INTEGRATED DATA ENVELOPMENT ANALYSIS - ML FRAMEWORK FOR GLOBAL UNIVERSITY EFFICIENCY ANALYSIS

DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF SCIENCE
IN
APPLIED MATHEMATICS

Submitted by

MEHAK GOYAL (23/MSCMAT/67)

Under the supervision of

PROF. ANJANA GUPTA

DEPARTMENT OF APPLIED MATHEMATICS
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

MAY, 2025

DEPARTMENT OF APPLIED MATHEMATICS
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **MEHAK GOYAL**, Roll No's – **23/MSCMAT/67** students of MSc. (**Applied Mathematics**),hereby declare that the project Dissertation titled "**Integrated DATA ENVELOPMENT ANALYSIS -ML Framework for Global University Efficiency Analysis**" which is submitted by us to the **Department of Applied Mathematics**, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of degree of Master of Science, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi Mehak Goyal

Date: 26.05.2025 23/MSCMAT/67

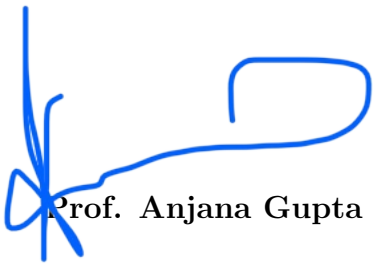
This is to certify that the student has incorporated all the corrections suggested by the examiners in the dissertation and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor **Signature of External Examiner**

DEPARTMENT OF APPLIED MATHEMATICS
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "**Integrated DATA ENVELOPMENT ANALYSIS -ML Framework for Global University Efficiency Analysis**" which is submitted by **MEHAK GOYAL**, Roll No's – **23/MSCMAT/67**, **Department of Applied Mathematics** ,Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Science, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.



Place: Delhi **Prof. Anjana Gupta**

Date: 26.05.2025 **SUPERVISOR**

DEPARTMENT OF APPLIED MATHEMATICS
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to **Prof. ANJANA GUPTA** for her continuous guidance and mentorship that she provided me during the project. She showed me the path to achieve the targets by explaining all the tasks to be done and explained me the importance of this project as well as its industrial relevance. She was always ready to help me and clear my doubts regarding any hurdles in this project. Without her constant support and motivation, this project would not have been successful.

Place: Delhi — Mehak Goyal

Date: 26.05.2025 — 23/MSCMAT/67

Abstract

This thesis evaluates the performance of 800 universities worldwide by integrating Data Envelopment Analysis (DEA) with machine learning techniques. Utilizing data from the 2016 global rankings, the study applies input-oriented CCR, BCC, and NIRS DEA models to measure technical, scale, and overall efficiency. The analysis treats the student-to-staff ratio as the input and employs teaching, citation, research, citation, industry income, and international outlook scores as outputs.

The DEA results indicate significant scope for efficiency improvement, with a mean overall (CCR) efficiency of approximately 0.108 and a mean technical (BCC) efficiency of 0.189. A predominant finding is that 86.75% of universities exhibit Increasing Returns to Scale (IRS), suggesting most were operating below optimal scale. Sensitivity analysis, conducted by altering output specifications, showed that while absolute efficiency scores and RTS distributions changed (Spearman rank correlation of ˜0.81 for BCC scores), the relative rankings of universities demonstrated considerable robustness.

K-Means clustering (K=2, determined via Silhouette analysis) grouped universities based on contextual variables (location, student numbers, female-male ratio, international student percentage), identifying a large primary cluster and a very small cluster of distinct mega-scale institutions. DEA performed within these clusters highlighted improved relative efficiency scores, especially for the smaller cluster, when benchmarked against more homogenous peers.

Finally, tuned Random Forest, LightGBM, and Gradient Boosting regression models were developed to explain technical efficiency. LightGBM performed best, achieving an R-squared of approximately 0.4725 in predicting BCC scores. Key contextual drivers identified were total student numbers, location, and percentage of international students. This multi-stage approach provides a nuanced understanding of university performance, offering actionable insights for strategic planning and policy development in the higher education sector.

Contents

List of Tables

List of Figures

List of Symbols

DEA	Data Envelopment Analysis
DMU	Decision Making Unit (e.g., a university in this study)
CCR	Charnes-Cooper-Rhodes DEA model (Constant Returns to Scale)
BCC	Banker-Charnes-Cooper DEA model (Variable Returns to Scale)
NIRS	Non-Increasing Returns to Scale DEA model
RTS	Returns to Scale
IRS	Increasing Returns to Scale
DRS	Decreasing Returns to Scale
CRS	Constant Returns to Scale
SE	Scale Efficiency
ML	Machine Learning
K-Means	K-Means Clustering Algorithm
LightGBM	Light Gradient Boosting Machine
R^2	Coefficient of Determination (Model Performance Metric)
MSE	Mean Squared Error
Silhouette Score	A metric to evaluate clustering quality
λ	Weight assigned to peer DMUs in DEA
θ	Efficiency score in DEA
`stats_number_students`	Total number of students (contextual feature)
`stats_female_male_ratio`	Female to male student ratio
`stats_pc_intl_students`	Percentage of international students
`location_Encoded`	Encoded geographical location of the university
`scores_teaching`	Teaching performance score
`scores_research`	Research performance score
`scores_citations`	Citations score (research impact)
`scores_industry_income`	Industry income score
K	Number of clusters in clustering analysis

Chapter 1

INTRODUCTION

The global higher education sector operates within an increasingly complex and competitive environment, demanding greater accountability and optimal utilization of resources. As multifaceted institutions, universities transform various inputs—such as academic staff, financial resources, and infrastructure—into multiple outputs, including educated graduates, research contributions, and societal engagement. Evaluating the efficiency with which these transformations occur is paramount for institutional self-improvement, strategic decision-making by administrators, and evidence-based policy formulation by governing bodies. Traditional performance metrics often fail to capture the holistic nature of university operations, particularly the interplay between multiple inputs and outputs.

Data Envelopment Analysis (DEA) provides a powerful, non-parametric approach to measure the relative efficiency of comparable units, or Decision-Making Units (DMUs) — in this case, universities. Introduced by Charnes, Cooper, and Rhodes in 1978, DEA works by constructing an "efficient frontier" based on the performance of the most efficient DMUs observed. It then assesses the efficiency of all other units against this frontier, crucially without assigning predetermined weights to inputs and outputs or making assumptions about the underlying production function.

However, while DEA is adept at identifying *what* level of efficiency a DMU achieves and *where* potential improvements lie (i.e., input reductions or output augmentations), it does not inherently explain *why* some DMUs are more efficient than others, particularly when considering contextual or environmental factors not directly included as inputs or outputs. Moreover, the significant heterogeneity among universities globally—in size, mission, funding, and operational context—can complicate direct comparisons and the interpretation of efficiency scores.

This thesis adopts an integrated, multi-stage analytical framework to address these

challenges. This framework combines the strengths of DEA for efficiency measurement with the capabilities of machine learning (ML) techniques. Specifically, ML is used to handle institutional heterogeneity through clustering and identify key contextual factors significantly associated with the DEA-derived efficiency scores. This approach aims to provide a more nuanced, robust, and comprehensive understanding of university performance.

1.1 BACKGROUND

The quest to measure and enhance efficiency in higher education institutions (HEIs) has been a consistent theme in academic research and policy discussions for several decades. Universities worldwide are under continuous pressure to deliver high-quality education and impactful research while managing resources effectively. Data Envelopment Analysis (DEA) has become a cornerstone methodology in this domain since its inception. Its ability to handle multiple inputs and outputs simultaneously without requiring a pre-defined production function makes it particularly suitable for analyzing complex organizations like universities.

Foundational DEA models, such as the CCR model (assuming Constant Returns to Scale, or CRS) and the BCC model (assuming Variable Returns to Scale, or VRS), enable the assessment of overall efficiency, pure technical efficiency, and scale efficiency. These approaches have been extensively used to compare universities across various regions, shedding light on their relative performance.

