in worldwide rankings. The project aims to:

Recognize the Effects of World Rankings:

- Examine and evaluate the significant impact that global rankings have on the higher education scene.
- Analyze why, despite the criticism that has been leveled at them, HEIs are keeping a close eye on and actively working to improve their rankings.

Identify and Address Ranking System Challenges:

- Examine the difficulties and intricacies of the current ranking systems while taking a variety of measures and characteristics into account.
- Gain an understanding of the drawbacks and objections to conventional exploratory data analysis (EDA) techniques to offer solutions that can be put into practice.

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

Propose a Comprehensive Methodology:

- Introduce a novel methodology that combines advanced regression-based algorithms, including the Random Forest Regressor and Gradient Boosting Regressor, with interpretable decision tree models.
- Provide a practical and data-driven approach to predict ranking changes accurately and understand the decision paths influencing these changes.

Adjust Probability Estimates for Confidence Levels:

- Innovate by adjusting probability estimates, offering a mechanism to evaluate the confidence levels associated with decision paths.
- Enhance the interpretability of the models and decision-making process.

Identify Optimal Predictive Models:

- Determine the best-fit predictive model for ranking data, highlighting Gradient Boosting Regression as the optimal choice for precision in predictions.
- Establish a basis for selecting effective algorithms in similar contexts.

Empower HEIs for Strategic Decision-Making:

- Provide higher education institutions with actionable insights derived from data science methodologies.
- Enable the development of a strategic roadmap and a long-term action plan for continuous improvement in rankings.

Quantitatively Measure Progress:

- Introduce metrics and mechanisms to quantitatively measure progress over time.
- Equip HEIs with tools for monitoring and evaluating the effectiveness of implemented strategies.

Enhance Communication and Documentation:

- Document the entire data science process, ensuring transparency and reproducibility.
- Communicate findings and recommendations in a clear and interpretable manner to facilitate informed decision-making by stakeholders in the higher education sector.

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

Establish a Robust Framework for Prediction and Improvement:

- Create a holistic and robust framework that goes beyond traditional analyses, offering institutions a systematic approach to navigate the competitive landscape of higher education rankings.

Through these aims, the project endeavors to contribute valuable insights and methodologies that empower HEIs to strategically navigate and improve their global rankings, fostering continuous enhancement in the quality and competitiveness of higher education institutions worldwide.

This data science project includes a number of activities targeted at offering a complete technique for higher education institutes (HEIs) to anticipate and improve their global rankings. The first part is methodically collecting and integrating different and relevant data on higher education rankings. This information is gathered from a variety of sources, and every effort is taken to assure data consistency and completeness.

Following data gathering, the next phase is data cleaning and preprocessing. This includes dealing with missing numbers, outliers, and inconsistencies in order to assure the dataset's quality and reliability. Standardization and feature preparation are carried out to ensure data consistency throughout the investigation.

To detect broad trends and patterns within the ranking data, the exploratory data analysis (EDA) phase employs techniques such as correlation analysis and visualization tools. The goal is to uncover crucial qualities and correlations that could have a substantial impact on rankings. Feature engineering is then used to generate additional features or transformations that improve the model's prediction capability, while also investigating domain-specific characteristics to provide a full knowledge of ranking changes.

Regression-based algorithms, notably the Random Forest Regressor and Gradient Boosting Regressor, are reviewed and chosen during the algorithm selection step based on their usefulness for predicting ranking data. Model training and validation follow, with historical ranking data incorporated and temporal trends and changes taken into account. Cross-validation procedures are used to guarantee that the selected models work well.

Interpretable decision trees are created to provide clear decision paths that aid in comprehending the elements that influence ranking changes. To find key contributors to the prediction model, a feature importance analysis is performed. The novel step of modifying probability estimates is used to evaluate confidence levels associated with choice paths, hence improving the model's and decision-making process's interpretability.

For regression tasks, model evaluation entails analyzing performance using applicable measures such as Mean Squared Error (MSE) or R-squared. To assess the robustness of the models, sensitivity studies are performed. Following that, the project enters the strategic roadmap

development phase, in which model results are turned into concrete initiatives for HEIs to enhance their rankings. A comprehensive strategy as well as a long-term action plan are developed.

The project creates measures for objectively monitoring progress over time to promote continual improvement. Institutions are given instruments to track and assess the success of implemented policies. Finally, the entire data science process is meticulously documented, including data sources, preprocessing methods, modeling details, and evaluation outcomes. Findings and recommendations are conveyed to stakeholders in the higher education sector in a straightforward and understandable manner. The project intends to provide a robust and actionable technique for anticipating ranking changes and promoting continual improvement in higher education rankings through these systematic data science tasks.

Data Science Tasks to be Followed in Project

Data science tasks in the context of predicting and improving higher education rankings involve a series of systematic steps to leverage data for actionable insights. Here is an organized list of these tasks:

1. **Data Collection and Integration:**
    - Gather diverse and relevant data on higher education rankings, encompassing various metrics and parameters.
    - Integrate datasets from multiple sources to create a comprehensive dataset, ensuring data consistency and completeness.
2. **Data Cleaning and Preprocessing:**
    - Handle missing values, outliers, and inconsistencies in the collected data to ensure data quality.
    - Standardize and preprocess features to maintain data consistency and prepare it for analysis.

3. **Exploratory Data Analysis (EDA):**
    - Utilize EDA techniques such as correlation analysis, visualizations, and statistical summaries to understand trends and patterns in ranking data.
    - Identify key features and relationships that may influence rankings.
4. **Feature Engineering:**
    - Derive new features or transformations to enhance the predictive power of the model.
    - Explore domain-specific features that may contribute to a better understanding of ranking changes.
5. **Algorithm Selection:**
    - Evaluate and choose appropriate regression-based algorithms, such as the Random Forest Regressor and Gradient Boosting Regressor, based on their suitability for predicting ranking data.
6. **Model Training and Validation:**
    - Train the selected models using historical ranking data, considering temporal trends and variations.
    - Validate the models using cross-validation techniques to ensure robust performance.
7. **Interpretable Decision Trees:**
    - Develop decision tree models to provide interpretable decision paths for understanding the factors influencing ranking changes.
    - Analyze feature importance to identify critical contributors to the predictive model.
8. **Adjusting Probability Estimates:**
    - Implement a strategy to adjust probability estimates, allowing for the evaluation of confidence levels associated with decision paths.
    - Enhance the interpretability of the models and decision-making process.
9. **Model Evaluation:**
    - Assess the performance of the models using relevant evaluation metrics, such as Mean Squared Error (MSE) or R-squared for regression tasks.

- Conduct sensitivity analyses to understand the robustness of the models.

10. Strategic Roadmap Development:

- Translate model insights into actionable strategies for higher education institutes (HEIs) to improve their rankings.
- Create a comprehensive roadmap and a long-term action plan based on data-driven findings.

11. Quantitative Measurement of Progress:

- Establish metrics for quantitatively measuring progress over time.
- Provide institutions with tools to monitor and evaluate the effectiveness of implemented strategies.

12. Documentation and Communication:

- Document the entire data science process, including data sources, preprocessing steps, modeling details, and evaluation results.
- Communicate findings and recommendations in a clear and interpretable manner for stakeholders in the higher education sector.

By systematically performing these data science tasks, the goal is to provide a rigorous methodology for predicting ranking changes accurately, offering actionable insights for strategic decision-making, and fostering continuous improvement in higher education rankings.

About Dataset

The World University Rankings 2023 dataset includes 1,799 universities across 104 countries and regions, making them the largest and most diverse university rankings to date. The table is based on 13 carefully calibrated performance indicators that measure an institution's performance across four areas: teaching, research, knowledge transfer, and international outlook. This year's ranking analyzed over 121 million citations across more than 15.5 million research publications and included survey responses from 40,000 scholars globally. Overall, we collected over 680,000 data points from more than 2,500 institutions that submitted data.

*Features *

This dataset includes the following 13 features:

1. University Rank
2. Name of University

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

3. Location
4. No of student
5. No of student per staff
6. International Student
7. Female: Male Ratio
8. Overall Score
9. Teaching Score
10. Research Score
11. Citations Score
12. Industry Income Score
13. International Outlook Score

Description

This cleaned version of the dataset has undergone rigorous preprocessing, including handling missing values and encoding categorical features, resulting in a dataset with enhanced usability and cleanliness. It now consists of 2,341 rows and 2,361 columns, providing valuable insights for data analysis, machine learning, and research in the field of higher education.

Original vs. Cleaned Version Comparison

Original Version

The original version of the "World University Rankings 2023" dataset was a comprehensive collection of data on 1,799 universities across 104 countries and regions. While it provided valuable insights into higher education worldwide, it presented some challenges due to missing values, inconsistencies, and a mix of data types.

Original Dataset Source:

World University Rankings 2023

Cleaned Version

In this cleaned version of the dataset, significant efforts have been made to enhance its quality and usability. The following improvements were made:

Handling Missing Values:

- All missing values, including NaN and Null values, have been meticulously addressed for every feature in the dataset.
- Specifically, missing values in the "Name of University" and "Location" columns have been replaced with meaningful placeholders: "Unknown University" and "Unknown Location," respectively.

