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.

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)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	de.overleaf.com 2%
2	Submitted works	IIT Delhi on 2019-05-28 <1%
3	Internet	dspace.dtu.ac.in:8080 <1%
4	Internet	arxiv.org <1%
5	Submitted works	Associatie K.U.Leuven on 2025-05-21 <1%
6	Internet	www.igi-global.com <1%
7	Submitted works	Visvesvaraya National Institute of Technology on 2021-07-06 <1%
8	Publication	Al Brashdi, Ahmed Yaqoob Mohammed. "Assessing the Resource Utilization and E... <1%
9	Publication	Dalmo Marchetti, Peter Wanke. "Brazil's rail freight transport: Efficiency analysis ... <1%
10	Submitted works	De Montfort University on 2023-01-29 <1%

11 Internet

pgsdspace.ictp.it <1%

12 Submitted works

University of Edinburgh on 2024-08-21 <1%

13 Internet

123dok.net <1%

14 Submitted works

Universiteit Utrecht on 2025-01-30 <1%

15 Submitted works

Abant İzzet Baysal Universitesi on 2017-12-19 <1%

16 Submitted works

Cranfield University on 2024-09-02 <1%

17 Internet

research.shahed.ac.ir <1%

18 Internet

trepo.tuni.fi <1%

19 Submitted works

University College London on 2024-08-29 <1%

20 Submitted works

Associatie K.U.Leuven on 2010-03-31 <1%

21 Submitted works

Colorado Technical University Online on 2025-05-19 <1%

22 Publication

Elsayed, Rayan. "Predicting Newborn Low Birth Weight: A Machine Learning Appr... <1%

23 Submitted works

Universiti Putra Malaysia on 2011-02-14 <1%

24 Internet

wrap.warwick.ac.uk <1%

25 Internet
www.coursehero.com <1%

26 Internet
www.frontiersin.org <1%

27 Submitted works
Associatie K.U.Leuven on 2025-05-21 <1%

28 Publication
Z. Yang, J.C. Paradi. "DEA Evaluation of a Y2K Software Retrofit Program", IEEE Tra... <1%

29 Internet
dspace.daffodilvarsity.edu.bd:8080 <1%

30 Internet
research-portal.uu.nl <1%

31 Publication
Azad, A.S.M. Sohel, Suzuki Yasushi, Victor Fang, and Amirul Ahsan. "Impact of poli... <1%

32 Publication
Derek D. Wang. "Performance-based resource allocation for higher education ins... <1%

33 Publication
Md. Ferdous Alam, Khondker Murshed-e-Jahan. "RESOURCE ALLOCATION EFFICIE... <1%

34 Submitted works
University of Leicester on 2010-08-27 <1%

35 Internet
cuir.car.chula.ac.th <1%

36 Internet
dlibrary.univ-boumerdes.dz:8080 <1%

37 Internet
www.sunshinecoastnews.com.au <1%

38 Internet
5dok.org <1%

39 Submitted works

Liverpool John Moores University on 2024-09-11 <1%

40 Submitted works

University of Edinburgh on 2023-08-24 <1%

41 Submitted works

University of Stellenbosch, South Africa on 2016-01-10 <1%

42 Submitted works

Wageningen University on 2024-12-17 <1%

43 Internet

businessdocbox.com <1%

44 Internet

www.cerem-review.eu <1%

45 Internet

www.researchgate.net <1%

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

A PROJECT REPORT

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

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

MAY, 2025

DEPARTMENT OF MECHANICAL ENGINEERING
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 Bachelor of Technology, 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: 25.05.2025 23/MSCMAT/67

DEPARTMENT OF MECHANICAL ENGINEERING
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 Bachelor of Technology, 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: 25.05.2025

SUPERVISOR

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

ACKNOWLEDGEMENT

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

Place: Delhi Mehak Goyal

Date: 25.05.2025 23/MSCMAT/67

Abstract

This thesis investigates the efficiency of global universities using an integrated Data Envelopment Analysis (DEA) and machine learning framework. Based on data for 800 universities from the 2016 rankings, this study employs input-oriented CCR, BCC, and NIRS DEA models to assess the technical, scale, and overall efficiencies, using the student-staff ratio as input, and scores for teaching, research, citations, industry income, and international outlook 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) offers a robust, non-parametric approach to measure the relative efficiency of a set of comparable Decision-Making Units (DMUs)—in this context, universities. Developed by Charnes, Cooper, and Rhodes (1978), DEA constructs an "efficient frontier" based on the observed best-performing DMUs and evaluates the efficiency of others relative to this frontier. It does so without requiring pre-specified weights for inputs and outputs or making assumptions about the underlying functional form of the production process.