However, DEA has its drawbacks: efficiency estimates can vary significantly depending on the selected inputs and outputs as well as the sample of institutions, and because DEA is a deterministic technique, it treats all deviations from the frontier as inefficiency without accounting for statistical noise or unmeasured environmental influences.

Researchers have increasingly focused on multi-stage analytical approaches to overcome some of these limitations. This often involves using DEA in the first stage to estimate efficiency scores and then employing statistical or machine-learning techniques in the second stage to regress these scores against potential explanatory variables. Machine learning algorithms like Random Forest, LightGBM, and clustering methods like K-Means are particularly well-suited for this task. They can identify complex non-linear

relationships, handle diverse data types, and group heterogeneous DMUs into more comparable subsets, enriching the insights derived from DEA. This thesis embraces such an integrated methodology, applying it to a contemporary global dataset of universities to offer fresh perspectives on their operational efficiency.

1.2 PROBLEM STATEMENT

Measuring and understanding university efficiency is critical, yet it presents significant analytical challenges. While DEA is a robust tool for assessing relative efficiency, its standard application may not fully account for the substantial heterogeneity among global universities. Institutions differ widely in their scale of operations, funding models, national contexts, and strategic priorities. Comparing a small, specialized institution with a large, comprehensive university using a single DEA model might yield results that are difficult to interpret or act upon.

Furthermore, identifying the factors driving efficiency is as important as measuring efficiency. DEA provides target improvements but does not inherently pinpoint which specific institutional characteristics or environmental factors (not used as direct inputs or outputs) are associated with higher or lower performance. This limits the ability of university managers and policymakers to develop targeted interventions.

Therefore, there is a clear need for an analytical framework that can:

1. Rigorously measure university efficiency while acknowledging and addressing institutional heterogeneity.

2. Assess the robustness of efficiency findings to variations in model specification.

3. Identify and quantify the key contextual drivers or correlates of university efficiency.

This research aims to tackle this problem by developing and applying an integrated DEA-clustering-ML framework to a global sample of universities, thereby providing a more comprehensive and actionable understanding of their performance.

1.3 OBJECTIVES

The central aim of this dissertation is to perform an in-depth evaluation of worldwide university efficiency by combining Data Envelopment Analysis (DEA) with clustering methods and machine learning techniques. The specific objectives are:

1. To define and preprocess a suitable dataset of global universities, specifying appropriate inputs and outputs for DEA based on the available data and relevant literature.

2. To apply input-oriented DEA models (CCR, BCC, and NIRS) to the full sample of universities to calculate their technical efficiency, overall efficiency, scale efficiency, and determine their respective returns to scale (IRS, CRS, DRS).

3. To conduct a sensitivity analysis on the primary DEA model by altering the output specification to evaluate the robustness of the efficiency results and returns to scale classifications.

4. To employ K-Means clustering, informed by Silhouette analysis and the Elbow method, to segment the universities into more homogenous groups based on selected contextual variables.

5. To profile the identified university clusters based on their input/output variables, contextual characteristics, and initial DEA scores to understand their distinct features.

6. To perform DEA (specifically the BCC model) within each identified cluster to assess the relative technical efficiency of universities against more comparable peers and compare these with their overall efficiency scores.

7. To develop, tune, and evaluate supervised machine learning models (Random Forest, LightGBM, Gradient Boosting) to identify and quantify the significant contextual university characteristics associated with the DEA-derived technical efficiency scores.

8. To synthesize the findings from the DEA, sensitivity analysis, clustering, and machine learning stages to provide a holistic understanding of university efficiency, its drivers, and to offer pertinent policy and managerial implications.

Chapter 2

LITERATURE REVIEW

This chapter surveys the core principles of Data Envelopment Analysis (DEA), examines how it has been applied to evaluate the efficiency of higher education institutions (HEIs), and explores recent advances in integrating machine learning methods to augment DEA-based performance assessments.

2.1 DATA ENVELOPMENT ANALYSIS (DEA) FUNDAMENTALS

Data Envelopment Analysis (DEA), initially developed by Charnes, Cooper, and Rhodes in 1978, is a non-parametric and mathematical programming method. Its purpose is to assess the relative efficiency of multiple decision-making units (DMUs) that convert various inputs into multiple outputs. This approach is especially suited to non-profit entities like universities, where market-based valuations of inputs and outputs are often unavailable or unsuitable. DEA constructs an empirical "best-practice" frontier from the performance of the most efficient DMUs and then assigns efficiency scores to all other units by comparing them against this benchmark.

2.1.1 CCR Model

The CCR model, named after Charnes, Cooper, and Rhodes (1978), assumes Constant Returns to Scale (CRS). This assumption implies that any proportional input change will result in an equi-proportional output shift. The CCR model measures overall efficiency, integrating technical efficiency (managerial ability to use inputs) and scale efficiency (op-

erating at the optimal size). DMUs with a CCR score of one are considered globally efficient, lying on the CRS frontier.

2.1.2 BCC Model

Banker, Charnes, and Cooper (1984) introduced the BCC model to relax the CRS restriction by permitting Variable Returns to Scale (VRS). Under this framework, the efficiency frontier can reflect increasing, constant, or decreasing returns to scale. By comparing each decision-making unit only to peers of comparable size, the BCC model isolates pure technical efficiency from scale effects. Therefore, a BCC efficiency score of 1 signifies that a unit is fully technically efficient, independent of its scale of operations.

2.1.3 Scale Efficiency and Returns to Scale (RTS)

The difference between the CCR and BCC efficiency scores for a DMU indicates the presence of scale inefficiency. We can calculated Scale Efficiency (SE) as the ratio of the CCR score to the BCC score (SE = CCR / BCC). An SE of 1 implies that the DMU is scale efficient (i.e., its technical efficiency under CRS and VRS is the same). Further analysis, often involving a Non-Increasing Returns to Scale (NIRS) model, helps distinguish between Increasing Returns to Scale (IRS– DMU is too small) and Decreasing Returns to Scale (DRS– DMU is too large). As used in this thesis, the NIRS model assumes that the sum of lambdas (peer weights) is less than or equal to 1.

2.1.4 Input and Output Orientation

DEA models can be input-oriented or output-oriented. An input-oriented model seeks to minimize inputs while maintaining current output levels, asking "To what extent can we proportionally decrease all inputs while still maintaining the same output levels?" An output-oriented model aims to maximize outputs for a given level of inputs. The choice of orientation depends on the managerial control context; this study employs input oriented models, assuming universities have more control over their resource utilization for given performance targets.

2.2 APPLICATIONS OF DEA IN HIGHER EDUCATION

DEA has been widely adopted for efficiency analysis in the higher education sector globally. Studies typically use inputs such as academic staff numbers, student enrollment (as a proxy for resources consumed or an input factor), research funding, and infrastructure. Outputs commonly include graduate numbers, research publications, citation counts, research income, and sometimes measures of teaching quality or societal impact. These studies vary in scope, from national systems to international comparisons, and often aim to identify benchmarks and policy implications for improving HEI performance.

2.3 ADDRESSING HETEROGENEITY: CLUSTERING IN DEA

A significant challenge in applying DEA to diverse sets of HEIs is their inherent heterogeneity. Universities differ in size, mission, funding, location, and student body composition. Applying DEA without considering such differences can lead to unfair comparisons. Clustering techniques, such as K-Means, are increasingly used as a precursor or an integrated step to group DMUs into more homogenous subsets based on relevant contextual variables. DEA is then performed within these clusters, allowing for more meaningful relative efficiency assessments against truly comparable peers. This study utilizes K-Means clustering based on variables like university location, student numbers, female-male ratio, and international student percentage.

2.4 EXPLAINING EFFICIENCY: SECOND-STAGE ANALYSIS WITH MACHINE LEARNING

While DEA identifies efficient and inefficient DMUs and quantifies the magnitude of inefficiency, it does not inherently explain the sources of these variations, particularly concerning factors not included in the input-output model. A common approach is to use a two-stage analysis, where DEA efficiency scores (obtained in the first stage) are regressed

against a set of explanatory environmental or contextual variables in the second stage.

