Prediction of NIRF Scores Using the Classification and Regression Tree (CART) Algorithm in Indian Engineering Institutions

Uzma Khanum[1], Ramya N. N[1], Nisarga H P[1], Ashin Krishna. C[1], Chaithra N[2]*

1. Research Scholar, Division of Medical Statistics, School of Life Sciences, JSS Academy of Higher Education and Research, Mysuru

2. Associate Professor, Division of Medical Statistics, School of Life Sciences, JSS Academy of Higher Education and Research, Mysuru

ABSTRACT

Background:
The National Institutional Ranking Framework (NIRF), introduced by the Ministry of Education, Government of India, in 2015, ranks higher education institutions across various disciplines, including engineering. It evaluates institutions based on five key parameters: Teaching, Learning & Resources (TLR), Research and Professional Practice (RPC), Graduation Outcomes (GO), Outreach and Inclusivity (OI), and Perception (PR). These rankings aim to promote transparency, accountability, and continuous quality improvement in the higher education sector.

Methodology:
A cross-sectional study was conducted using NIRF data from 799 engineering institutions in India, spanning the years 2016 to 2023. The Classification and Regression Tree (CART) algorithm was applied to predict total NIRF scores based on five predictor variables: TLR, RPC, GO, OI, and PR. Model performance was evaluated using classification metrics such as sensitivity, specificity, accuracy, Cohen's Kappa, and the ROC curve. Feature importance analysis was conducted to identify the most influential predictors. Statistical analysis was performed using IBM SPSS 22.0.

Results:
The Classification and Regression Tree (CART) model predicted Total Scores primarily using RPC and TLR, generating 13 nodes with 7 terminal nodes at a depth of 3. The model showed strong performance, achieving 89% accuracy, 88% sensitivity, and 90% specificity, with a risk estimate of 19.91%. A high AUC value of 0.95 indicates a strong ability to distinguish between high and low scores. Based on variable importance, RPC was identified as the most influential predictor, followed by PR and TLR.

Conclusion:
The CART algorithm provides an interpretable and data-driven approach to identifying critical performance indicators among engineering institutions. By highlighting the most impactful parameters, it can support evidence-based policy decisions and institutional strategies aimed at improving national ranking outcomes.

Keywords:
Engineering Institutions, NIRF Parameters, Predictive Modelling, CART algorithm, Performance Evaluation

INTRODUCTION

The National Institutional Ranking Framework (NIRF) was established by the Indian Ministry of Education on September 29, 2015, initially under the Ministry of Human Resource Development, which is now known as the Ministry of Education. This framework aims to rank higher education institutions across the nation [1]. In India, higher education encompasses various establishments, including colleges, universities, and institutes that offer undergraduate and graduate degree programs in fields such as arts, sciences, commerce, engineering, medicine, and law [2]. Ensuring institutional effectiveness in higher education involves the systematic collection, analysis, and application of data to support the institution's mission and goals, thus guaranteeing the quality of education [3].

Global rankings, like those provided by Quacquarelli Symonds (QS) and Times Higher Education (THE), are highly respected and serve as useful comparative tools for researchers, decision-makers, and students [4]. India, with its extensive higher education system comprising 42,343 colleges and 1,043 universities, ranks third globally among publicly funded university systems. The Ministry of Education has been striving to improve the Gross Enrolment Ratio and elevate the quality of education in the country [5]. Engineering education, a pivotal part of India's higher education landscape, has significantly evolved since the 19th century, especially following market liberalization, which accelerated technological advancements [6].

The NIRF assesses institutions based on five principal parameters: Teaching, Learning & Resources (TLR), Research and Professional Practices (RP), Graduation Outcomes (GO), Outreach and Inclusivity (OI), and Perception. Each parameter includes several sub-parameters for a thorough evaluation of institutional performance [7]. The decision tree is a popular predictive method for solving various data mining tasks, including classification, prediction, regression and estimation. It is also used for data description, visualization, and dimensionality reduction. As a non-parametric method, it employs inductive reasoning and supervised learning to model data without assuming a specific distribution [8]. the aim of the study is to identify the key Predictors for improving the Total Scores in NIRF.