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.

5 Foundational DEA models, including the CCR model, which assumes Constant Returns to Scale (CRS), and the BCC model, which assumes Variable Returns to Scale (VRS), allow for assessing overall efficiency, pure technical efficiency, and scale efficiency. These models have been widely applied to benchmark universities across different regions and contexts, providing valuable insights into their relative performance.

45 Despite its utility, DEA is not without its limitations. Efficiency scores can be sensitive to the choice of variables and the sample of DMUs. Furthermore, DEA is a deterministic method that attributes any deviation from the frontier to inefficiency without explicitly accounting for statistical noise or the impact of external contextual factors that are not part of the input-output specification.

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

24

The primary objective of this thesis is to conduct a comprehensive analysis of global university efficiency by integrating Data Envelopment Analysis (DEA) with clustering 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 reviews the foundational concepts of Data Envelopment Analysis (DEA), its application in assessing the efficiency of higher education institutions (HEIs), and the emerging role of machine learning techniques in enhancing DEA-based performance evaluations.

2.1 DATA ENVELOPMENT ANALYSIS (DEA) FUNDAMENTALS

Data Envelopment Analysis (DEA), first introduced by Charnes, Cooper, and Rhodes in 1978, is a non-parametric mathematical programming technique used for measuring the relative efficiency of a collection of Decision Making Units (DMUs) that consume multiple inputs to produce multiple outputs. It is particularly useful for evaluating non-profit organizations, such as universities, where market prices for inputs and outputs are often unavailable or inappropriate. DEA constructs an empirical "best-practice" frontier based on the observed performance of DMUs and calculates efficiency scores for other DMUs relative to this frontier.

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 change in inputs will result in an equi-proportional change in outputs. The CCR model provides a measure of overall efficiency, which integrates both technical efficiency (managerial ability to use

inputs) and scale efficiency (operating at the optimal size). DMUs with a CCR score of 1 are considered globally efficient, lying on the CRS frontier.

2.1.2 BCC Model

Recognizing that the CRS assumption might be too restrictive, Banker, Charnes, and Cooper (1984) developed the BCC model, which assumes Variable Returns to Scale (VRS). The BCC model allows the production frontier to exhibit increasing, constant, or decreasing returns to scale. It measures pure technical efficiency by comparing a DMU only to other DMUs of similar scale, thus isolating managerial efficiency from scale effects. A BCC score of 1 indicates that a DMU is technically efficient, irrespective of its size.

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. Scale Efficiency (SE) can be calculated 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). The NIRS model, as used in this thesis, 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 "by how much can input quantities be proportionally reduced without changing the output quantities produced?" 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, but machine learning (ML) models are gaining prominence due to their ability to handle non-linear relationships, complex interactions, and offer robust feature importance measures without stringent distributional assumptions. Tree-based ensemble methods like Random Forest and gradient boosting machines like LightGBM and XGBoost are particularly powerful. They can be used for both 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 details 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 study follows a quantitative, multi-stage analytical research design. The core objective is to provide a comprehensive assessment of university efficiency by systematically integrating DEA with clustering and supervised machine learning. The key stages are:

1. **Data Acquisition and Preprocessing:** Sourcing the university rankings data and preparing it for analysis through cleaning, transformation, and imputation.

2. **Primary DEA Modeling:** Applying input-oriented CCR, BCC, and NIRS models to the entire dataset to compute overall efficiency, technical efficiency, scale efficiency, and determine returns to scale. This stage also includes a sensitivity analysis of the DEA model.

3. **Clustering of DMUs:** Segmenting universities into more homogenous groups using K-Means clustering based on key contextual variables. The optimal number of clusters is determined using established statistical methods.

4. **Contextualized DEA Modeling:** Re-applying DEA models (specifically BCC) within each identified cluster to assess relative efficiency against more comparable

peers.