Encoding and Transformation:

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

- One-hot encoding has been applied to the "Name of University" and "Location" columns, converting categorical data into a numerical format suitable for analysis and modeling.
- The "Female Ratio" and "Male Ratio" columns have been separated, allowing for more straightforward analysis of gender ratios.
- "OverAll Score" has been divided into "OverAll Score Min" and "OverAll Score Max" columns, providing insights into the range of scores.
- "International Student" values have been encoded as fractional values, making it easier to interpret and analyze.
- Several features, including "Female Ratio," "Male Ratio," "OverAll Score Min," "OverAll Score Max," "No of Student," and "International Student," have been encoded as numerical values, improving their compatibility with data analysis and modeling techniques.

These enhancements have transformed the dataset into a cleaned and well-structured resource for data analysis, machine learning, and research in the field of higher education. Researchers and data enthusiasts can now explore and gain valuable insights from this improved dataset with confidence.

Whether you are conducting exploratory data analysis, building predictive models, or conducting research, this cleaned version of the dataset provides a solid foundation for your analytical endeavors.

Project Structure/Layout

This project employs a data-driven and interdisciplinary approach, utilizing Python.

Approaches:

1. **Data-Driven Approach:**
    - Utilize data to drive decision-making and strategy formulation.
    - Leverage historical ranking data and diverse metrics to understand patterns and trends.
2. **Predictive Modeling Approach:**
    - Employ regression-based algorithms, including Random Forest Regressor and Gradient Boosting Regressor, for accurate prediction of ranking changes.
    - Use interpretable decision trees to understand the factors influencing rankings.
3. **Interdisciplinary Approach:**
    - Combine expertise from data science, education, and domain-specific knowledge to develop a holistic understanding of ranking dynamics.
    - Facilitate collaboration between data scientists and education experts for a comprehensive solution.

Tools:

1. **Python Programming Language:**
    - Utilize Python for its extensive libraries and frameworks in data science, including scikit-learn, pandas, and NumPy.
2. **Scikit-Learn:**
    - Employ the scikit-learn library for implementing machine learning algorithms, model training, and evaluation.
3. **Jupiter Notebooks:**
    - Use Jupiter Notebooks for interactive and collaborative development, allowing for documentation of code, analyses, and visualizations.
4. **Matplotlib and Seaborn:**
    - Create visualizations using Matplotlib and Seaborn for effective communication of insights.
5. **Graphviz:**
    - Visualize decision trees using Graphviz to enhance interpretability.
6. **GitHub:**
    - Utilize version control through platforms like GitHub for collaborative development and code management.

Techniques:

1. **Regression-Based Modeling:**
    - Apply Random Forest Regressor and Gradient Boosting Regressor for predicting continuous ranking changes.
2. **Exploratory Data Analysis (EDA):**
    - Employ correlation analysis, box plots, and other EDA techniques to understand data patterns and relationships.
3. **Feature Engineering:**
    - Derive new features or transformations to enhance the predictive power of the model.
4. **Cross-Validation:**
    - Validate models using cross-validation techniques to ensure robust performance and generalization.

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

5. **Adjusting Probability Estimates:**
    - Implement a strategy to adjust probability estimates, providing a measure of confidence in decision paths.
6. **Sensitivity Analysis:**
    - Conduct sensitivity analyses to understand the impact of variations and uncertainties in the models.
7. **Strategic Roadmap Development:**
    - Translate model insights into actionable strategies for HEIs to improve their rankings.
    - Develop a comprehensive roadmap and a long-term action plan based on data-driven findings.
8. **Quantitative Measurement of Progress:**
    - Establish metrics for quantitatively measuring progress over time, enabling the monitoring of strategy effectiveness.
9. **Documentation and Communication:**
    - Document the entire data science process using Jupiter Notebooks and communicate findings through visualizations and reports.
    - Utilize GitHub for collaborative documentation and version control.

By combining these approaches, tools, and techniques, the project aims to provide a robust and interdisciplinary solution for predicting and improving higher education rankings, fostering continuous improvement in the higher education landscape.

Proposed Methodology

In addressing our research query, we employed the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. This widely recognized data science standard, often employed by decision-makers in business settings, comprises six key phases: (1) Business understanding, (2) Data understanding, (3) Data Preparation, (4) Modelling, (5) Evaluation, and (6) Deployment. These phases were elucidated, implemented, and deliberated within the context of analyzing the World University Ranking dataset.

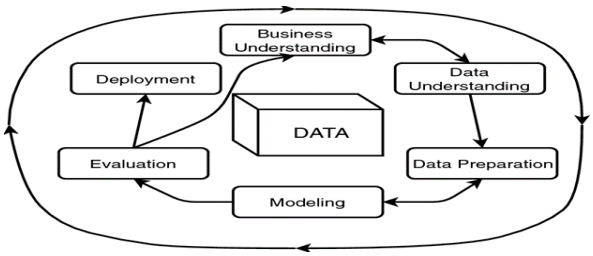


Figure 1. Layout of project workflow

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

Business Understanding

World University Rankings 2023 involves recognizing and comprehending the practical and strategic implications of these rankings for various stakeholders. This understanding can encompass several aspects:

Institutional Performance Benchmarking: Businesses and employers often consider university rankings to assess the quality of education and research facilities. Understanding the impact of these rankings on benchmarking the performance of educational institutions is crucial.

Talent Acquisition and Retention: For businesses, these rankings might influence their talent acquisition strategies. Understanding the rankings helps in attracting and retaining top graduates from highly-ranked universities.

Research and Innovation Collaboration: High-ranking universities often have a robust research environment. Businesses may seek collaborations or partnerships for innovation, R&D initiatives, and technology transfer.

Reputation and Partnerships: A university's ranking can significantly influence its reputation and global partnerships. Understanding how these rankings impact the formation of alliances, academic collaborations, and international partnerships is vital for business growth.

Educational Quality and Market Demand: The ranking may indicate the educational quality and the market demand for specific skills and knowledge. This understanding can guide businesses in aligning their hiring needs and skill development programs.

Investment and Funding Opportunities: Higher-ranked universities might attract more investment opportunities for infrastructure, research, and development. Understanding this aspect can provide insights into potential investment avenues for businesses.

Impact on Economic Development: Recognizing how the rankings influence local and national economies by attracting talent, and investments, and fostering innovation is also a key business consideration.

Understanding these aspects allows businesses to align their strategies and operations, tapping into the opportunities presented by highly-ranked universities and adapting to the changing landscape of education and research.

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

Business Questions

a) What factors contribute most to a university's high ranking?

b) How do universities from different regions perform in the rankings?

c) Is there a correlation between specific indicators (teaching quality, research output, etc.) and the overall reputation score of a university?

d) Are there specific countries that dominate the top rankings?

e) Analyze countries and institutions based on their scores (e.g., academic scores, employer reputation, faculty/student ratios)?

f) How do universities strategically improve their scores over time?

g) Explore relationships and correlations between different features and their impact on overall ranking?

h) How do universities leverage their rankings for international student recruitment and partnerships with other institutions?

After the data mining process, we followed a process of data cleaning. The data cleaning process for the World University Ranking 2023 dataset on Kaggle typically involves several steps to ensure the dataset is accurate, complete, and ready for analysis. Here's an outline of the general data-cleaning process:

Initiating the Project

1. Importing the Necessary Libraries

In this section of the code, we are importing several Python libraries for data analysis, visualization, and modeling. Let's break down the key components:

Data Analysis and Visualization Libraries:

- numpy and pandas are used for numerical operations and data manipulation.
- matplotlib and seaborn are employed for creating visualizations.
- plotly.graph_objects and plotly.express are utilized for interactive visualizations.
- missingno is used for visualizing missing data patterns.

Styling and Configuration:

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

- Stylistic choices are made using plt.style.use() and sns.set_theme() to configure the appearance of plots.
- warnings library is employed to suppress warning messages during code execution.

Label Encoding:

- LabelEncoder from sklearn.preprocessing is instantiated to encode categorical variables into numerical format.

Modeling Libraries:

- train_test_split from sklearn.model_selection is used to split the dataset into training and testing sets.
- LinearRegression, RandomForestRegressor, GradientBoostingRegressor, and DecisionTreeRegressor from sklearn.ensemble are imported for regression modeling.
- GridSearchCV is employed for hyperparameter tuning using grid search.

Evaluation Metrics:

- r2_score, mean_absolute_error, and mean_squared_error from sklearn. metrics are imported to evaluate the performance of regression models.

The code emphasizes a dark-themed visualization style and utilizes various libraries to handle data, create insightful visualizations, and implement regression models for analysis. Additionally, it configures settings for a visually appealing and informative representation of data.

2. Exploratory Data Analysis (EDA)

Steps for EDA Involves:

- **Reading Dataset:** Reads the World University Rankings 2023 dataset from the specified CSV file path on Kaggle.
- **Displaying Dataset:** Prints the first few rows of the dataset using the head() function.