Traditional statistical methods like Tobit regression have been used for this second stage. Still, machine learning models have become increasingly popular because they can capture non-linear patterns and complex interactions and provide reliable assessments of feature importance without relying on strict distributional assumptions. Tree-based ensemble methods like Random Forest and gradient boosting machines like LightGBM and XGBoost are potent. They can be used for regression (predicting continuous efficiency scores) and classification (predicting discrete efficiency categories). This thesis employs Random Forest, LightGBM, and Scikit-learn's Gradient Boosting for this second-stage analysis.

2.5 INTEGRATED DEA-ML FRAMEWORKS

The integration of DEA with ML techniques forms a synergistic analytical framework. This study contributes by applying a systematic multi-stage methodology that encompasses:

1. Robust DEA modeling (CCR, BCC, NIRS) with sensitivity analysis to test model stability.
2. Data-driven clustering (K-Means with optimal K selection) to address heterogeneity.
3. Contextualized DEA-within-cluster analysis for fairer peer comparisons.
4. Application of tuned ML models (Random Forest, LightGBM, Gradient Boosting) for explaining efficiency based on contextual variables.

This comprehensive approach is designed to yield more nuanced, robust, and actionable insights into university performance than standalone methods.

Chapter 3

METHODOLOGY

This chapter outlines the methodological framework employed in this research to assess the efficiency of global universities and identify factors associated with their performance. It covers the research design, data sources and variable selection, data preprocessing steps, the specific Data Envelopment Analysis (DEA) models used, clustering techniques, and the machine learning approaches for second-stage analysis.

3.1 RESEARCH DESIGN

The investigation adopts a quantitative, multi-stage analytical strategy, integrating Data Envelopment Analysis (DEA), clustering, and supervised machine learning. Its principal objective is to deliver a holistic assessment of institutional efficiency and to identify the characteristics that most strongly influence performance.

1. **Data Acquisition and Preprocessing:** First, data for 800 universities are collected from the 2016 world rankings. These raw data undergo a rigorous preprocessing pipeline: variables are cleansed (e.g., removal of non-numeric characters), missing values are imputed using median strategies, categorical fields (such as country) are encoded numerically, and inputs and outputs are adjusted to satisfy DEA's requirement for strictly positive values.

2. **Primary DEA Modeling:** Next, three input-oriented DEA models—CCR (Constant Returns to Scale), BCC (Variable Returns to Scale), and NIRS (Non-Increasing Returns to Scale)—are applied to the entire dataset. These models produce overall, pure technical, and scale efficiency scores for each university. A sensitivity analysis follows, in which output specifications are altered to assess the robustness of

efficiency estimates and Returns to Scale classifications.

3. **Clustering of DMUs:** K-means clustering is employed on key contextual variables (e.g., student numbers, international student share, encoded location, female-male ratio) to address institutional heterogeneity. Optimal cluster count is determined via the Elbow and Silhouette methods, resulting in more homogeneous subgroups for subsequent comparison.

4. **Contextualized DEA Modeling:** The input-oriented BCC model is rerun within each resulting cluster to obtain efficiency scores relative to comparable peers. This sub-group DEA highlights how benchmarks shift when universities are evaluated among similar institutions rather than across the global spectrum.

5. **Second-Stage Machine Learning Analysis:** Developing and evaluating Random Forest, LightGBM, and Gradient Boosting regression models to identify and quantify the association between contextual university characteristics and DEA-derived technical efficiency scores.

3.2 DATA SOURCE AND VARIABLE SELECTION

The data for this study are derived from the 2016 World University Rankings, specifically the `2016_rankings.csv` file. After necessary cleaning and ensuring completeness for DEA variables, the dataset comprises 800 universities, which serve as the Decision Making Units (DMUs).

3.2.1 Input and Output Variables for DEA

The selection of inputs and outputs is crucial for meaningful DEA. Based on the data and common practice in HEI efficiency literature, the following variables were defined:

- **Input Variable (1):**
  - `stats_student_staff_ratio`: Student-staff ratio.
- **Output Variables (5):**
  - `scores_teaching`: Teaching score.

- `scores_research`: Research score.
- `scores_citations`: Citations score.
- `scores_industry_income`: Industry income score.
- `scores_international_outlook`: International outlook score.

3.2.2 Contextual Variables for Clustering and Machine Learning

To explore the impact of institutional characteristics beyond the direct DEA inputs/outputs, the following contextual variables were selected:

- `DMU_NAME_COLUMN` ('name'): Name of the university.
- `COUNTRY_COLUMN` ('location'): Country of the university. This was label encoded as `location_Encoded` for use in models.
- `NUM_STUDENTS_COLUMN` ('stats_number_students'): Total number of students.
- `FEMALE_MALE_RATIO_COLUMN` ('stats_female_male_ratio'): Ratio of female to male students.
- `PC_INTL_STUDENTS_COLUMN` ('stats_pc_intl_students'): Percentage of international students.

3.3 DATA PREPROCESSING

The raw dataset was subjected to a comprehensive preprocessing pipeline to ensure accuracy and compatibility with DEA requirements:

1. **Data Cleaning and Type Conversion.**
   - Removed thousands-separator commas from `stats_number_students` and converted the result to integer.
   - Stripped the "%" symbol from `stats_pc_intl_students` and divided by 100 to obtain a proportion.

- Parsed the `stats_female_male_ratio` string (e.g., "40:60") into a numeric female-share value (0.40) and stored it as a float.

2. **Categorical Encoding.** The university's country label (`location`) was transformed into a numeric code (`location_Encoded`) using `sklearn.preprocessing.LabelEncoder`, allowing it to serve as an input to machine-learning algorithms.

3. **Missing-Value Treatment.**

   - Any university record missing one of the core DEA variables (`essential_dea_cols`) was discarded via `DataFrame.dropna()`; in this run, no records were removed, preserving all 800 DMUs.
   - Remaining gaps in auxiliary contextual variables (e.g. the 52 missing entries in `stats_female_male_ratio`) were imputed with the column median using `sklearn.impute.SimpleImputer(strategy='median')`.

4. **Ensuring Strict Positivity.** Since DEA requires all inputs and outputs to be strictly positive, any zero or negative values—most notably in `scores_industry_income` were replaced with a small epsilon constant ($\varepsilon = 1 \times 10^{-6}$) to avoid infeasibility in the linear programming solver.

5. **Index Reinitialization.** After all cleaning, encoding, and imputation steps, the DataFrame's index was reset (`.reset_index(drop=True)`) to ensure a continuous 0–799 index range for downstream analysis.

3.4 DEA MODELING

This study adopts an input-oriented perspective—universities are assumed to seek the minimum possible use of resources while maintaining their current output levels. To accelerate computations across 800 decision-making units (DMUs), the DEA routines were executed in parallel using `joblib` in conjunction with `psutil` for efficient CPU management.

3.4.1 CCR, BCC, and NIRS Formulations

We implemented three linear-programming formulations, all solved via the Highs solver in `scipy.optimize.linprog`, with efficiency scores constrained to the unit interval [0,1]:

- `dea_ccr_input_oriented` This model assumes *Constant Returns to Scale* (CRS). It minimizes the scalar θ—representing the proportional reduction in all inputs—subject to:

$$\sum_j \lambda_j X_{ij} \;\le\; \theta\, X_{ik} \quad foreachinputi, \quad \sum_j \lambda_j Y_{rj} \;\ge\; Y_{rk} \quad foreachoutputr.$$

  A DMU on the CRS frontier attains $\theta = 1$.

- `dea_bcc_input_oriented` Extending the CCR model, this formulation enforces the convexity constraint

$$\sum_j \lambda_j = 1,$$

  thereby permitting *Variable Returns to Scale* (VRS). The resulting efficiency score reflects pure managerial performance, independent of scale effects.