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 data underwent extensive preprocessing:

1. **Data Cleaning and Type Conversion:** Columns such as `stats_number_students` (commas removed), `stats_pc_intl_students` (percentage signs removed, converted to proportion), and `stats_female_male_ratio` (string "F:M" parsed to female proportion) were cleaned and converted to numeric types.
2. **Label Encoding:** The categorical `COUNTRY_COLUMN` ('location') was converted into a numerical representation (`location_Encoded`) using `LabelEncoder`.
3. **Handling Missing Values:**
   - DMUs with missing values in any of the essential DEA input or output columns (`essential_dea_cols`) were subject to `dropna`. In this specific run, 0 rows were dropped at this stage for the 800 DMUs.
   - Missing values in other numerical contextual columns (e.g., `stats_female_male_ratio` which had 52 missing values after conversion) were imputed using the median strategy with `SimpleImputer`).
4. **Ensuring Positivity for DEA:** DEA inputs and outputs must be positive. Variables like `scores_industry_income` that contained non-positive values were clamped to a small positive epsilon (1×10^{-6}).
5. **Index Reset:** The DataFrame index was reset after cleaning.

3.4 DEA MODELING

This study utilizes input-oriented DEA models, assuming universities aim to minimize resource consumption for given output levels. The calculations for each DMU were parallelized using `joblib` and `psutil` for efficiency.

3.4.1 CCR, BCC, and NIRS Models

The core DEA models implemented are:

- `dea_ccr_input_oriented`: Calculates overall efficiency assuming Constant Returns to Scale (CRS). The objective is to minimize θ (proportional input reduction) subject to $\sum \lambda_j X_{ij} \leq \theta X_{ik}$ for inputs and $\sum \lambda_j Y_{rj} \geq Y_{rk}$ for outputs.
- `dea_bcc_input_oriented`: Calculates pure technical efficiency assuming Variable Returns to Scale (VRS). It adds the convexity constraint $\sum \lambda_j = 1$ to the CCR formulation.
- `dea_nirs_input_oriented`: Calculates efficiency assuming Non-Increasing Returns to Scale. This is similar to BCC but uses the constraint $\sum \lambda_j \leq 1$.

All models use the 'highs' solver in `scipy.optimize.linprog`. Efficiency scores are bounded between 0 and 1.

3.4.2 Scale Efficiency (SE) and Returns to Scale (RTS) Determination

Scale Efficiency (SE) for each DMU is computed as:

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

Returns to Scale (RTS) are determined using the logic):

- CRS: If `DEA_NIRS_Score` $\approx$ `DEA_BCC_Score` $\approx$ `DEA_CCR_Score`.
- IRS: If `DEA_NIRS_Score` $\approx$ `DEA_BCC_Score` and `DEA_BCC_Score` $>$ `DEA_CCR_Score`.
- DRS: If `DEA_NIRS_Score` $<$ `DEA_BCC_Score`.

A tolerance of 1×10^{-5} is used for these floating-point comparisons.

3.4.3 DEA Sensitivity Analysis

To assess the robustness of DEA results, a sensitivity analysis was performed by running the DEA models with an alternative output specification: `output_cols_scenario2` = [`'scores_teaching'`, `'scores_research'`, `'scores_citations'`]. Efficiency scores, RTS distributions, and Spearman rank correlations between original and scenario BCC scores were compared.

3.4.4 Peer (Benchmark) Analysis

For inefficient DMUs (BCC score ¡ 0.5, for example), efficient peers are identified using the lambda (λ_j) values from the BCC model output. A DMU is considered a peer if its lambda value is greater than 1×10^{-5}.

3.5 CLUSTERING OF UNIVERSITIES

K-Means clustering was used to group universities based on contextual variables.

3.5.1 Feature Selection and Scaling for Clustering

The features used for clustering were `location_Encoded`, `stats_number_students`, `stats_female_male_` and `stats_pc_intl_students`. These features were imputed using median strategy and then standardized using `StandardScaler` before applying K-Means.

3.5.2 Determining the Optimal Number of Clusters (K)

26 The optimal K was determined by:

1. **Elbow Method:** Plotting Within-Cluster Sum of Squares (WCSS) against K (for K from 2 to 10).
2. **Silhouette Analysis:** Calculating the average Silhouette score for K from 2 to 10.