After the data mining process, we followed a process of data cleaning. The data cleaning process for the World University Ranking 2023 dataset on Kaggle typically involves several steps to ensure the dataset is accurate, complete, and ready for analysis. Here's an outline of the general data-cleaning process:

- **Handling Missing Values:** Check for any missing or null values in the dataset. Depending on the significance of the missing data, you can either remove rows or columns with missing values or impute the missing data using techniques like mean, median, or mode replacement.
- **Data Validation:** Validate the integrity of the data. This includes checking for outliers, incorrect data types, and inconsistent data entries that might affect the analysis.
- **Removing Duplicates:** Identify and eliminate any duplicate entries to prevent redundancy in the dataset.
- **Standardizing Data:** Ensure uniformity by standardizing units, formats, and representations. This step involves converting data into a consistent format, such as ensuring date formats are the same, standardizing categorical variables, or converting data to a common scale if needed.
- **Addressing Inconsistencies:** Address any inconsistencies in the data. For example, ensuring that the values in the 'Female: Male Ratio' column are correctly formatted and represent the intended data.
- **Feature Selection:** Select the relevant features for analysis. Not all columns or attributes in the dataset might be required for analysis. Choose the most significant attributes that contribute to the analysis and remove irrelevant or redundant columns.
- **Data Transformation:** If necessary, perform data transformations such as normalization or scaling for numerical attributes.
- **Creating Derived Features:** Create new features or variables that might be useful for analysis based on existing data.
- **Data Integrity Checks:** Perform final checks on the dataset to ensure data integrity and correctness after the cleaning process.

The task involved preprocessing the dataset for World University Rankings to handle missing values, performing data encoding for categorical values, and scaling numeric attributes.

I have added an image below which shows the weights for the respective subject disciplines.

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

	Indicator	Overall	Arts and Humanities	Social Sciences	Business and Economics	Clinical and Health	Life Sciences	Physical Sciences	Engineering	Computer Science	Psychology	Law	Education
C1	**Citations**	30.00%	15.00%	25.00%	25.00%	35.00%	35.00%	35.00%	27.50%	27.50%	35.00%	25.00%	27.50%
E1	**Industry Income/Staff**	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	5.00%	5.00%	2.50%	2.50%	2.50%
T1	**Teaching Reputation**	15.00%	25.30%	21.10%	21.10%	17.90%	17.90%	17.90%	19.50%	19.50%	17.90%	21.00%	20.00%
T2	**Students to Staff Ratio**	4.50%	3.80%	3.30%	3.30%	2.80%	2.80%	2.80%	3.00%	3.00%	2.80%	4.50%	4.50%
T3	**PhD/Bachelors**	2.25%	1.80%	1.60%	0.00%	1.40%	1.40%	1.40%	1.50%	1.50%	1.40%	0.00%	0.00%
T4	**PhD/Staff**	6.00%	4.60%	4.80%	4.90%	4.00%	4.00%	4.00%	4.50%	4.50%	4.00%	4.90%	6.00%
T5	**Income/Staff**	2.25%	1.90%	1.60%	1.60%	1.40%	1.40%	1.40%	1.50%	1.50%	1.40%	2.30%	2.20%
R1	**Research Reputation**	18.00%	30.00%	22.80%	22.80%	19.30%	19.30%	19.30%	21.00%	21.00%	19.30%	21.00%	20.00%
R2	**Research Income/Staff**	6.00%	3.80%	4.90%	4.90%	4.10%	4.10%	4.10%	4.50%	4.50%	4.10%	4.90%	4.90%
R3	**Papers/Staff**	6.00%	3.80%	4.90%	4.90%	4.10%	4.10%	4.10%	4.50%	4.50%	4.10%	4.90%	4.90%
I1	**International Students**	2.50%	2.50%	2.50%	3.00%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	3.00%	2.50%
I2	**International Staff**	2.50%	2.50%	2.50%	3.00%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	3.00%	2.50%
I3	**International collaboration**	2.50%	2.50%	2.50%	3.00%	2.50%	2.50%	2.50%	2.50%	2.50%	2.50%	3.00%	2.50%
	Total	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Figure 2. Weights for the respective subject disciplines of World University Ranking 2023

We executed the provided steps to ensure the data was thoroughly cleansed and validated. Below are the specific actions undertaken:

Handling Missing Values:

- Empty strings were replaced with NaN values for columns with missing data. The count of missing values in each feature was checked.
- Missing values in the 'Name of University' and 'Location' columns were filled with "Unknown University" and "Unknown Location" labels, respectively.
- The mean values were used to replace missing data in the 'No of students per staff', 'Teaching Score', 'Research Score', 'Citations Score', 'Industry Income Score', and 'International Outlook Score' columns.

Data Encoding:

- The 'No of student' feature, initially containing numeric values with commas, was converted to a numeric format by removing commas.
- The 'International Student' feature, which included percentage values, was converted to a fraction after removing the '%' symbols. Empty strings were assigned NaN values.
- 'Female: Male Ratio' data was split into separate 'Female Ratio' and 'Male Ratio' columns and converted to floating-point numbers.
- The 'Overall Score' column contained both ranged and single numeric values. It was split into the 'Overall Score Min' and 'Overall Score Max' columns.
- One-hot encoding was applied to the 'Name of University' and 'Location' columns to convert categorical values.

Handling Missing Values of the Rest of the Features:

- The remaining missing values in columns were filled with the mean values of their respective features.
- Overall, these preprocessing steps were conducted to ensure the dataset was clean and suitable for subsequent analysis or modeling tasks, aiming to enhance the accuracy and reliability of the research involving World University Rankings.

3. Feature Engineering and Data Visualization

These feature engineering strategies served to refine the dataset, making it more suitable for modeling, analysis, and drawing accurate insights from the World University Ranking 2023 data. These methods played a pivotal role in ensuring the dataset's robustness and effectiveness for subsequent analysis. By utilizing these visualization techniques, stakeholders can gain comprehensive insights into the performance and trends within the World University Ranking 2023 dataset. These visualizations assist in identifying key patterns, making comparisons, and understanding the factors that contribute to the ranking positions of various universities.

The process of developing, manipulating, and selecting features that can increase the performance of machine learning models is known as feature engineering; yet, it can be difficult, time-consuming, and complex. This article will look at various ways of visualizing features in feature engineering to help you better visualize the features you're working with. Histograms and box plots are used to investigate feature distributions, scatter plots and pair plots to investigate feature relationships, heat maps, and correlation matrices to identify potential outliers, anomalies, and redundancies, and dimensionality reduction and clustering to analyze the impact of features on the target variable, and feature importance and partial dependence plots to determine which features

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

have a greater influence. Using these approaches, we can acquire significant insights into your data by using feature engineering.

Plots/Visualizations Categories:

Various charts can be useful in visualizing and comprehending data in a project predicting university rankings using machine learning models such as Random Forest Regressor, Gradient Boosting, and Linear Regression. Here's a quick rundown of the key types of plots that may be relevant to your project:

Scatter Plots:

Scatter plots depict the relationship between two continuous variables. Each point represents a university, and the x-y coordinates reflect the values of two attributes, allowing potential relationships to be observed.

Histograms:

Visualize numerical variable distributions with histograms.

Box Plots:

Identify outliers and assess data spread using box plots.

Correlation Analysis:

Explore relationships with correlation matrices and heatmaps.

Pair Plots:

Visualize relationships among multiple numerical variables.

Value Counts:

Understand the distribution of categorical variables using value_counts().

Bar Charts:

Create bar charts to visualize categorical variable distributions.

Heatmaps:

Visualize patterns and relationships in a matrix of numerical data.