METHODOLOGY

Study Design

The study was an annual observational cross-sectional analysis conducted on data from 799 higher education institutions in India, as part of the NIRF assessment from 2016 to 2023. Institutions were evaluated using multiple performance metrics, and these metrics were analyzed in relation to demographic factors such as institution type, geographic location, and student population size. This approach provides actionable insights for stakeholders and promotes continuous improvement in higher education.

Inclusion Criteria

Institutions were included if they has completed the NIRF survey and provided complete data on the performance indicators: Teaching, Learning & Resources (TLR), Research and Professional Practices (RPC), Graduation Outcomes (GO), Outreach and Inclusivity (OI), and Perception (PR). Complete demographic information, including institution type, location, and student enrolment, was also required.

Exclusion Criteria

Institutions were excluded if data on any of the key performance metrics (TLR, RPC, GO, OI, PR) were missing. Additionally, institutions providing incomplete or unreliable data or those that did not participate in the NIRF survey were excluded to ensure data accuracy and completeness.

Statistical Methods

The Classification and Regression Tree (CART) model was used to predict total NIRF scores based on the independent variables TLR, RPC, GO, OI, and PR. CART is a non-parametric method capable of capturing non-linear relationships by segmenting data into subgroups. Model performance was evaluated using a confusion matrix and metrics including sensitivity, specificity, and accuracy, which demonstrated moderate predictive capability. Feature importance analysis identified the most influential predictors. Node summaries and decision tree visualizations were used to illustrate relationships between variables and their impact on predicted scores. Statistical analyses were conducted using IBM SPSS version 22.0 and R Studio.

Classification and Regression Trees (CART) Algorithm

The Classification and Regression Trees (CART) algorithm is a widely used method for constructing decision trees to predict a target variable based on one or more input variables. CART performs two primary tasks: Classification, which assigns observations to discrete categories, and Regression, which estimates continuous numerical outcomes [9]. CART employs a binary tree structure for decision-making. Each node represents a decision point based on the value of an input attribute, where the dataset is tested and split. The edges or branches indicate the outcomes of these decisions, leading to subsequent nodes or terminal nodes [10]. The terminal nodes (leaves) represent the final prediction or classification outcome. At each node, CART evaluates all candidate attributes and selects the one that produces the most homogeneous subsets of data, thereby optimizing prediction accuracy [11].

For classification tasks, CART commonly uses the Gini index to assess the quality of splits.

The Gini index measures the probability of incorrectly classifying a randomly chosen observation and is calculated as:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

where pi is the probability of an observation belonging to class I

- n is the total number of classes.

The Gini index ranges from 0 to 1: a value of 0 indicates perfect purity (all elements belong to a single class), while a value of 1 indicates maximum impurity (elements are evenly distributed across classes). For an equally distributed classes, the Gini value is 1 - 1/n occurs when elements are uniformly distributed among n classes (e.g., for two classes, Gini = 1 - 1/2 = 0.5 [12]. In essence, the Gini impurity quantifies the likelihood of misclassification assuming random selection based on class probabilities.

The CART algorithm relies on several key assumptions:

1. Independence: Observations are assumed to be independent.
2. Homogeneity within Nodes: Data within each node or leaf is assumed to be homogeneous, meaning the target variable is similar for all observations in that node.
3. Binary Splits: Each node is divided into exactly two child nodes.
4. Recursive Partitioning: Data are recursively partitioned into increasingly homogeneous subsets.

5. Greedy Algorithm: The best split is chosen at each step without considering future splits.
6. Pruning: To prevent overfitting, branches that contribute minimally to predictive accuracy are pruned [13].

RESULTS AND DISCUSSION

The CART decision tree model was constructed to predict the total scores using the independent variables TLR, RPC, GO, OI, and PR. The model was built with a maximum tree depth of 3, a minimum of 50 cases in parent nodes, and 25 cases in child nodes, without applying any validation method. As shown in Tabsle 1, the final tree included RPC and TLR as the primary predictors, generating 13 nodes in total, of which 7 were terminal nodes, with an overall depth of 3. This structure indicates that RPC and TLR play a central role in predicting total scores, segmenting the dataset into 13 distinct groups and producing 7 refined terminal subgroups. The model thus reflects a moderate level of complexity, achieving a balance between predictive accuracy and interpretability.