Based on the Silhouette analysis, K=2 was chosen.

3.5.3 Cluster Profiling

Once clusters were formed (`N_CLUSTERS_CHOSEN = 2`), they were profiled by calculating the mean values of numeric features (including original DEA scores and clustering variables) and the mode of 'Returns_to_Scale' for each cluster. ANOVA was used to test for statistical differences in means for `DEA_BCC_Score` across clusters.

3.5.4 DEA-within-Clusters

The input-oriented BCC DEA model was applied separately to DMUs within each of the K=2 clusters to obtain contextualized efficiency scores (`DEA_BCC_Score_Within_Cluster`). These were then compared to the overall BCC scores.

3.6 SECOND-STAGE MACHINE LEARNING ANALYSIS

Supervised machine learning models were developed to explain the overall `DEA_BCC_Score` using contextual variables as predictors.

3.6.1 Target and Explanatory Variables

- **Target Variable (Regression):** Continuous `DEA_BCC_Score`.
- **Explanatory Variables (Features):** `location_Encoded`, `stats_female_male_ratio`, `stats_number_students`, `stats_pc_intl_students`. These features were scaled using `StandardScaler`.

3.6.2 Machine Learning Models

The following regression models were implemented and tuned:

1. **Random Forest Regressor** (`sklearn.ensemble.RandomForestRegressor`).
2. **LightGBM Regressor** (`lightgbm.LGBMRegressor`).
3. **Gradient Boosting Regressor** (`sklearn.ensemble.GradientBoostingRegressor`).

3.6.3 Model Training, Tuning, and Evaluation

1. **Data Splitting:** Data was split into training (70%) and testing (30%) sets.
2. **Hyperparameter Tuning:** `GridSearchCV` with 3-fold cross-validation was used to find optimal hyperparameters for each model, using 'r2' as the scoring metric.
3. **Model Evaluation (Regression):** Performance on the test set was evaluated using Mean Squared Error (MSE) and R-squared (R^2).
4. **Feature Importance:** Feature importances were extracted from the best-performing regressor to identify key drivers of efficiency.

3.7 SOFTWARE AND LIBRARIES

The analysis was conducted in Python using several key libraries: `pandas` for data handling, `numpy` for numerical operations, `scipy.optimize.linprog` for DEA, `matplotlib` and `seaborn` for visualizations, `scikit-learn` for preprocessing, clustering, ML models and evaluation, `joblib` and `psutil` for parallel processing. `lightgbm` was used for the LightGBM model.

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.

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 PRIMARY DATA ENVELOPMENT ANALYSIS (DEA) RESULTS

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

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 mean overall efficiency (CCR score) was 0.1077, while the mean pure technical efficiency (BCC score) was 0.1892. The distributions of these scores are depicted in **Fig. 4.4**.

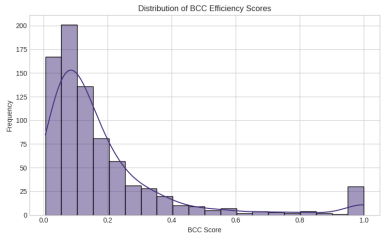

Figure 4.1: (a) BCC Scores

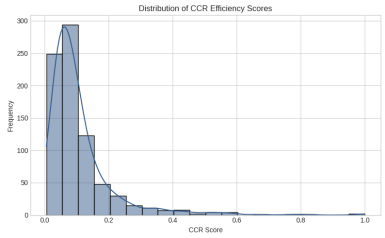

Figure 4.2: (b) CCR Scores

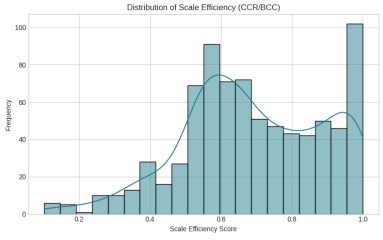

Figure 4.3: (c) Scale Efficiency

Figure 4.4: Distributions of DEA Scores

4.2.2 Returns to Scale (RTS)

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

A significant majority (86.75%) of universities exhibited Increasing Returns to Scale (IRS).