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

➢ **Key Insights with visualization are observed through feature Engineering**

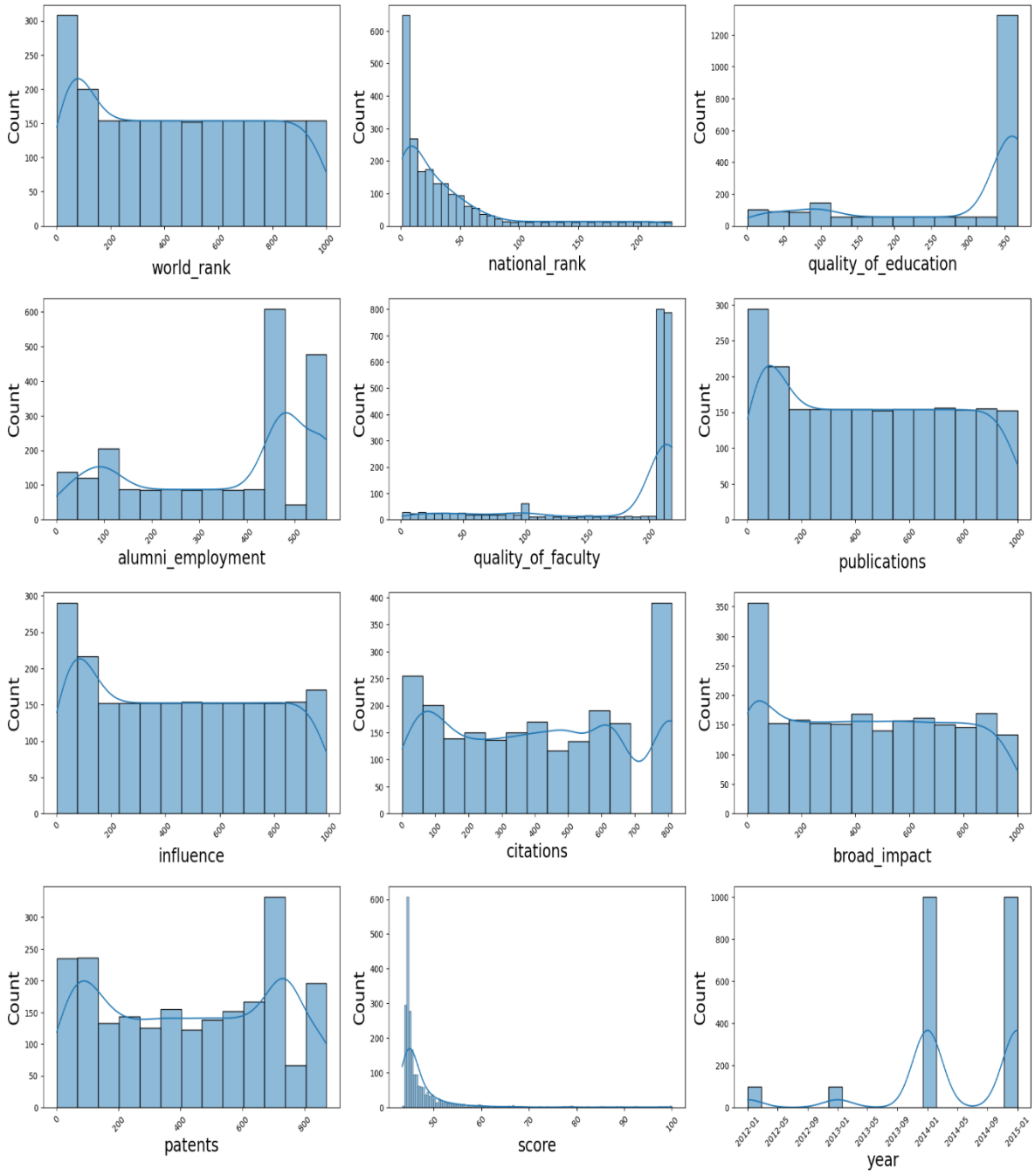
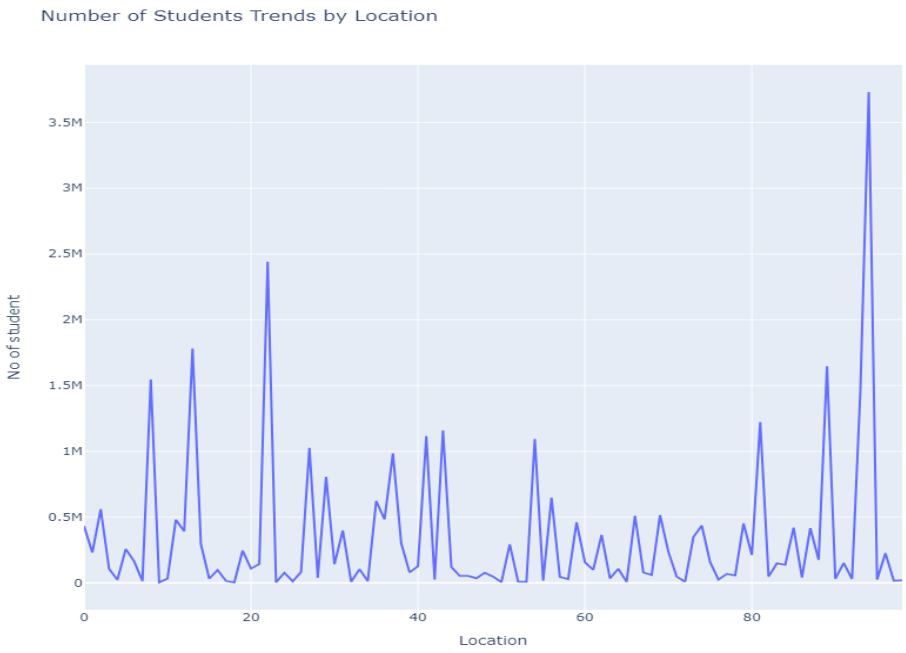
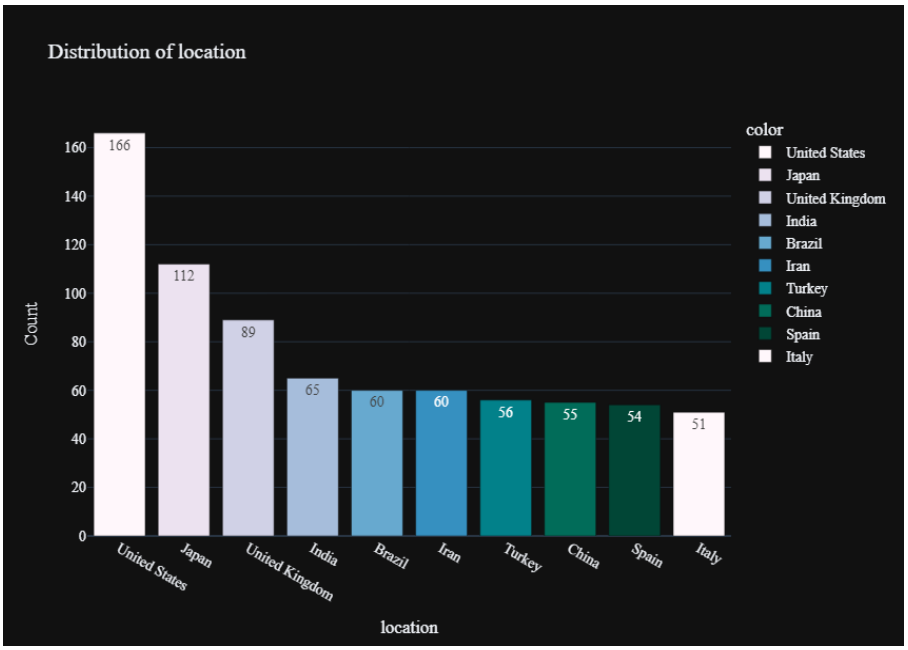


Figure 3: The histogram shows the value(frequency) distributed of the numeric features.

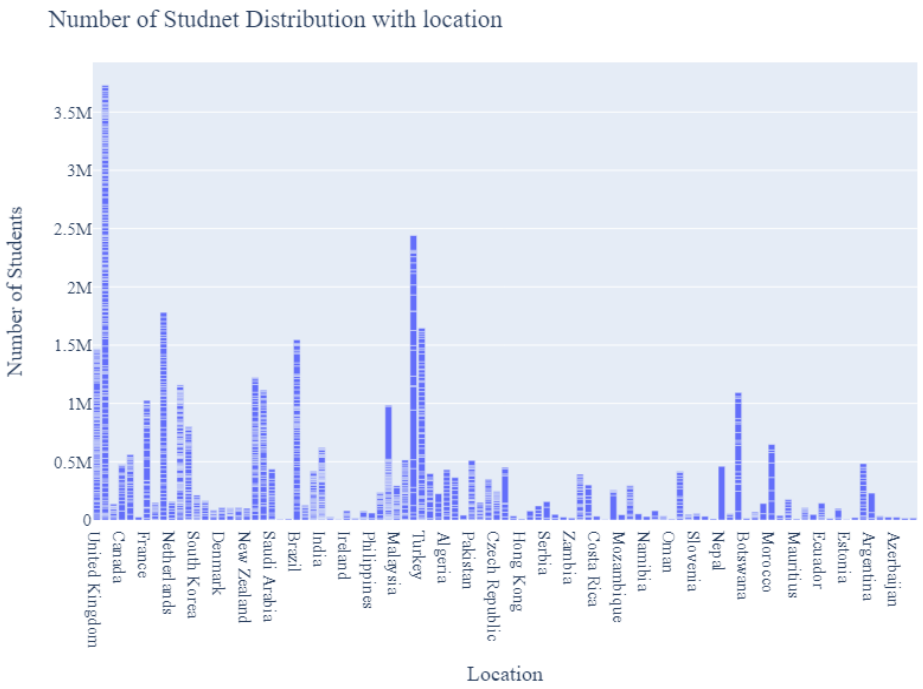
Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

