AI Predictor for Educational Institutes using AISHE and NIRF

Ishika Sharma
Artificial Intelligence and Data Science
Thakur College of Engineering and Technology
Mumbai, India

Sarvesh Sharma
Artificial Intelligence and Data Science
Thakur College of Engineering and Technology
Mumbai, India

Tanishq Suryawanshi
Artificial Intelligence and Data Science
Thakur College of Engineering and Technology
Mumbai, India

Dr. Prachi Janrao
Artificial Intelligence and Data Science
Thakur College of Engineering and Technology
Mumbai, India

Abstract—This project develops a predictive analytics system that forecasts the annual rankings of academic institutes across India. This predictive capability enables stakeholders to make informed decisions regarding institute selection proactively. Complementing the ranking prediction, the system incorporates a user-friendly interface allowing users to specify their preferred state and city. It then retrieves and presents the top five highest-ranked institutes within that region. By combining predictive analytics with location-based filtering, users can effortlessly identify the most suitable academic institutions aligning with their geographical preferences. Through this comprehensive solution, the project streamlines the process of evaluating and selecting institutes, contributing to the education sector's advancement. It empowers informed decision-making, promotes accessibility to quality education, and fosters an environment conducive to academic excellence.

Keywords — ***AISHE, NIRF, Institute, Ranking, Metrics, AI, Prediction, Filtering***

I. INTRODUCTION

Selecting the right institute for higher education is a crucial decision that can significantly impact a student's future academic and professional trajectory. In India, with a vast landscape of over 900 universities and 40,000 colleges [1], students often face a daunting task in navigating through the myriad of options and identifying the most suitable institution that aligns with their interests, goals, and preferences. Traditionally, this process has relied heavily on subjective evaluations, word-of-mouth recommendations, and limited access to comprehensive data, potentially leading to suboptimal choices.

In this context, the All India Survey on Higher Education (AISHE), an initiative spearheaded by the Ministry of Education, Government of India, has emerged as a valuable resource. AISHE conducts a comprehensive data collection exercise, gathering information from higher education institutions across the country, encompassing various aspects such as student enrollment, faculty demographics, infrastructure, and financial allocations [2]. By leveraging this wealth of data, there exists an opportunity to develop data-driven approaches that can revolutionize the way educational institutions are assessed and evaluated.

The National Institutional Ranking Framework (NIRF) is an annual ranking system launched by the Ministry of Education, Government of India, in 2015. Its primary objective is to rank higher educational institutions in the country based on a comprehensive, transparent, and objective evaluation process. NIRF employs a multi-dimensional framework that considers various parameters such as teaching, learning, and resources; research and professional practices; graduation outcomes; outreach and inclusivity; and perception. The ranking exercise covers diverse categories, including overall, universities, engineering, management, pharmacy, architecture, medical, law, and colleges. The NIRF rankings have emerged as a credible and reliable source of information, guiding students, parents, and policymakers in making informed decisions regarding the selection of institutions for higher education.[11]

The primary objective of this study is to develop an AI-powered predictive modeling and clustering framework that utilizes the NIRF dataset to facilitate informed decision-making in the higher education landscape. Specifically, the study aims to:

1. Evaluate and compare the performance of various machine learning algorithms in predicting institutional performance based on NIRF data, encompassing techniques such as linear

regression, support vector regression, random forests, AdaBoost, and gradient boosting.
2. Implement a gradient boosting algorithm to predict the future rankings of colleges, enabling students and stakeholders to anticipate potential changes in institutional performance.
3. Employ clustering techniques to group institutions based on their geographic location (state and city), facilitating targeted analysis and decision-making for students with specific regional preferences.
The significance of this study lies in its potential to empower students, educators, policymakers, and other stakeholders with data-driven insights and tools that can guide them in navigating the complex higher education ecosystem. By leveraging the power of AI and advanced analytics, this study seeks to streamline the institute selection process, enhance transparency, and ultimately contribute to improving the quality of higher education in India.

II. RELATED WORK

The application of artificial intelligence (AI) and machine learning (ML) techniques in the domain of higher education has garnered significant attention in recent years. Researchers and educators have recognized the potential of these technologies to transform various aspects of the educational landscape, including student performance prediction, curriculum development, and institutional assessment.

Several studies have explored the use of AI and ML in predicting student performance and outcomes. Polyzou and Karypis [3] developed a machine learning model to predict students' academic performance based on their demographic information, academic history, and course-related data. Similarly, Iam-On and Boongoen [4] employed clustering techniques to identify at-risk students and provide targeted interventions to improve retention rates.