- `dea_nirs_input_oriented` Similar to the BCC model but with the relaxed constraint

$$\sum_j \lambda_j \;\le\; 1,$$

  this *Non-Increasing Returns to Scale* (NIRS) variant captures situations where scaling up inputs cannot yield more than proportional output increases.

3.4.2 Scale Efficiency and Returns to Scale Classification

Once CCR and BCC scores are obtained, the *Scale Efficiency* (SE) of each DMU is calculated as the ratio:

$$SE \;=\; \frac{\texttt{DEA\_CCR\_Score}}{\texttt{DEA\_BCC\_Score}}. \tag{3.1}$$

An SE value of 1 indicates no scale inefficiency.

To classify the Returns to Scale (RTS) status of each DMU, we compare the three efficiency measures with a numerical tolerance $\varepsilon = 10^{-5}$:

- **CRS:** If `DEA_NIRS_Score` $\approx$ `DEA_BCC_Score` $\approx$ `DEA_CCR_Score`, the DMU operates under constant returns.
- **IRS:** If `DEA_NIRS_Score` $\approx$ `DEA_BCC_Score` and `DEA_BCC_Score` $>$ `DEA_CCR_Score`, the DMU exhibits increasing returns, suggesting it is too small relative to the optimal scale.
- **DRS:** If `DEA_NIRS_Score` $<$ `DEA_BCC_Score`, the DMU demonstrates decreasing returns, indicating potential over-sizing.

This tiered approach distinguishes pure technical efficiency from scale effects and provides a clear taxonomy of scale-related performance for each university.‘‘‘

3.4.3 DEA Sensitivity Analysis

To verify the stability of our results, we conducted a sensitivity analysis by re-running the DEA models using a reduced set of outputs:

$$\texttt{output\_cols\_scenario2} = \{\texttt{scores\_teaching}, \texttt{scores\_research}, \texttt{scores\_citations}\}.$$

We then compared the resulting efficiency scores, the distribution of RTS classifications, and calculated the Spearman rank correlation between the original BCC scores and those from the scenario analysis.

3.4.4 Peer (Benchmark) Analysis

For DMUs identified as inefficient (for instance, those with a BCC score below 0.5), we extracted their reference peers from the BCC model's λ_j weights. A peer is defined as any DMU for which $\lambda_j > 10^{-5}$. These peers serve as benchmarks, guiding inefficient units toward best-practice performance.

3.5 Clustering of Universities

To account for heterogeneity across institutions, we applied K-Means clustering to group universities into more homogeneous subsets.

3.5.1 Feature Selection and Scaling

We selected four contextual variables for clustering:

{`location_Encoded`, `stats_number_students`, `stats_female_male_ratio`, `stats_pc_intl_students`}

Missing values in these features were imputed with the median, and all variables were standardized using `StandardScaler` to ensure equal weighting in the distance calculations.

3.5.2 Determining the Optimal Number of Clusters (K)

The appropriate number of clusters was chosen using two complementary methods:

1. **Elbow Method:** We plotted the total within-cluster sum of squared distances (WCSS) against K (ranging from 2 to 10). The "elbow" point—where additional clusters yield diminishing returns in WCSS reduction—suggests an optimal K.

2. **Silhouette Analysis:** For each K, we computed the average Silhouette score, which measures how similar each university is to its own cluster versus other clusters. The K with the highest average Silhouette score indicates the most cohesive clustering structure.

3.5.3 Cluster Profiling

After applying K-Means with $K = 2$, each university was assigned to one of two clusters. To characterize these groups, we computed:

- **Cluster Centroids (Mean Values):** For each numeric variable—including the original DEA efficiency scores (CCR, BCC, NIRS), scale efficiency, and the four clustering features (`location_Encoded`, `stats_number_students`, `stats_female_male_ratio`, `stats_pc_intl_students`)—we calculated the arithmetic mean within each cluster. These centroid values summarize the typical profile of universities in each subgroup.

- **Dominant Returns to Scale:** We determined the most frequent RTS category (IRS, CRS, or DRS) in each cluster by computing the statistical mode of the `Returns_to_Scale` variable.

- **Statistical Comparison via ANOVA:** To assess whether cluster membership corresponds to significant differences in technical efficiency, we conducted a one-way ANOVA on the `DEA_BCC_Score` across the two clusters. A p-value below the chosen significance level (e.g., 0.05) would indicate that the mean BCC scores differ meaningfully between clusters.

These profiling steps reveal how the clusters differ in terms of both contextual attributes and baseline efficiency performance, providing insight into the characteristics of the more and less efficient groups.

3.5.4 DEA-Within-Clusters Analysis

To obtain efficiency benchmarks tailored to each subgroup's context, we re-applied the input-oriented BCC model separately within Cluster 1 and Cluster 2. This produced a new efficiency metric, `DEA_BCC_Score_Within_Cluster`, for each university. By comparing these within-cluster scores against the original, global BCC scores, we can evaluate the extent to which universities perform differently when measured against peers of similar scale and environment. Such contextualized efficiency estimates help identify institutions that benefit most from homogeneous benchmarking and those whose global rankings may mask underlying performance nuances.

3.6 Second-Stage Machine Learning Analysis

In the final stage of the framework, we employ supervised learning techniques to uncover which contextual attributes best explain the variation in pure technical efficiency (`DEA_BCC_Score`) across universities.

3.6.1 Definition of Variables

- **Response Variable:** The continuous `DEA_BCC_Score`, representing each institution's technical efficiency under Variable Returns to Scale.
- **Predictor Variables:**
  - `location_Encoded`: Numeric encoding of the university's country.

- `stats_female_male_ratio`: Proportion of female students.
- `stats_number_students`: Total student enrollment.
- `stats_pc_intl_students`: Share of international students.

Prior to modeling, all predictors were standardized (zero mean, unit variance) using `StandardScaler` to ensure comparability and avoid dominance by variables with larger scales.

3.6.2 Model Selection

We compare three ensemble-based regression algorithms renowned for capturing complex, non-linear patterns:

1. **Random Forest Regressor** Implemented via `sklearn.ensemble.RandomForestRegressor`. Aggregates predictions from multiple decision trees to reduce overfitting.
2. **LightGBM Regressor** Using `lightgbm.LGBMRegressor`. A gradient-boosting framework optimized for speed and memory usage, particularly effective on large tabular datasets.
3. **Gradient Boosting Regressor** Via `sklearn.ensemble.GradientBoostingRegressor`. Builds an additive model in a forward stage-wise fashion to minimize prediction error.

3.6.3 Model Development Workflow

1. **Train/Test Split:** The full dataset was randomly partitioned into a training subset (70%) for model fitting and a hold-out test subset (30%) for final evaluation, ensuring no data leakage.
2. **Hyperparameter Optimization:** We conducted exhaustive grid searches using `GridSearchCV` with 3-fold cross-validation on the training set, optimizing for the highest mean R^2 score. This process tuned parameters such as the number of trees, maximum depth, and learning rate.

3. **Performance Assessment:** Each tuned model was evaluated on the test set, with performance quantified by:

    - *Mean Squared Error (MSE)*: Average squared difference between predicted and actual `DEA_BCC_Score`.
    - R^2 *(Coefficient of Determination)*: Proportion of variance in the efficiency scores explained by the model.

4. **Variable Importance Analysis:** From the best-performing regressor, we extracted feature importance values to rank the contextual predictors by their relative contribution to explaining efficiency variance.

3.7 Software and Libraries

The analysis was implemented in Python, leveraging the following major packages:

- `pandas`, `numpy`: Data manipulation and numerical operations.
- `scipy.optimize.linprog`: Solving the linear programs underlying the DEA models.
- `scikit-learn`: Preprocessing (encoding, imputation, scaling), clustering (K-Means), and machine-learning algorithms (Random Forest, Gradient Boosting, model validation tools).
- `lightgbm`: High-performance gradient boosting implementation.
- `joblib`, `psutil`: Parallel computation of DEA scores to accelerate processing of multiple DMUs.
- `matplotlib`, `seaborn`: Visualization of results and diagnostic plots.