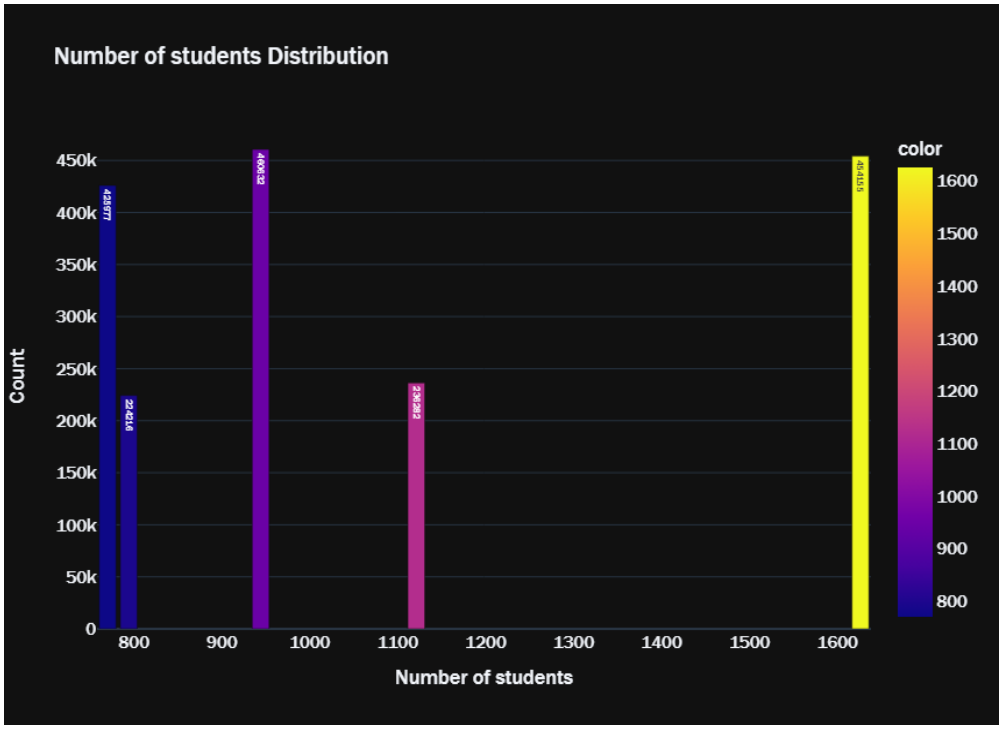
Insights using Visualization



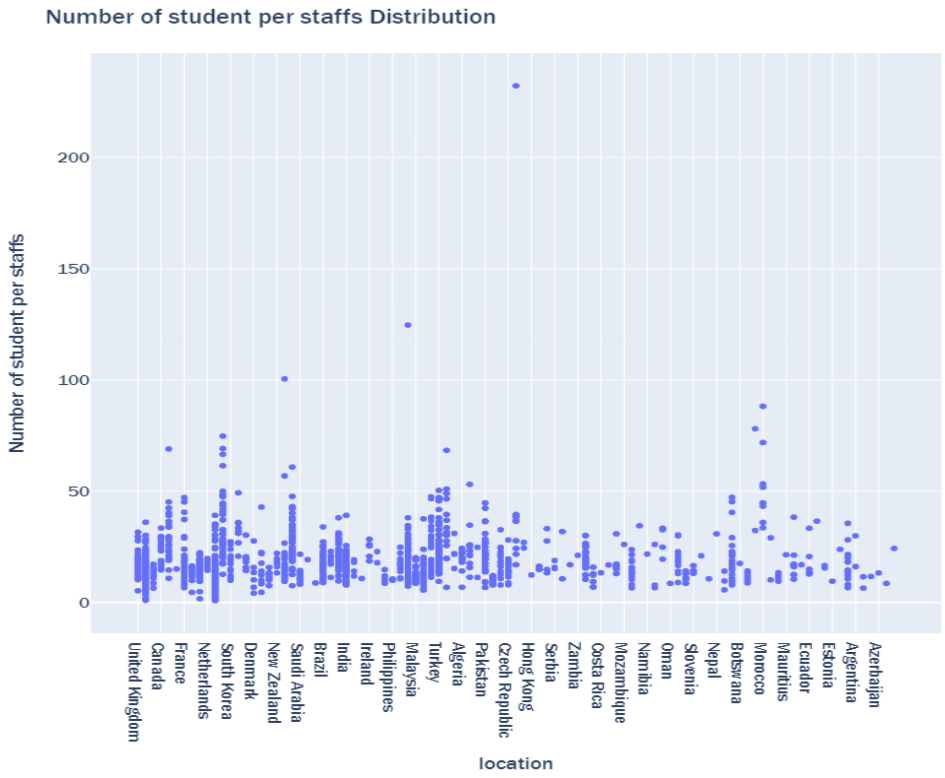


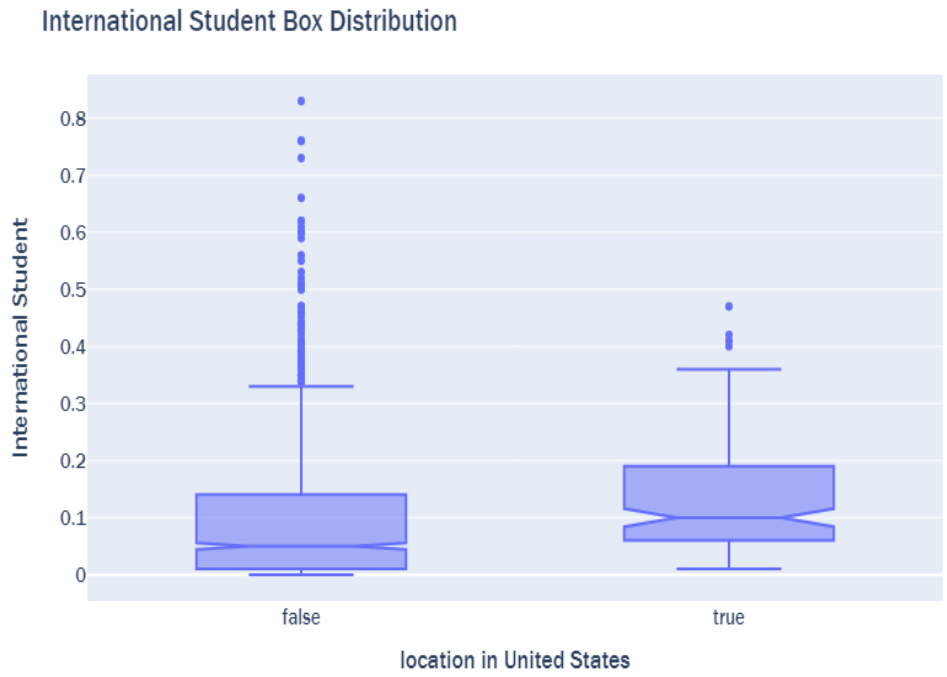
Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400