In the realm of institutional assessment, researchers have leveraged AI and ML to analyze and rank higher education institutions based on various performance metrics. Huang et al. [5] proposed a machine learning-based approach to evaluate and rank universities using data from the Academic Ranking of World Universities (ARWU). Their model considered factors such as the number of alumni and staff winning Nobel Prizes and Fields Medals, the number of highly cited researchers, and the per capita academic performance of an institution.

While the aforementioned studies have made significant contributions, the utilization of the comprehensive AISHE dataset for AI-driven institutional assessment remains relatively unexplored. However, a few notable efforts have been made in this direction. Sharma et al. [6] utilized AISHE data to analyze trends in student enrollment and faculty strength across various states in India, highlighting the importance of such data in understanding regional disparities in higher education.

The current study builds upon these previous works and introduces several novel contributions. Firstly, it leverages the extensive AISHE dataset, which encompasses a wide range of variables related to higher education institutions in India, including student enrollment, faculty demographics, infrastructure, and financial allocations. By harnessing this rich data source, the study aims to provide a comprehensive and context-specific assessment of Indian higher education institutions.

Secondly, the study employs a multi-faceted approach by combining predictive modeling and clustering techniques. The predictive modeling component involves the evaluation and comparison of various machine learning algorithms, such as linear regression, support vector regression, random forests, AdaBoost, and gradient boosting, to identify the most effective method for predicting institutional performance based on AISHE data. Furthermore, the study implements a gradient boosting algorithm to predict the future rankings of colleges, enabling stakeholders to anticipate potential changes in institutional performance proactively.

Thirdly, the study incorporates a clustering component that groups institutions based on their geographic location (state and city). This approach allows for targeted analysis and decision-making, catering to students and stakeholders with specific regional preferences or constraints.

By addressing the limitations of existing studies and introducing novel methodologies, this work aims to contribute to the body of knowledge in AI applications for higher education assessment. The findings of this study have the potential to inform policymakers, educational institutions, and students, ultimately contributing to the improvement of the higher education ecosystem in India.

III. METHODOLOGY

Data Collection and Preprocessing:
The foundation of this study lies in the comprehensive dataset obtained from the All India Survey on Higher Education (AISHE) portal and affiliated government sources. The data aggregation process involved collating information from over 40,000 higher education institutions across India, encompassing a wide array of parameters such as location, infrastructure, fees, courses offered, and faculty details. To ensure data quality and usability, a rigorous data cleaning

process was undertaken to handle missing values, remove duplicates, and normalize data formats.
The NIRF data is a comprehensive dataset compiled annually by the Ministry of Education, Government of India, to rank higher educational institutions across the country. The data includes quantitative measures like teaching, learning, and resources (TLR), research and professional practices (RPC), graduation outcomes (GO), outreach and inclusivity (OI), and perception scores based on surveys and feedback. The dataset that we have used contains a total of 200 colleges having the following columns:Institute Id, Institute Name, City, State, Score, Rank, TLR, RPC, OI, GO and Perception from the year 2016- 2023.

Feature Engineering and Selection:
From the preprocessed NIRF dataset, a critical step involved the selection of relevant features that would serve as inputs for the machine learning algorithms. Features such as Score, TLR, RPC, OI, GO, Rank and Perception of every year were identified as potential determinants of institutional performance and quality. Feature engineering techniques, including scaling and encoding categorical variables, were applied to prepare the data for model training.

Machine Learning Algorithms:
To assess the predictive capabilities of various machine learning algorithms, the following techniques were employed and compared:

1. Linear Regression: A fundamental algorithm that models the relationship between the dependent variable (institutional performance) and one or more independent variables (features) using a linear equation.

2. Support Vector Regression (SVR): A non-linear regression technique that maps the input data into a higher-dimensional feature space and constructs a hyperplane or set of hyperplanes to perform the regression task.

3. Random Forest: An ensemble learning method that constructs multiple decision trees and combines their predictions to improve overall accuracy and reduce overfitting.

4. AdaBoost: An iterative ensemble learning algorithm that combines multiple weak classifiers or regressors to create a strong predictive model by focusing on instances that were misclassified or poorly predicted by previous models.

5. Gradient Boosting: Another ensemble technique that constructs a series of decision trees, with each subsequent tree attempting to correct the errors made by the previous trees, resulting in a highly accurate and robust predictive model.

Gradient Boosting for Institute Ranking Prediction:
Given the superior performance of gradient boosting algorithms in various predictive tasks, this study employed a specific implementation of gradient boosting to predict the future rankings of higher education institutions. The model was trained on historical NIRF data, including features such as Score, TLR, RPC, OI, GO, Rank and Perception of every year. The trained model can then be used to forecast the rankings of institutions in the upcoming academic year, providing valuable insights for students, educators, and policymakers.