Chapter 4

RESULTS AND DISCUSSION

This chapter presents the empirical results derived from the application of the methodology outlined in Chapter 3. The findings are structured according to the analytical stages: primary DEA results, sensitivity analysis, clustering outcomes, DEA-within-cluster analysis, and machine learning model insights.

4.1 DESCRIPTIVE STATISTICS OF KEY VARIABLES

After preprocessing, the dataset comprised 800 universities. **Table 4.1** summarizes the descriptive statistics for the primary DEA input, outputs, and key contextual variables. It summarizes the central tendencies and dispersion of our primary variables. The student–staff ratio averages around 19 with a wide spread (0.6 to 162.6), indicating diverse faculty intensities across institutions. Output indicators (teaching, research, citations, industry income, international outlook) exhibit means between 28 and 51, reflecting varied performance scales. Contextual factors show a large span in total enrollment (462 to 379,231) and internationalization (0% to 82%), underscoring the heterogeneity in size and global engagement.

4.2 PRIMARY DATA ENVELOPMENT ANALYSIS (DEA) RESULTS

Input-oriented CCR, BCC, and NIRS DEA models were applied to all 800 DMUs.

Table 4.1: Descriptive Statistics of Key Variables

Variable	Mean	Std Dev	Min	Max
stats_student_staff_ratio	19.09	12.50	0.60	162.60
scores_teaching	31.64	15.03	9.90	95.60
scores_research	28.25	19.58	2.90	99.00
scores_citations	51.31	27.05	1.20	100.00
scores_industry_income	44.96	22.43	0.00	100.00
scores_international_outlook	48.50	23.69	7.10	99.90
stats_number_students	24071.77	22494.25	462.00	379231.00
stats_female_male_ratio	0.49	0.13	0.00	1.00
stats_pc_intl_students	0.13	0.11	0.00	0.82

4.2.1 Efficiency Score Distributions

Table 4.2 presents the descriptive statistics for the calculated DEA scores.

Table 4.2: Descriptive Statistics of DEA Scores (CCR, BCC, NIRS, SE)

Score	Count	Mean	Std Dev	Min	25%	50%	75%	Max
DEA_CCR	800	0.1077	0.1078	0.0052	0.0478	0.0767	0.1195	1.0000
DEA_BCC	800	0.1892	0.2150	0.0055	0.0626	0.1185	0.2157	1.0000
DEA_NIRS	800	0.1890	0.2151	0.0052	0.0622	0.1185	0.2157	1.0000
Scale Eff.	800	0.6900	0.1991	0.1016	0.5558	0.6724	0.8630	1.0000

The DEA models reveal low average efficiency: the mean CCR score is only 0.108, and pure technical efficiency (BCC) averages 0.189 (Table 4.2). This suggests most universities operate substantially below the DEA frontier. Figure 4.4 illustrates that only a handful of institutions attain scores near 1.0, while the bulk cluster below 0.2, indicating room for improvement in resource utilization.

4.2.2 Returns to Scale (RTS)

The RTS analysis results are summarized in **Table 4.3** and visualized in **Fig. 4.5**.

Scale analysis (Table 4.3) shows 86.75% of universities under Increasing Returns to Scale (IRS), implying that enlarging size tends to improve efficiency. Only 13.00% exhibit Decreasing Returns (DRS), and a negligible 0.25% demonstrate Constant Returns (CRS). The bar chart in Figure 4.5 emphasizes the dominance of IRS, suggesting that many

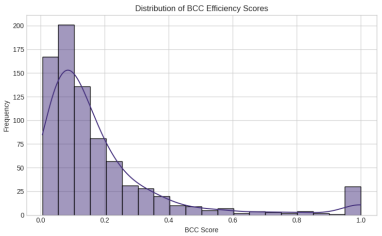


Figure 4.1: (a) BCC Scores

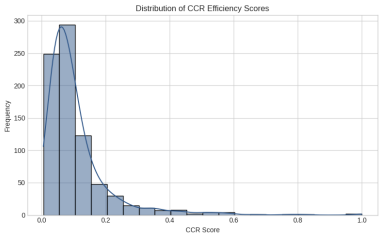


Figure 4.2: (b) CCR Scores

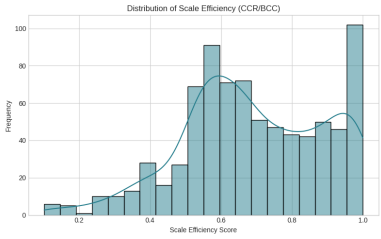


Figure 4.3: (c) Scale Efficiency

Figure 4.4: Distributions of DEA Scores

Table 4.3: Distribution of Returns to Scale

Returns to Scale Type	Percentage (%)
Increasing (IRS)	86.75
Decreasing (DRS)	13.00
Constant (CRS)	0.25

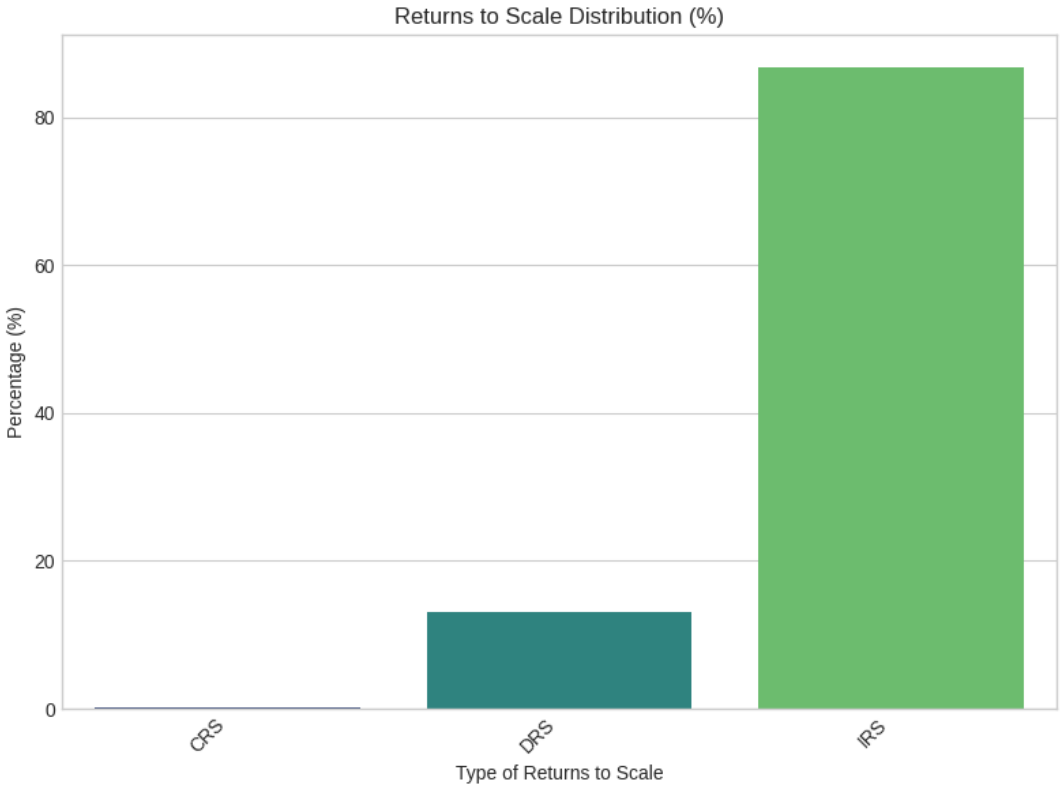


Figure 4.5: Returns to Scale Distribution (%)

institutions may benefit from strategic expansion.

4.2.3 Peer (Benchmark) Analysis

For selected inefficient DMUs (BCC $Score < 0.5$):

- University of California, Berkeley (BCC Score: 0.405) had California Institute of Technology (Lambda: 0.875) and Duke University (Lambda: 0.125) as peers.
- University of California, Los Angeles (BCC Score: 0.409) had California Institute of Technology (Lambda: 0.179), Johns Hopkins University (Lambda: 0.804), and Duke University (Lambda: 0.017) as peers.
- Cornell University (BCC Score: 0.356) had Johns Hopkins University (Lambda: 0.966) and Yale University (Lambda: 0.034) as peers.