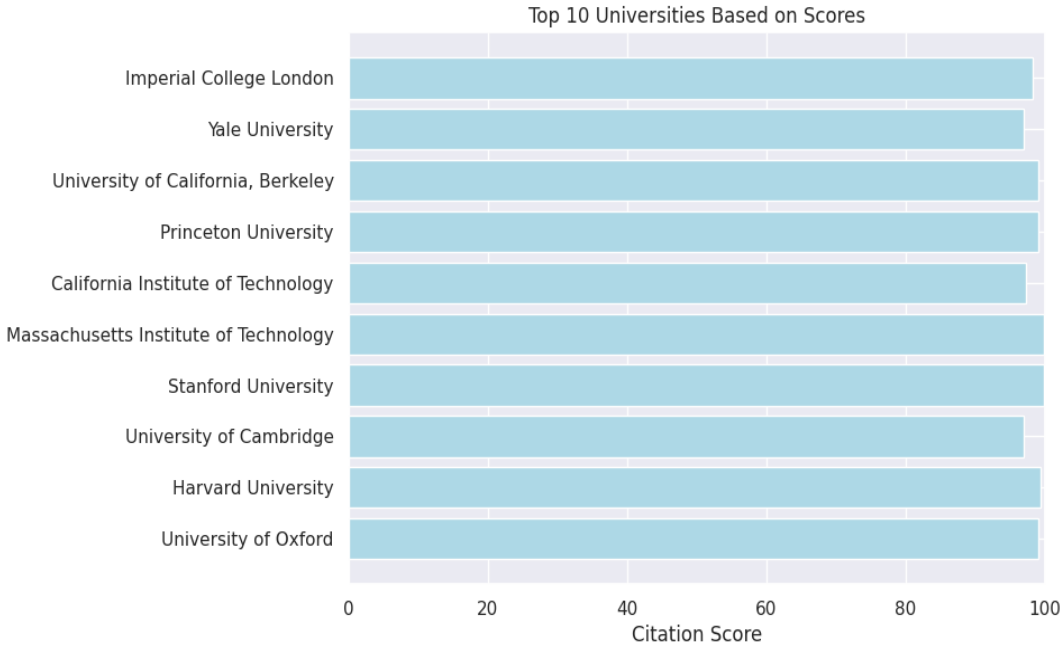


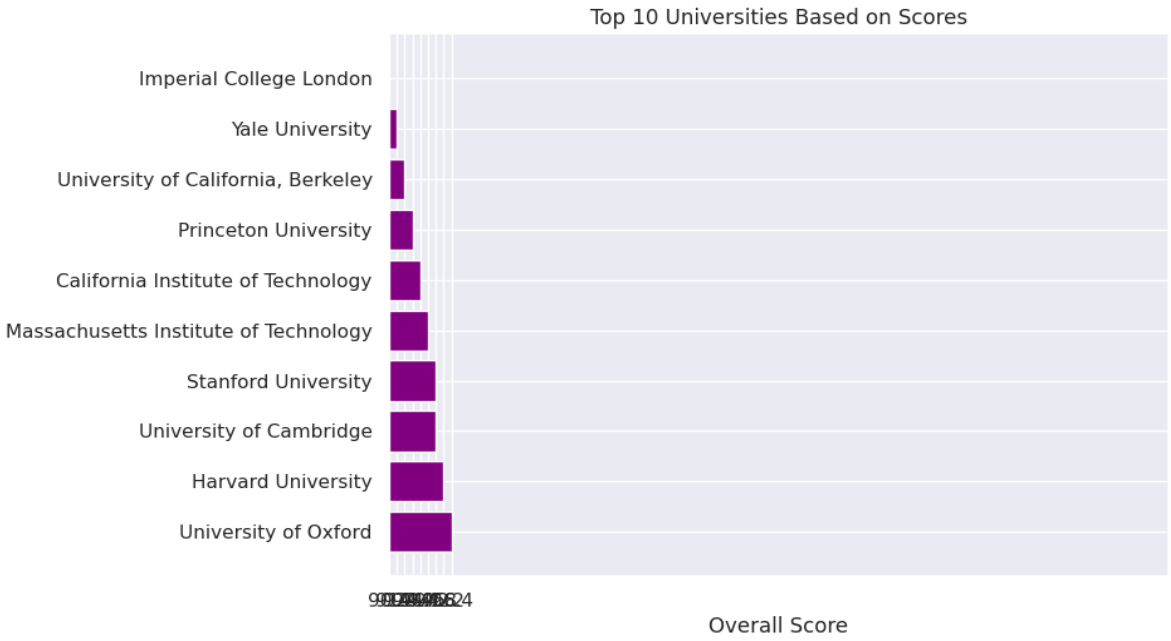
Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400



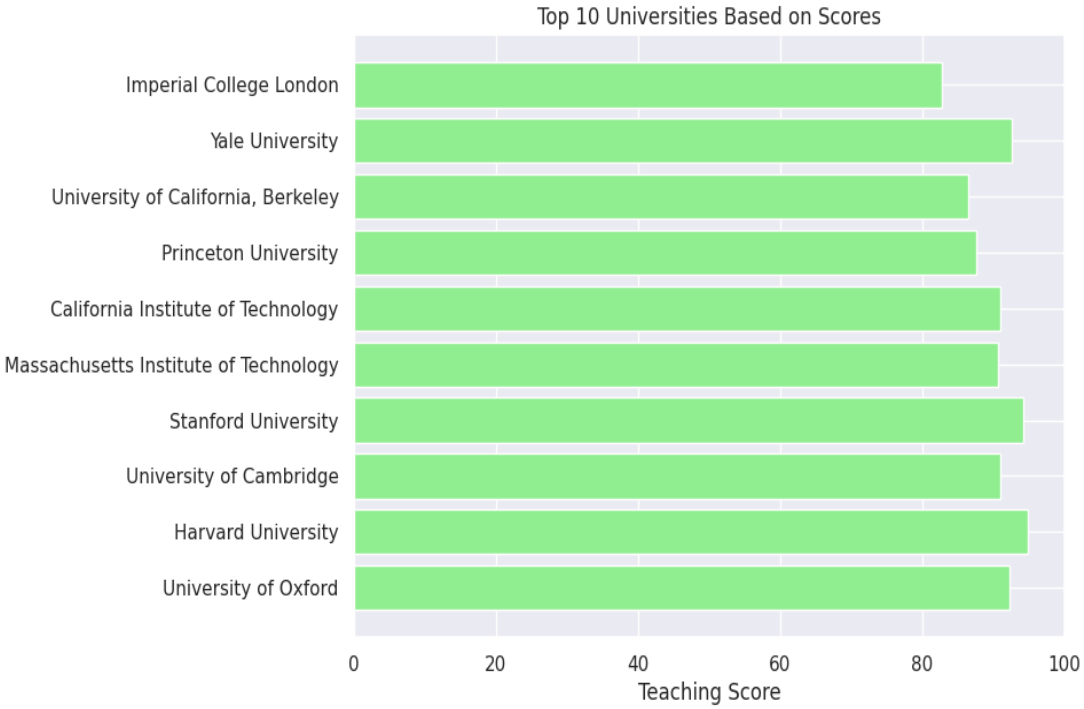


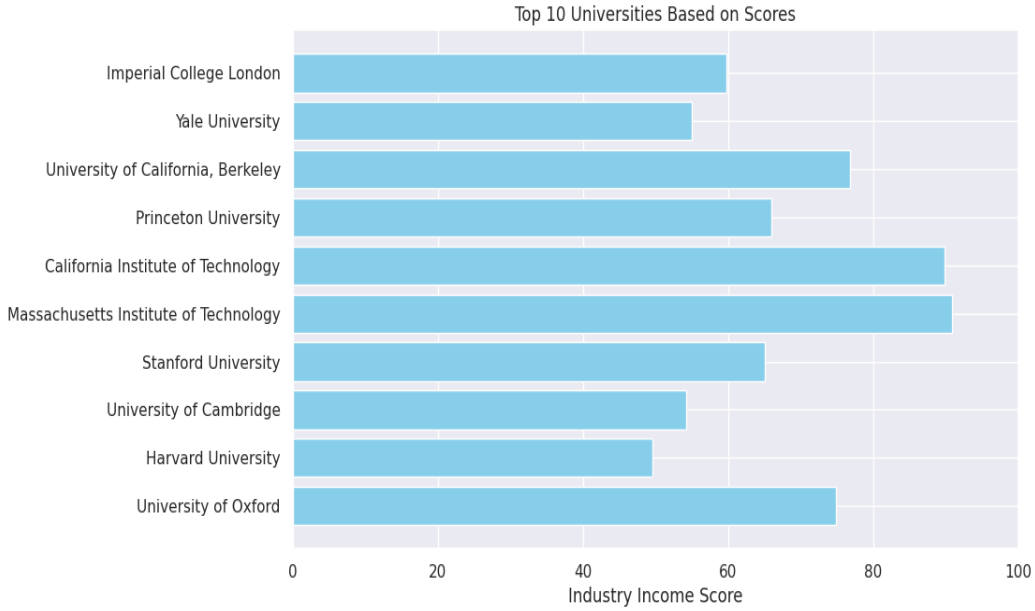
Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400