Model Summary		
Specifications	**Growing Method**	**CART**
	Dependent Variable	Total scores
	Independent Variables	TLR, RPC, GO, OI, PR
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	50
	Minimum Cases in Child Node	25
Results	Independent Variables Included	RPC, TLR
	Number of Nodes	13
	Number of Terminal Nodes	7
	Depth	3

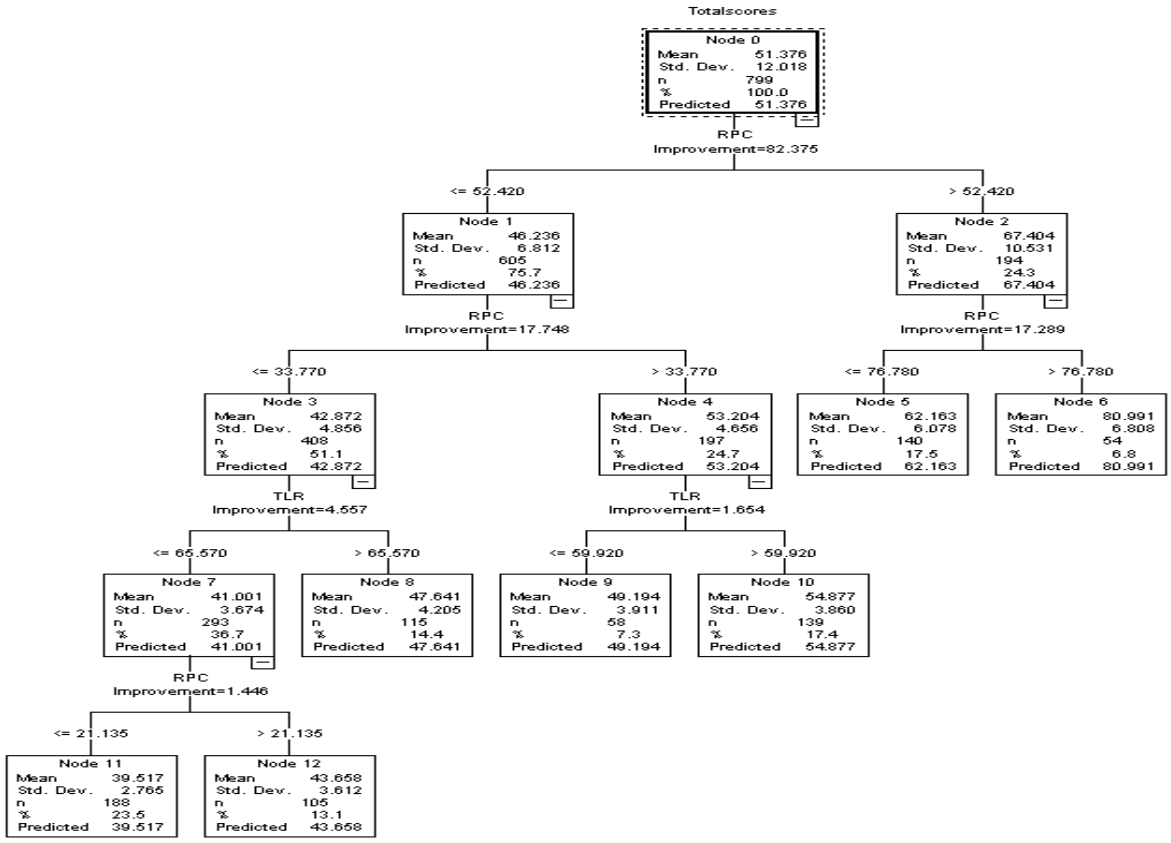

Figure 1: Decision Tree Model for Predicting Total Scores

Figure 1 illustrates the decision tree model, which predicts total scores based on successive splits of the predictor variables. The tree begins with the root node, encompassing all 799 observations, with a mean score of 51.376 and a standard deviation of 12.018. The first split occurs at a score of 52.420, resulting in Node 1, which contains 605 observations (75.7%) with a mean score of 46.236, and Node 2, which contains 194 observations (24.3%) with a mean score of 67.404. Node 1 further splits at 33.770, producing Node 3 (mean score 42.872) and Node 4 (mean score 53.204). Node 3 undergoes an additional split at 65.570, resulting in Node 7 (mean score 41.001) and Node 8 (mean score 47.641). Node 4 splits at 59.920, creating Node 9 (mean score 49.194) and Node 10 (mean score 54.877). Node 2 splits at 76.780, generating Node 5 (mean score 62.163) and Node 6 (mean score 80.991).