4.2.4 Geographical Variations in Efficiency

Fig. 4.6 illustrates the BCC technical efficiency scores distributions for the top 7 countries. Noticeable variations emerge: some nations, like the United Kingdom and Australia, show higher median efficiency, while others display wider dispersion, highlighting differing national policies and resources.

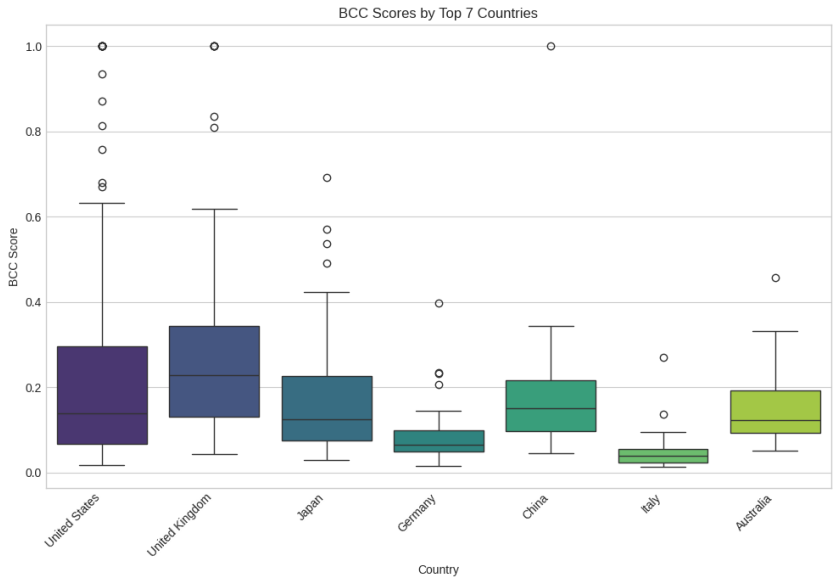


Figure 4.6: Box Plot of BCC Scores by Top 7 Countries

4.3 DEA SENSITIVITY ANALYSIS

A sensitivity analysis ("`Scenario_FewerOutputs`") used only `scores_teaching`, `scores_research`, and `scores_citations` as outputs. **Table 4.4** compares key metrics.

Table 4.4: DEA Sensitivity Analysis Results Summary

Metric	Original Model	Scenario_FewerOutputs
Mean BCC Score	0.1892	0.1095
Mean CCR Score	0.1077	0.0915
Mean Scale Efficiency	0.6900	0.9042
% IRS	86.75%	22.75%
% DRS	13.00%	76.88%
% CRS	0.25%	0.38%
Spearman Corr. (BCC)	-	0.8096

By restricting outputs to teaching, research, and citations, mean BCC efficiency falls to 0.1095, and CCR to 0.0915 (Table 4.4). Scale efficiency increases to 0.9042, and DRS becomes prevalent (76.88%), reversing the original pattern. A Spearman correlation of 0.8096 between the two BCC score sets confirms that efficiency rankings remain largely stable but are sensitive to the choice of outputs.The RTS distribution changed significantly, with DRS becoming dominant.

4.4 CLUSTERING RESULTS (K=2)

K-Means clustering used `location_Encoded`, `stats_number_students`, `stats_female_male_ratio`, and `stats_pc_intl_students`.

4.4.1 Optimal Number of Clusters

Fig. 4.9 Shows the Elbow method and Silhouette scores.

Silhouette analysis suggested K=2 as optimal.

4.4.2 Cluster Sizes and Profiles

Clustering with K=2 resulted in: Cluster 0: 793 DMUs; Cluster 1: 7 DMUs. **Table 4.5** details the profiles. Cluster 0 (793 universities) comprises the majority, with moderate contextual values and mean BCC of 0.1905. Cluster 1 (7 mega-scale institutions) shows

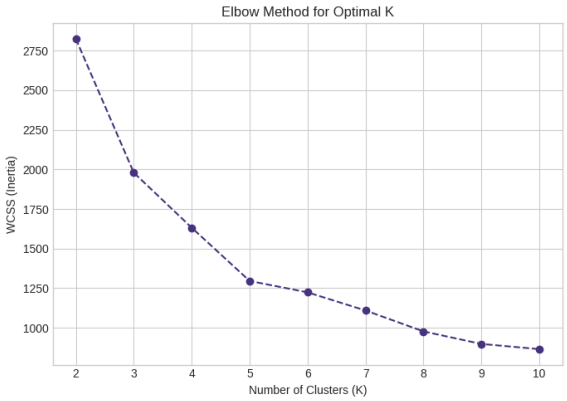


Figure 4.7: (a) Elbow Method

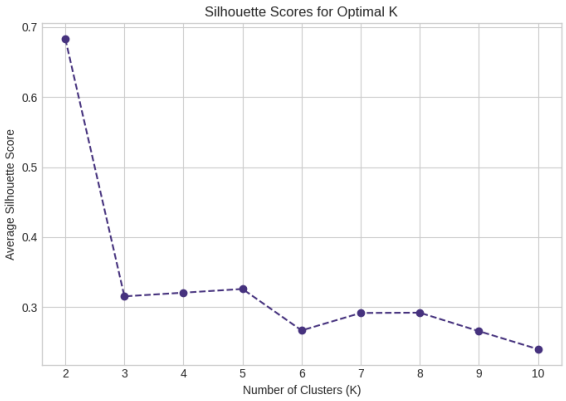


Figure 4.8: (b) Silhouette Scores

Figure 4.9: Optimal K Determination

dramatically higher enrollment (186,000 students) but lower original BCC (0.0442), suggesting scale-driven inefficiencies (Table 4.5). Both clusters predominantly operate under IRS. ANOVA for `DEA_BCC_Score` across clusters showed a P-value of 0.07296.

Table 4.5: University Cluster Profiles (K=2) - Mean Values (Mode for RTS)

Variable	Cluster 0 (n=793)	Cluster 1 (n=7)
DEA_BCC_Score	0.1905	0.0442
DEA_CCR_Score	0.1084	0.0388
Scale_Efficiency	0.6884	0.8736
location_Encoded	41.06	31.00
scores_citations	51.59	19.30
scores_industry_income	44.96	44.81
scores_international_outlook	48.67	29.49
scores_research	28.32	19.96
scores_teaching	31.71	23.79
stats_female_male_ratio	0.4933	0.5829
stats_number_students	22641.84	186063.14
stats_pc_intl_students	0.1277	0.0386
stats_student_staff_ratio	18.73	59.36
Returns_to_Scale_Mode	IRS	IRS

4.4.3 DEA-within-Clusters (K=2)

BCC DEA was run within each cluster. Running BCC DEA within each cluster yields a mean within-cluster efficiency of 0.1941 (Table 4.6), slightly above the overall aver-

age. Figure 4.10 and Figure 4.11 illustrate that most institutions improve their relative standing when compared only to peers of similar context.**Table 4.6** provides descriptive statistics. The mean `DEA_BCC_Score_Within_Cluster` was 0.1941.

Table 4.6: Descriptive Statistics of Within-Cluster BCC Scores (Overall Sample)

Statistic	Value
Count	800.000000
Mean	0.194095
Std	0.220496
Min	0.014315
25%	0.063403
50%	0.122550
75%	0.219844
Max	1.000000

Fig. 4.10 and **Fig. 4.11** visualize these scores.