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

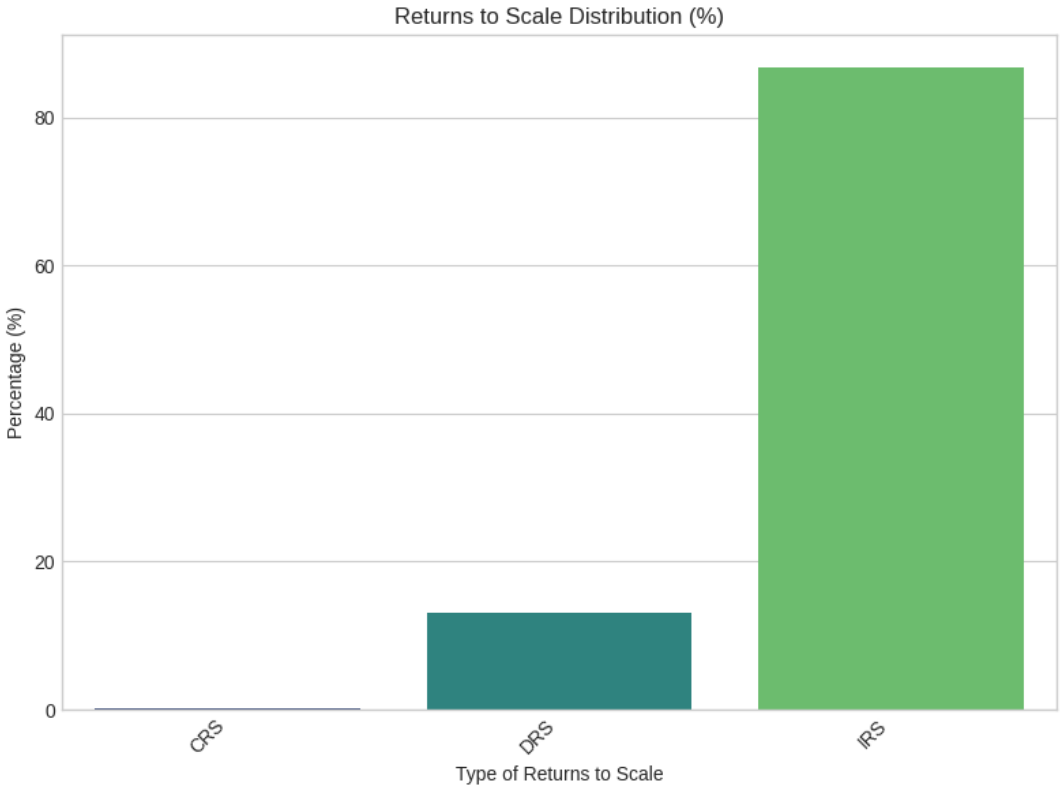

Figure 4.5: Returns to Scale Distribution (%)

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.

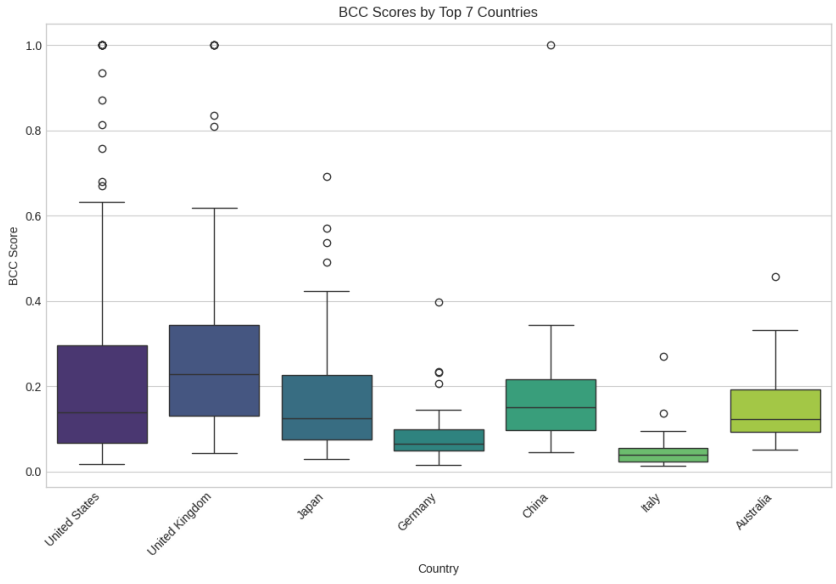

Figure 4.6: Box Plot of BCC Scores by Top 7 Countries

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

The RTS distribution changed significantly, with DRS becoming dominant. The Spearman correlation was 0.8096.