Table 2: Statistical Analysis of Risk Estimation

Risk	
Estimate	**Std. Error**
19.907	1.448
Growing Method: CART Dependent Variable: Total scores	

The table 2 represents CART decision tree model developed for predicting `Total scores` has a risk estimate of 19.907 with a standard error of 1.448. This risk estimate indicates that the model has an overall misclassification rate of approximately 19.907%. The standard error of 1.448 suggests that the precision of the risk estimate is fairly high, implying that the variability around this estimate is low. This indicates that while the model is moderately effective in predicting the `Total scores`, about 19.907% of the predictions are expected to be incorrect. The relatively low standard error provides confidence in the stability and reliability of the risk estimate, indicating that the model's performance is consistent across the dataset.

Table 3: Classification table

Observed	**Predicted**		
	High	**Low**	**Percent Correct**
High	70	9	88.60%
Low	8	72	90%
Overall percentage	49%	50.90%	89.3

The table 3 represents the confusion matrix which evaluates the classification performance of a predictive model on a binary outcome, with categories "High" and "Low." The model correctly identified 70 out of 79 actual "High" cases, yielding a sensitivity (true positive rate) of 88.6%, and correctly identified 72 out of 80 actual "Low" cases, resulting in a specificity (true negative rate) of 90%. The overall accuracy of the model is 89.3%, meaning that nearly 9 out of 10 predictions align with actual outcomes. The balanced performance across both classes, with a slight increase in accuracy for "Low" predictions, suggests that the model performs well in distinguishing between "High" and "Low" classes without significant bias toward either class. The near-even split of predictions across "High" (49%) and "Low" (50.9%) further indicates that the model is appropriately calibrated for balanced data.

Table 4: Prediction Performance Measures

Metrices	Value
True Positive (Sensitivity)	0.88
True Negative (Specificity)	0.90
Positive Predicted Value	0.89
Negative Predicted Value	0.88
Accuracy	0.89
Kappa	0.78
Precision	0.89
F-Score	0.89
Gini	0.90
Area under (ROC)	0.95

The table 4 represents the strong reliability in differentiating positive and negative cases. With a Sensitivity of 0.88, it accurately identifies 88% of true positives, while a Specificity of 0.90 means it correctly detects 90% of true negatives. The Positive Predictive Value (PPV) of 0.89 indicates that 89% of positive predictions are accurate, and a Negative Predictive Value (NPV) of 0.88 confirms 88% accuracy for negative predictions. With an Accuracy of 0.89, the model makes correct predictions 89% of the time. The Cohen's Kappa of 0.78 indicates substantial agreement beyond chance, while the Precision and F-Score both at 0.89 show a balanced performance in precision and recall. A Gini Coefficient of 0.90 and an AUC of 0.95 further highlight the model's strong ability to distinguish between outcomes, making it a reliable choice for effective classification in real-world applications.