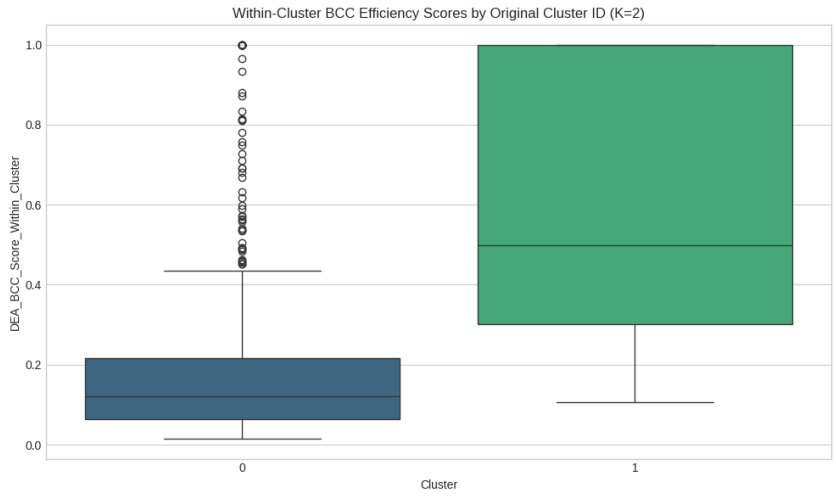


Figure 4.10: Box Plot of Within-Cluster BCC Efficiency Scores by Original Cluster ID

4.5 MACHINE LEARNING RESULTS (EXPLAINING EFFICIENCY)

Regression models predicted `DEA_BCC_Score` using contextual features.

4.5.1 Regression Model Performance

Table 4.7 summarizes test set performance. It shows that LightGBM achieves the best predictive power ($R^2 = 0.4725$), explaining nearly half the variance in DEA-BCC scores.

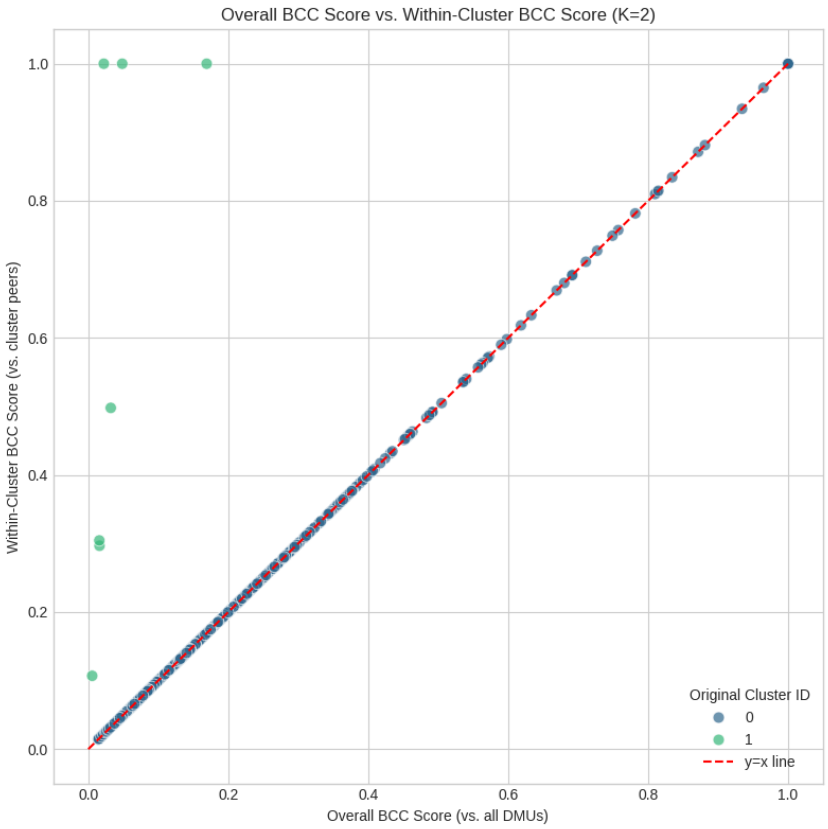


Figure 4.11: Scatter Plot of Overall BCC Score vs. Within-Cluster BCC Score

This indicates a moderate ability to estimate efficiency from contextual variables alone.

Table 4.7: Performance of Regression Models on Test Set

Model	Test R^2	Test MSE
Random Forest	0.4594	0.0297
LightGBM	0.4725	0.0290
Gradient Boosting (sklearn)	0.4673	0.0293

4.5.2 Feature Importances (LightGBM Regressor)

Fig. 4.12 displays feature importances from LightGBM.

The most important features were *stats_number_students*, *location_Encoded*, *stats_pc_intl_students*, and *stats_female_male_ratio*. Figure 4.12 highlights that total enrollment and country encoding are the top predictors, followed by international student share and gender ratio. This suggests that scale and geographic context are primary drivers of technical efficiency.

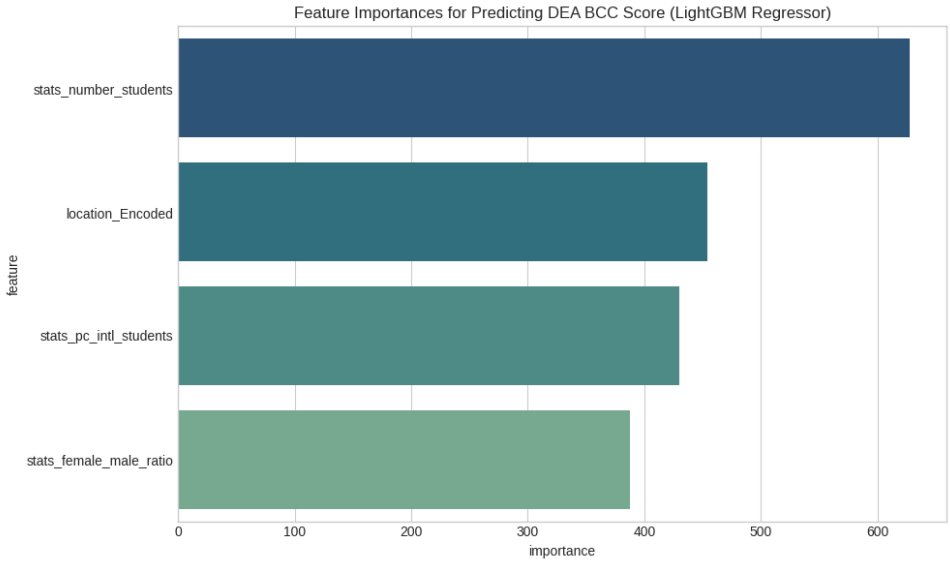


Figure 4.12: Feature Importances for Predicting DEA BCC Score (LightGBM Regressor)

The results underscore the generally low technical efficiency of universities globally, with strong evidence that larger institutions benefit from economies of scale. However, sensitivity checks reveal that efficiency interpretations can shift when focusing on core academic outputs. Clustering demonstrates that contextual segmentation yields more equitable comparisons, and subsequent within-cluster DEA slightly elevates average efficiency scores. Finally, machine learning models confirm that key contextual factors—particularly size and location—substantially influence efficiency, offering a predictive framework for policymakers and institutional leaders.

In the next chapter, we draw conclusions, discuss policy implications, and outline directions for future research.

Chapter 5

CONCLUSION AND FUTURE SCOPE

This chapter summarizes the key findings of the research, draws conclusions based on the analyses performed, discusses the limitations of the study, and outlines potential directions for future research. It also reflects on the broader social impact and practical implications of the work.

5.1 CONCLUSION

This thesis successfully applied an integrated multi-stage framework, combining Data Envelopment Analysis (DEA) with clustering and machine learning, to assess the efficiency of 800 global universities from the 2016 rankings.

The primary DEA revealed an average technical (BCC) efficiency of 0.189, highlighting significant potential for improvement. A key finding was the prevalence of Increasing Returns to Scale (IRS) across 86.75% of institutions, suggesting most were operating below their optimal scale given the broad range of outputs considered. Sensitivity analysis confirmed the robustness of relative rankings (Spearman correlation ˜0.81) despite changes in absolute efficiency scores and RTS distributions under different output specifications.