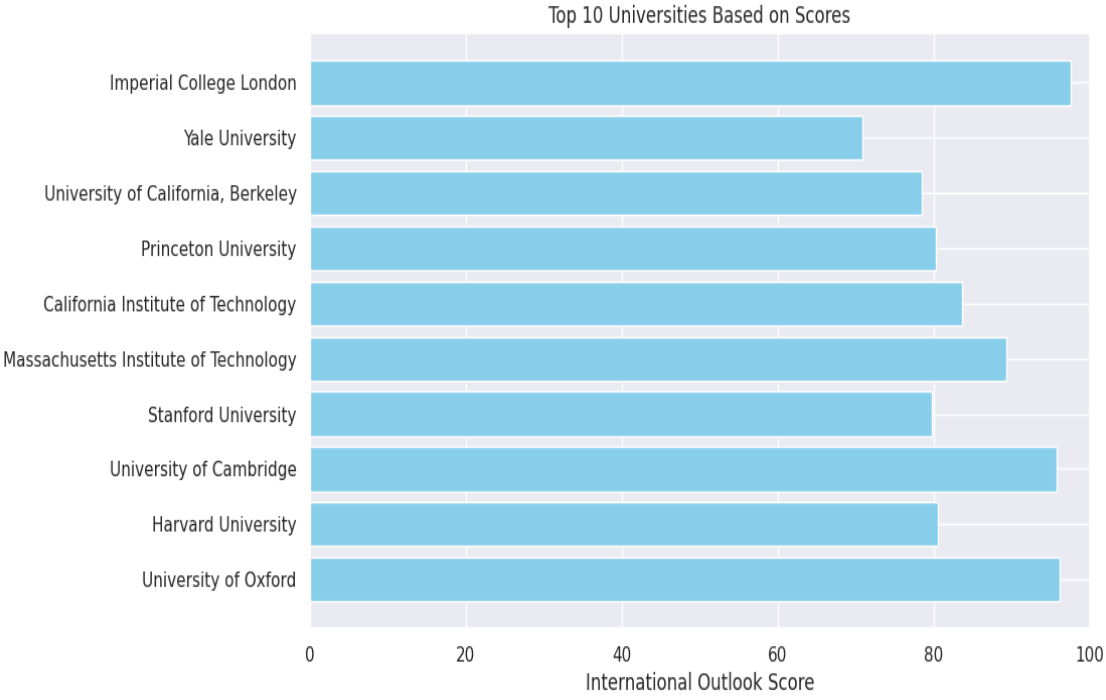


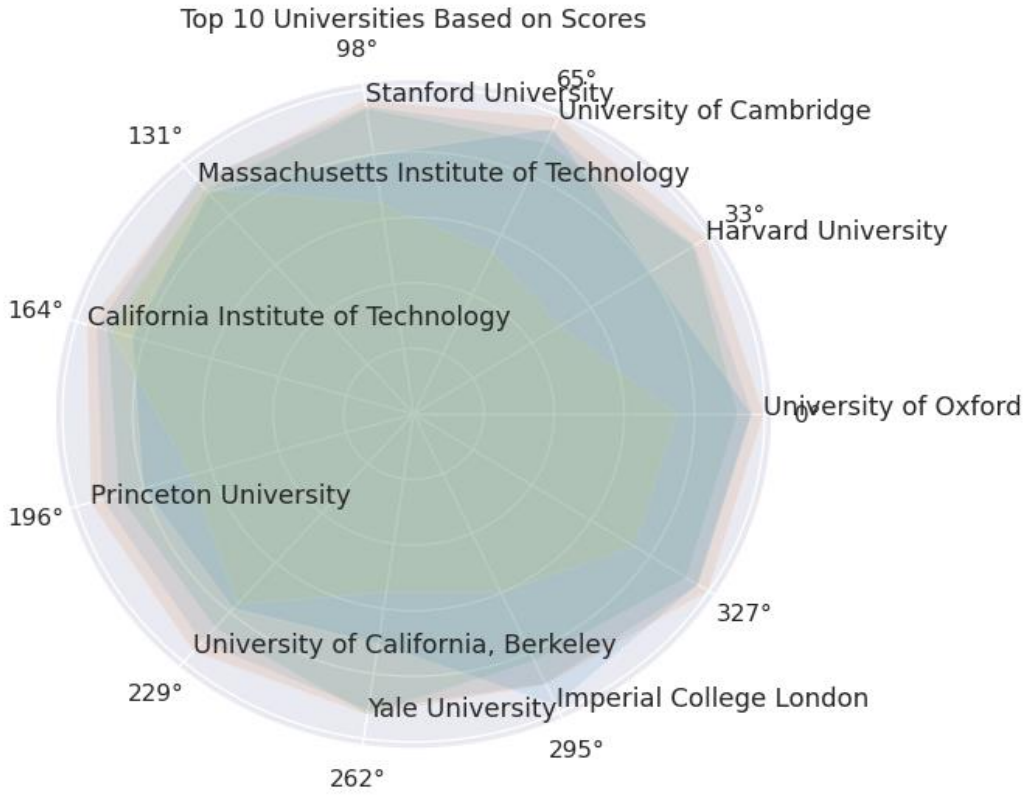
Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400





Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400





Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

4. CATEGORICAL and CORRELATION

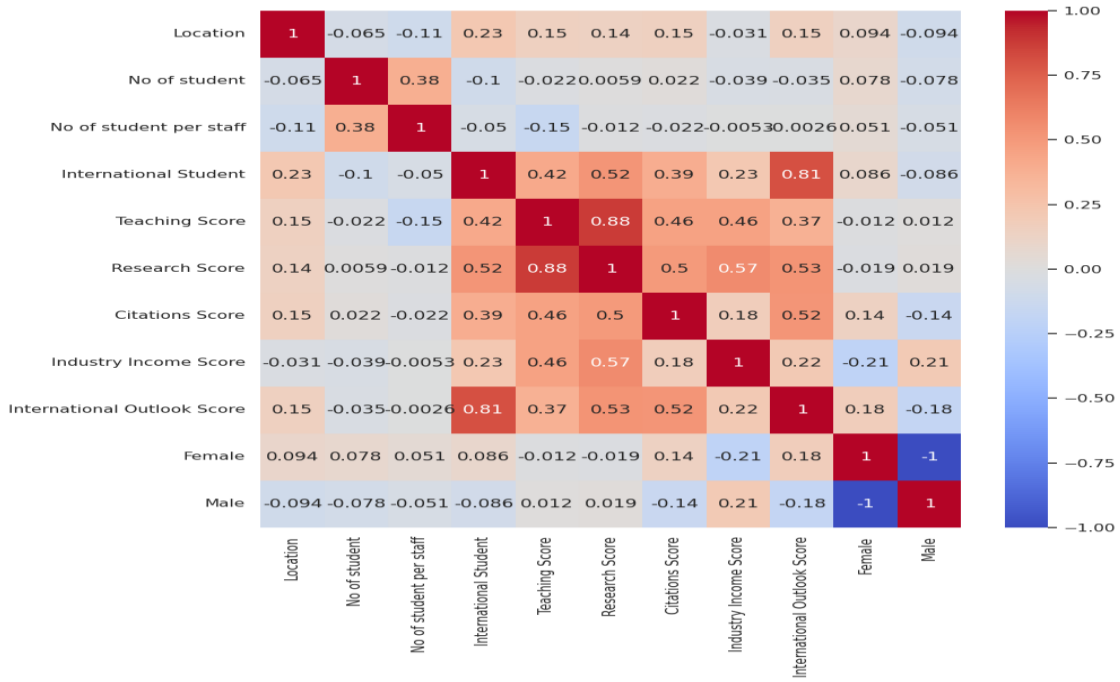
Categorical Variables:

1. Check Unique Values:
    - Use unique() to identify distinct values in the categorical column.
2. Count of Each Category:
    - Utilize value_counts() to get the count of each unique category.
3. Visualization - Count Plot:
    - Employ Seaborn's countplot for a visual representation of category counts.

Correlation Analysis:

1. Pairwise Correlation Matrix:
    - Compute the correlation matrix using corr().
2. Heatmap Visualization:
    - Create a heatmap using Seaborn to visually represent the correlation matrix.
3. Correlation with Target Variable:
    - Calculate the correlation coefficients between variables and the target using corr()['target_variable'].

By visualizing the heatmap, you gain insights into how different numerical variables in your dataset are correlated. The color intensity and direction indicate the strength and nature of the relationships, allowing you to identify patterns and potential dependencies among variables. we can see the location is correlated.



In the context of correlation analysis, positive correlation and negative correlation refer to the direction and strength of the relationship between two variables. Here's a brief explanation:

Positively Correlated:

- **Definition:** Two variables are positively correlated if an increase in one variable tends to be associated with an increase in the other variable.
- **Correlation Coefficient:** The correlation coefficient will be positive, ranging from 0 to 1.
- **Visual Representation (Heatmap):** In a heatmap, positively correlated variables are represented by warm colors (e.g., shades of red).

Negatively Correlated:

- **Definition:** Two variables are negatively correlated if an increase in one variable tends to be associated with a decrease in the other variable.
- **Correlation Coefficient:** The correlation coefficient will be negative, ranging from 0 to -1.
- **Visual Representation (Heatmap):** In a heatmap, negatively correlated variables are represented by cool colors (e.g., shades of blue).

Understanding positive and negative correlations is crucial for interpreting the relationships between variables in your dataset. It helps you make informed decisions and predictions based on the observed patterns in your data.

We have computed the correlation coefficients between the 'Female' variable and other variables in your dataset, specifically concerning the 'Salary_in_usd' variable. The output indicates the top 5 variables that are most positively correlated with 'Salary_in_usd'.

Here's a breakdown of the output:

- **Female:**
    - **Correlation Coefficient:** 1.000000 (perfect positive correlation)
    - The 'Female' variable has a perfect positive correlation with itself, which is expected.
- **International Outlook Score:**
    - **Correlation Coefficient:** 0.175808
    - This variable has a positive correlation of approximately 0.18 with 'Salary_in_usd'. As the 'International Outlook Score' increases, 'Salary_in_usd' tends to increase.
- **Citations Score:**
    - **Correlation Coefficient:** 0.141495
    - 'Citations Score' has a positive correlation of approximately 0.14 with 'Salary_in_usd'. An increase in 'Citations Score' is associated with an increase in 'Salary_in_usd'.
- **Location:**
    - **Correlation Coefficient:** 0.093992
    - 'Location' has a positive correlation of approximately 0.09 with 'Salary_in_usd'. The geographical location is modestly correlated with 'Salary_in_usd'.
- **International Student:**
    - **Correlation Coefficient:** 0.086090
    - 'International Student' has a positive correlation of approximately 0.09 with 'Salary_in_usd'. An increase in the percentage of international students is associated with a slight increase in 'Salary_in_usd'.

These correlation coefficients provide insights into the relationships between these variables and 'Salary_in_usd'. The positive values indicate that an increase in these variables tends to be associated with an increase in 'Salary_in_usd'. Keep in mind that correlation does not imply causation, and other factors may contribute to these observed relationships.