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.

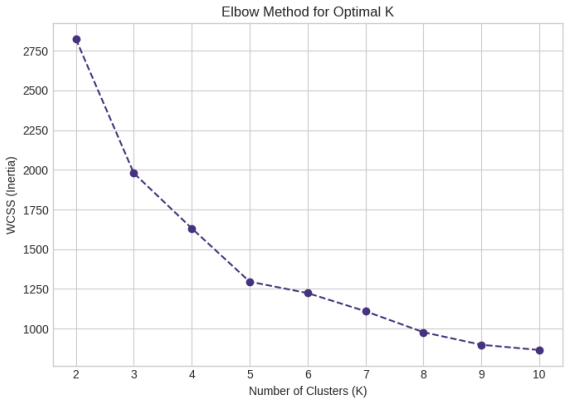

Figure 4.7: (a) Elbow Method

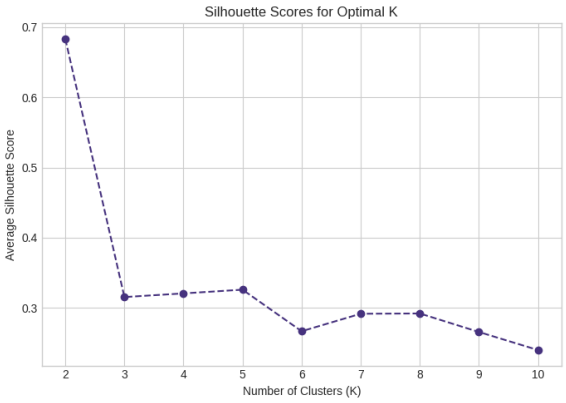

Figure 4.8: (b) Silhouette Scores

Figure 4.9: Optimal K Determination

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.

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

ANOVA for `DEA_BCC_Score` across clusters showed a P-value of 0.07296.

4.4.3 DEA-within-Clusters (K=2)

BCC DEA was run within each cluster. **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.

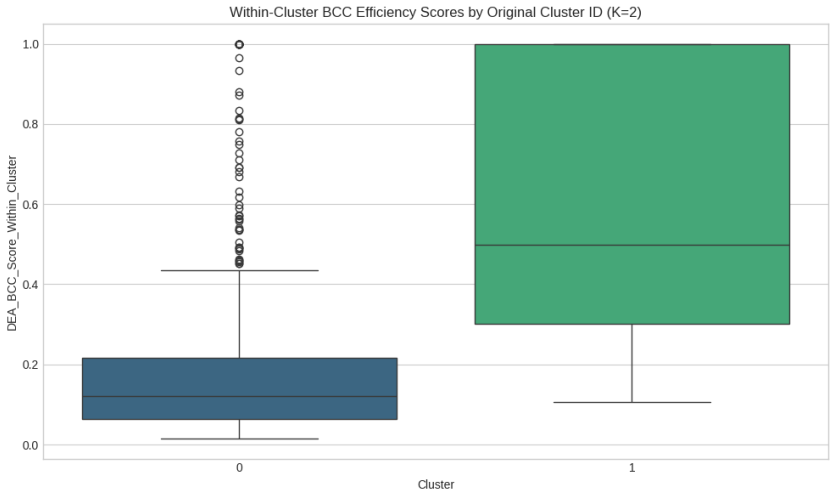

Figure 4.10: Box Plot of Within-Cluster BCC Efficiency Scores by Original Cluster ID