K-Means clustering (K=2) effectively segmented the universities into a large mainstream group and a very small, distinct group of mega-scale institutions with lower overall efficiency but higher scale efficiency. Contextualized DEA-within-clusters demonstrated that universities, especially those in the unique smaller cluster, achieve considerably higher relative efficiency when benchmarked against more homogenous peers, underscoring the importance of contextualized comparisons. Machine learning models, particularly LightGBM (test $R^2 \approx 0.4725$), identified `stats_number_students`, `location_Encoded`, and

stats_pc_intl_students as significant predictors of technical efficiency. This indicates that institutional size, geographical context, and internationalization are key factors associated with variations in university efficiency.

The multi-stage methodology employed provides a more nuanced and comprehensive understanding of university performance than standalone approaches. It not only quantifies relative efficiencies and scale characteristics but also sheds light on important contextual drivers, offering a robust basis for strategic planning and policy interventions in the higher education sector.

5.2 LIMITATIONS OF THE STUDY

Several limitations should be acknowledged:

- **Data Scope and Variables:** The analysis is based on a single year (2016), precluding dynamic efficiency analysis. The variables are from a public rankings dataset, which may not capture all nuanced inputs (e.g., granular financial data) or outputs (e.g., long-term graduate impact, direct teaching quality measures). The "scores" used as outputs are themselves composite indicators.

- **DEA Model Assumptions:** DEA is deterministic and attributes all deviations from the frontier to inefficiency, without accounting for statistical noise. The choice of input orientation and specific models (CCR, BCC, NIRS) carry inherent assumptions. Sensitivity analysis highlighted how output definitions impact RTS results.

- **Clustering Limitations:** K-Means has assumptions (e.g., spherical clusters), and the optimal K, while guided by Silhouette analysis, remains a choice. The identified K=2 resulted in highly imbalanced cluster sizes (793 vs. 7), suggesting one group is very distinct or that chosen features strongly separate this small group.

- **Machine Learning Model Limitations:** The R^2 of ˜0.47 means a portion of efficiency variance remains unexplained by the selected contextual variables. Feature importances indicate association, not causation.

- **Generalizability:** Findings are based on universities included in the 2016 rankings, which may not represent all HEIs globally.

5.3 FUTURE SCOPE

This research opens several avenues for future work:

- **Longitudinal Analysis:** Our current study uses single-year data, limiting insights into efficiency trends. Future work could use panel data DEA (e.g., Malmquist Productivity Index, Window DEA) to analyze efficiency evolution over time, revealing how universities adapt to changes in policy or strategy.

- **Alternative DEA Formulations:** Beyond standard CCR, BCC, and NIRS models, exploring advanced formulations like the Slacks-Based Measure (SBM) could offer more precise assessments by accounting for non-radial inefficiencies. Network DEA could also be valuable for analyzing universities as multi-stage systems (e.g., teaching and research processes).

- **Expanded Variable Sets:** The current analysis uses public ranking data. Future studies should integrate more granular data, including detailed financial inputs, qualitative teaching indicators, and diverse research outputs (e.g., patents, industry collaborations) for a more comprehensive evaluation.

- **Advanced Machine Learning and Econometric Techniques:** While Random Forest, LightGBM, and Gradient Boosting provided a strong basis, future research could explore deep learning or Bayesian techniques for more complex relationships. Incorporating causal inference methods (e.g., propensity score matching) or econometric models like Tobit regression would also help establish stronger causal links between contextual variables and efficiency outcomes.

- **Qualitative Case Studies:** Complementing quantitative findings with qualitative case studies of selected universities (efficient and inefficient) could reveal organizational practices, leadership strategies, and institutional cultures contributing to performance. Such insights are crucial for translating analytical results into actionable recommendations.

- **Sub-group Analysis and Local Contexts:** Applying the framework to specific university sub-groups (e.g., by country, region, or institutional type) would provide context-specific insights. This stratified analysis would enable policymakers to tailor improvement strategies that align with local challenges and priorities.

5.4 SOCIAL IMPACT

The insights from this thesis can contribute to several positive social impacts:

- **Enhanced University Management:** Provides objective self-assessment tools and benchmarks, enabling data-driven strategic planning and resource allocation for university leaders.
- **Informed Higher Education Policy:** Offers critical insights on scale efficiency and key efficiency drivers to inform policymakers' decisions regarding funding models, expansion, and strategies for a more effective education sector.
- **Improved Accountability and Resource Utilization:** Enhances transparency in fund utilization, allowing stakeholders to understand how resources generate outcomes, leading to more responsible allocation and greater public trust.
- **Fairer Performance Evaluation:** Promotes equitable comparisons by accounting for diverse institutional contexts, moving beyond simple rankings to highlight true operational effectiveness.
- **Contribution to Economic and Societal Development:** Fosters more efficient universities, which are better positioned to produce high-quality graduates, conduct impactful research, drive innovation, and address societal challenges, thereby boosting national competitiveness.

By offering a comprehensive and nuanced approach to evaluating university performance, this research aims to support ongoing efforts to strengthen the global higher education sector and its contributions to society.

Bibliography

[1] A. Charnes, W. W. Cooper, and E. L. Rhodes, "Measuring the efficiency of decision making units," *European Journal of Operational Research*, vol. 2, no. 6, pp. 429–444, 1978.

[2] R. D. Banker, A. Charnes, and W. W. Cooper, "Some models for estimating technical and scale inefficiencies in data envelopment analysis," *Management Science*, vol. 30, no. 9, pp. 1078–1092, 1984.

[3] W. W. Cooper, L. M. Seiford, and K. Tone, *Data envelopment analysis: A comprehensive text with models, applications, references and DEA-solver software*, 2nd ed. Springer Science & Business Media, 2007.

[4] J. Johnes, "Data envelopment analysis and its application to the measurement of efficiency in higher education," *Economics of Education Review*, vol. 25, no. 3, pp. 273–288, 2006.

[5] N. K. Avkiran, "Investigating technical and scale efficiencies of australian universities through data envelopment analysis," *Socio-Economic Planning Sciences*, vol. 35, no. 1, pp. 57–80, 2001.

[6] L. Simar and P. W. Wilson, "Estimation and inference in two-stage, semi-parametric models of production processes," *Journal of Econometrics*, vol. 136, no. 1, pp. 31–64, 2007.

[7] E. Thanassoulis, *Introduction to the theory and application of data envelopment analysis: A foundation text with integrated software.* Kluwer Academic Publishers, 2001.

[8] T. J. Coelli, D. S. P. Rao, C. J. O'Donnell, and G. E. Battese, *An introduction to efficiency and productivity analysis*, 2nd ed. Springer, 2005.

[9] J. Wolszczak-Derlacz and A. Parteka, "Efficiency of european public higher education institutions: A two-stage multicountry approach," *Scientometrics*, vol. 89, no. 3, pp. 887–917, 2011.

[10] K. De Witte and L. López-Torres, "Efficiency in education: A review of literature and a way forward," *Journal of the Operational Research Society*, vol. 68, no. 4, pp. 339–363, 2017.

[11] E. Thanassoulis, M. Kortelainen, G. Johnes, and J. Johnes, "Costs and efficiency of higher education: a review of the evidence," *Oxford Review of Economic Policy*, vol. 27, no. 4, pp. 598–621, 2011.

[12] A. Emrouznejad and K. De Witte, "Cooper-framework: A unified platform for dea-based literature search," in *Working Paper Series, COWPER COLES (Formerly Aston Business School), No. RP1003*, 2010.

[13] O. S. Olanrewaju, M. A. Hossain, N. Whiteside, and P. Mercieca, "Application of data envelopment analysis and machine learning in healthcare: A systematic review," *Annals of Operations Research*, 2021.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

THESIS (1).pdf

Delhi Technological University

Document Details

Submission ID
trn:oid:::27535:97238943

Submission Date
May 22, 2025, 9:51 PM GMT+5:30

Download Date
May 22, 2025, 9:52 PM GMT+5:30

File Name
THESIS (1).pdf

File Size
547.3 KB

39 Pages

7,518 Words

44,193 Characters

14% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 10 words)

Match Groups

53 Not Cited or Quoted 14%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

10% Internet sources
5% Publications
12% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.