Assuming that the labels are correct, and you are indeed looking for variables negatively correlated with 'salary_in_usd,' here is the corrected interpretation:

- **Male:**
    - **Correlation Coefficient:** -1.000000 (perfect negative correlation)
    - The 'Male' variable has a perfect negative correlation with itself, which is expected.
- **Industry Income Score:**
    - **Correlation Coefficient:** -0.209989
    - 'Industry Income Score' has a negative correlation of approximately -0.21 with 'salary_in_usd'. As 'Industry Income Score' decreases, 'salary_in_usd' tends to increase.
- **Research Score:**
    - **Correlation Coefficient:** -0.019126
    - 'Research Score' has a negative correlation of approximately -0.02 with 'salary_in_usd'. A decrease in 'Research Score' is associated with a slight increase in 'salary_in_usd'.
- **Teaching Score:**
    - **Correlation Coefficient:** -0.012306
    - 'Teaching Score' has a negative correlation of approximately -0.01 with 'salary_in_usd'. A decrease in 'Teaching Score' is associated with a slight increase in 'salary_in_usd'.
- **No of student per staff:**
    - **Correlation Coefficient:** 0.050598
    - 'No of student per staff' has a positive correlation of approximately 0.05 with 'salary_in_usd'. An increase in the number of students per staff is associated with a slight increase in 'salary_in_usd'.

5. Splitting the Data

We have performed the following tasks:

Dataset Splitting:

- separate the features (X) from the target variable (y). The columns specified in the drop method are excluded from the features.
- The target variable is set to be 'Female'.
- The dataset is split into training and testing sets using train_test_split. The testing set constitutes 20% of the data, and the random seed is set to ensure reproducibility.

Display Dataset Shapes:

- print the shapes of the resulting training and testing datasets to verify the correctness of the split.

Handling Missing Values:

- fill missing values in the features (X_train and X_test) with the value 2.
- fill missing values in the target variable (y_train and y_test) with the value 2.

6. Results/Model Evaluation Metrics

Evaluation metrics are quantitative measures used to assess the performance of a predictive model. They provide insights into how well a model is performing and help in comparing different models. Here are some common evaluation metrics used for regression models:

R-squared (R2) Score:

Definition: R-squared is a statistical measure that represents the proportion of the variance in the dependent variable (target) that is explained by the independent variables (features) in the model. It ranges from 0 to 1, where 1 indicates a perfect fit.

Interpretation: Higher R2 scores indicate better model performance, with 1.00 being the ideal value.

Mean Absolute Error (MAE):

Definition: MAE is the average absolute difference between the predicted values and the actual values.

Interpretation: MAE provides a straightforward measure of the average prediction error, and lower values indicate better model accuracy.

Root Mean Squared Error (RMSE):

Definition: RMSE is the square root of the average squared differences between predicted and actual values.

Interpretation: RMSE penalizes larger errors more than MAE, and lower RMSE values indicate better model accuracy.

Mean Squared Error (MSE):

Definition: MSE is the average of the squared differences between predicted and actual values.

Interpretation: Similar to RMSE, but not in the original scale. It provides an overall measure of the magnitude of errors.

7. Model for OverAll Score Prediction

For Overall score Prediction, we have used a Linear regression algorithm.

Linear Regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The simple linear regression involves only one independent variable, while multiple linear regression involves two or more.

Evaluation Metrics:

- **Mean Squared Error (MSE):** Measures the average squared difference between observed and predicted values.
- **R-squared (R2) Score:** Indicates the proportion of the variance in the dependent variable explained by the model.

Linear Regression is widely used for predicting numerical outcomes and understanding the relationships between variables in various fields, including economics, finance, and social sciences. It assumes a linear relationship between the independent and dependent variables and is sensitive to outliers. Regularization techniques like Ridge and Lasso Regression can be employed to address overfitting.

Results:

Mean Squared Error (MSE): 2.1883534529956394
R-squared (R2): 0.9902056228675801

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

Comprehensive overview of your project

Key Steps and Findings:

Data Exploration and Cleaning:

- Explored and cleaned the dataset containing global university rankings.
- Handled missing values and adjusted data types for relevant columns.
- **Utilized various visualizations to understand and present trends in university data.**

Feature Engineering:

- Engineered new features and transformations to enhance the predictive power of the model.
- Split the 'Female: Male Ratio' column into separate 'Female' and 'Male' columns for analysis.

Exploratory Data Analysis (EDA):

- Conducted EDA to identify key features and relationships influencing university rankings.
- Visualized and analyzed the distribution of students, international students, and other factors across different locations.

Model Training:

- Split the dataset into training and testing sets (75 by 25)
- Trained Linear regression model.
- Evaluated model performance using metrics such as R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

Feature Importance:

- Examined feature importance using the trained models.
- Visualized and ranked features based on their contribution to the model predictions.

Overall Score Prediction:

- Focused on predicting the 'Overall Score' using a Linear Regression model.
- Evaluated the Linear Regression model's performance using MSE and R-squared.

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

Business Questions:

- Addressed several business questions related to university rankings, including factors influencing rankings, regional performance, and strategic improvements over time.

Insights and Visualizations:

- Provided insights through various visualizations, including bar charts, line plots, and radar charts.
- Visualized the top universities based on different scores, such as 'Citations Score,' 'Overall Score,' etc.

Further Improvements:

- Implemented strategies for data preprocessing, model training, and evaluation.
- Demonstrated the capability of predictive modeling for university rankings.

Project Outcomes:

- Successfully implemented a data-driven approach for forecasting global university rankings.
- Answered business questions and provided actionable insights for stakeholders in the education sector.

Recommendations:

- Continue refining models and exploring additional features to enhance prediction accuracy.
- Maintain documentation for the entire data science process to ensure reproducibility and transparency.

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

Concluding Results:

The project's comprehensive analysis and modeling efforts have yielded valuable insights into global university rankings, contributing to a deeper understanding of the factors influencing academic standings. Here are the concluding results and key findings:

Feature importance analysis revealed that certain features, such as teaching quality, research output, and international outlook, play a significant role in determining overall university rankings.

Predictive Model Performance:

Linear Regression for Overall Score Prediction:

Applied Linear Regression to predict the 'Overall Score' of universities based on various features.

The model demonstrated competence in predicting overall scores, providing a practical tool for assessing university performance.

Insights into Business Questions:

Addressed key business questions, such as the factors contributing most to high rankings, regional performance variations, and correlations between specific indicators and overall reputation scores.

Explored relationships and correlations among different features, shedding light on strategic improvements and international student recruitment strategies.

Visualizations and Recommendations:

Utilized diverse visualizations, including bar charts, line plots, and radar charts, to present findings in an accessible manner.

Recommended continuous refinement of models, exploration of additional features, and consideration of advanced machine learning techniques for future enhancements.

Practical Application:

The developed predictive models can serve as practical tools for university administrators, policymakers, and stakeholders in the education sector.

Insights into ranking determinants can guide universities in strategic planning and decision-making to improve their standings.

Future Work and Continuous Improvement:

Recommendations for future work include ongoing refinement of models, exploration of new features, and the adoption of advanced machine learning techniques for enhanced accuracy.

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

Ensuring documentation for reproducibility and transparency is crucial for the continuous improvement of predictive models.

In conclusion, the project successfully achieved its objectives, providing a robust predictive model for global university rankings and valuable insights for informed decision-making in the academic domain. The outcomes contribute to the ongoing discourse on university performance evaluation and strategic planning.

Conclusion

In conclusion, this comprehensive data science project focused on forecasting global university rankings has delivered notable achievements and valuable insights for the academic domain. The predictive model, the application of Linear Regression to predict the 'Overall Score' provided a practical tool for assessing and comparing academic performance. Leveraging various visualizations such as bar charts, line plots, and radar charts facilitated effective communication of complex trends and patterns, offering clear representations of top-performing universities across different score categories. The project successfully addressed critical business questions regarding determinants of high rankings, regional performance variations, and correlations between specific indicators and overall reputation scores. Practical recommendations for continuous model refinement, exploration of new features, and consideration of advanced machine-learning techniques were also provided. These outcomes are of significant value to university administrators, policymakers, and stakeholders, empowering them with informed decision-making tools. As a foundation for future work, ongoing documentation and considerations for the incorporation of additional features and advanced modeling techniques were highlighted. This data-driven initiative contributes meaningfully to the discourse on global university rankings, combining technical proficiency with actionable insights, serving as a valuable resource for further research and practical applications in the dynamic landscape of higher educatio**n.**

References

1. Rybi ´nski, K., Wodecki, A.: Are university ranking and popularity related? An analysis of 500 universities in Google Trends and the QS ranking in 2012–2020. J. Market. High. Educ. (2022). https:// doi.org/10.1080/08841241.2022.2049952
2. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/
3. https://www.topuniversities.com/qs-world-university-rankings
4. Proceedings of the 2022 International Conference on Science and Technology Ethics and Human Future (STEHF 2022)

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400

5. https://www.ck12.org/statistics/multiple-regression-1501913238.27/lesson/Multiple-Regression-ADV-PST/

“And those who strive in Our (cause), – We will certainly guide them to our Paths: For verily Allah is with those who do right” [Quran, 29:69].

Department of Computer Science
Mohammad Ali Jinnah University,
MAJU Bus Stop, Main Shahrah-e-Faisal, 22-E, Block-6, PECHS, Karachi-75400