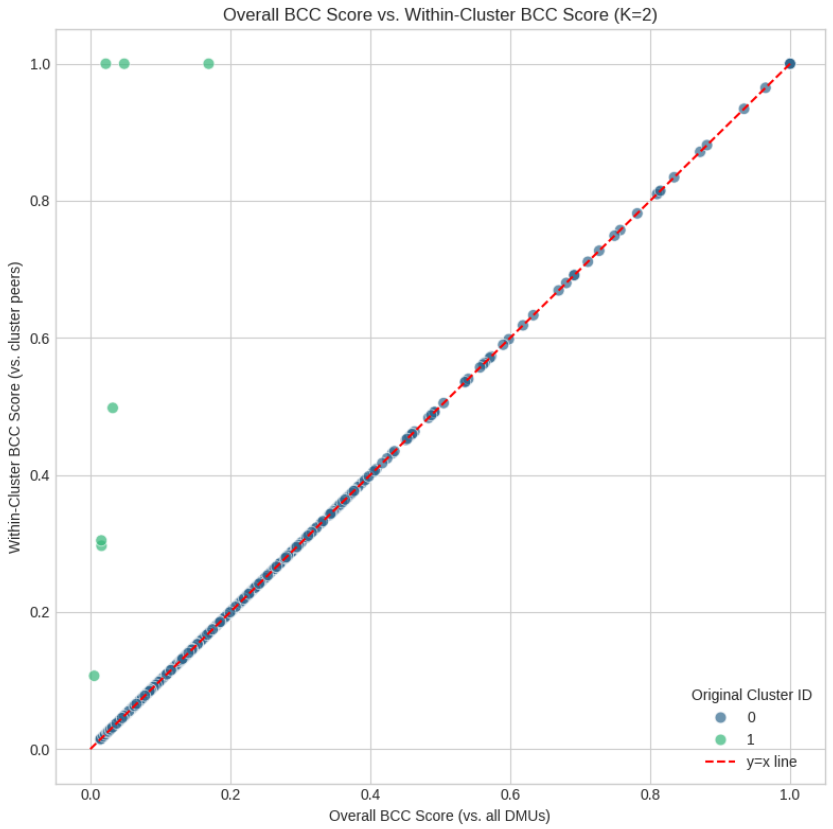

Figure 4.11: Scatter Plot of Overall BCC Score vs. Within-Cluster BCC Score

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.

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

LightGBM was the best regressor ($R^2 = 0.4725$).

4.5.2 Feature Importances (LightGBM Regressor)

Fig. 4.12 displays feature importances from LightGBM.

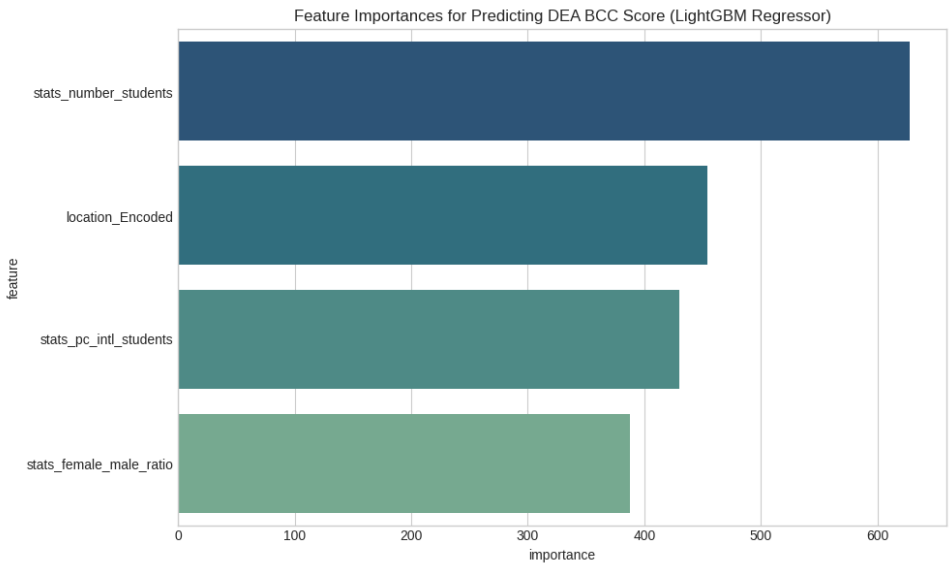

Figure 4.12: Feature Importances for Predicting DEA BCC Score (LightGBM Regressor)

The most important features were *stats_number_students*, *location_Encoded*, *stats_pc_intl_students*, and *stats_female_male_ratio*.

Chapter 5

CONCLUSION AND FUTURE SCOPE

18 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.