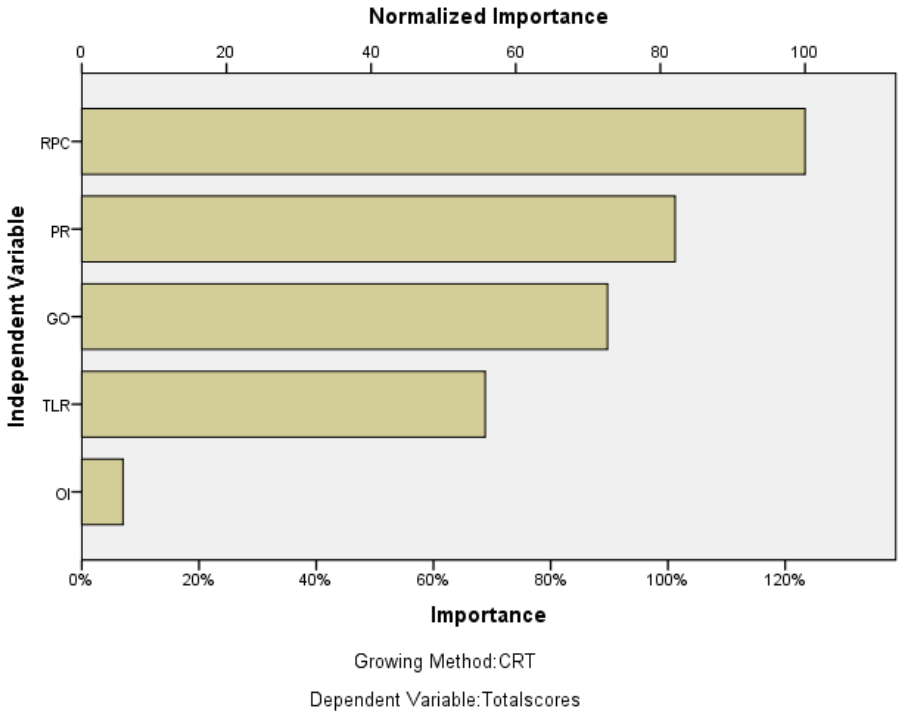

Figure 2: Features Importance in Predicting Total scores Using CRT Method

The figure 2 represents bar chart that illustrates the relative importance of five independent variables - RPC, PR, GO, TLR, and OI in predicting the dependent variable "Total scores," based on the CRT (Classification and Regression Tree) method. RPC holds the highest importance with a normalized value of 100%, making it the strongest predictor of Total scores. PR is the next most influential, with an importance around 80%. GO and TLR have moderate predictive power, with importance values near 60% and 40%, respectively. OI has the least influence, at about 10%. This suggests that RPC and PR are the primary contributors to predicting Total scores, while GO, TLR, and especially OI have progressively lesser impacts.

CONCLUSION

The CART model identifies RPC and TLR as key predictors of total NIRF scores, offering actionable insights to improve institutional performance and guide targeted interventions. While it effectively classifies institutions into performance categories, its precision in predicting exact scores is limited. The model is sensitive to data variations and may overfit if pruning is not applied. Despite these limitations, it remains a valuable tool for highlighting important performance factors and informing future research and quality improvement initiatives.

REFERENCES

1. Milanović M, Stamenković M. CHAID Decision Tree: Methodological Frame and Application. Economic Themes. 2016 Dec 1;54(4):1-24.

2. Charbuty B, Abdulazeez A. Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends. 2021 Mar 24;2(01):1-9.

3. Hood C, Dixon R, Beeston C. Rating the Rankings: Assessing International Rankings of Public Service Performance. International Public Management Journal. 2008 Aug 25;11(3):1–8.

4. Mukherjee, B., Ranking Indian Universities through Research and Professional Practices of National Institutional Ranking Framework (NIRF): A case study of Selected Central Universities in India. Journal of Indian Library Association, 52(4), 2017: 1-15.

5. Sivaperumal S, Abudhahir A. Top 100 Ranked Indian Institutions by NIRF 2021 in Engineering: An Interesting Analysis of Individual/Combined Metrics. Journal of Engineering Education Transformations. 2023 Jan 28;36(Special Issue 2)1-7.

6. Muniappan Ramaraj, Dhandapani Sabareeswaran, V. Vijayalaksmi, Chembath Jothish, N. Thangarasu, Govindaraj Manivasagam. Sophisticated CPBIS methods applied for FBISODATA clustering algorithm using with real time image database. Indonesian journal of electrical engineering and computer science. 2023 Apr 1;30(1):1–4.

7. Mishra BP, Dash R, Rath B. ML Use for Forecasting the NIRF Ranking of Engineering Colleges in India and PCA To Find the Correct Weightage for The Best Result. Webology (ISSN: 1735-188X). 2021;18(6)1-12.

8. Prathap G. Making scientometric sense out of NIRF scores. Current Science. 2017 Mar 25:1240-2:1-3.

9. CART (Classification and Regression Tree) in Machine Learning [Internet]. GeeksforGeeks. 2022. Available from: https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/

10. Biggs D, De Ville B, Suen E. A method of choosing multiway partitions for classification and decision trees. Journal of Applied Statistics. 1991 Jan;18(1):1-11.

11. CART (Classification and Regression Tree) in Machine Learning [Internet]. GeeksforGeeks. 2022. Available from: https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/

12. Quinlan JR. Induction of decision trees. Machine learning. 1986 Mar; 1:1-26.

13. Polamuri S. How the CART Algorithm (Classification and Regression Trees) Works - Dataaspirant [Internet]. 2023. Available from: https://dataaspirant.com/cart-algorithm/