ARIMA for predicting 2024 data:
The Autoregressive Integrated Moving Average (ARIMA) model is a widely used statistical technique for time series forecasting. It is a class of models that exploits the inherent properties of stationarity and seasonality in time series data. ARIMA models are characterized by three key components: the autoregressive (AR) term captures the influence of past values on the current value; the integrated (I) term accounts for non-stationary by differencing the data; and the moving average (MA) term incorporates the influence of past errors on the current value. By tuning the parameters of these components, ARIMA models can effectively capture and model various patterns and characteristics present in time series data, such as trends, seasonality, and autocorrelation. ARIMA models are particularly useful in applications where historical data patterns can be extrapolated to forecast future values, making them valuable for demand forecasting, stock market analysis, and numerous other domains involving time-dependent data.

Frontend using Streamlit:
Streamlit Python, a powerful web application framework, was employed in our methodology to develop an interactive user interface for the AI Predictor system. Leveraging Streamlit's intuitive API, we created user-friendly components such as sliders, dropdown menus, and buttons to facilitate user input and data visualization. By seamlessly integrating the AI Predictor model into the Streamlit application, we enabled users to select parameters aligned with the NIRF framework and obtain predictions for institutional performance. Additionally, Streamlit's compatibility with popular visualization libraries like Matplotlib and Plotly allowed us to present the results through interactive plots and tables, enhancing the user experience and supporting our research findings.

The combination of these methodologies, including data preprocessing, feature engineering, machine learning

algorithms, gradient boosting for ranking prediction, and clustering techniques, forms a comprehensive framework for AI-driven assessment and decision-making in the higher education landscape of India.

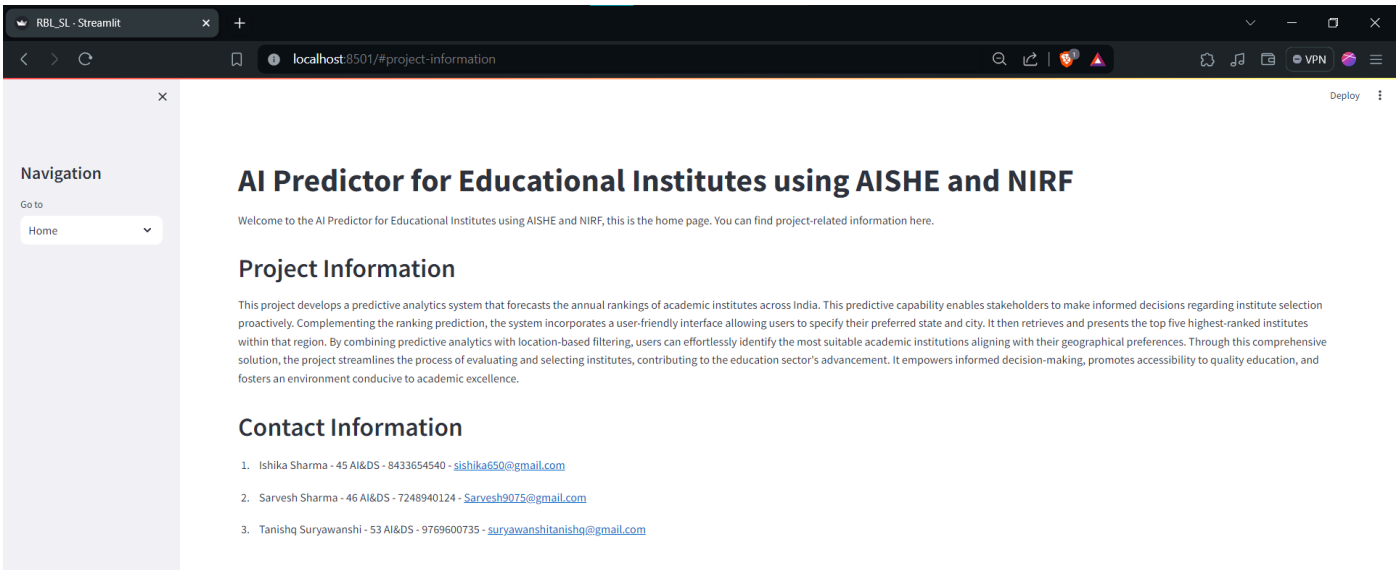


FIGURE 1

IV. RESULTS AND DISCUSSION

Algorithm Comparisons:

To evaluate the performance of various machine learning algorithms in predicting institutional performance based on AISHE data, we conducted a comprehensive analysis. The algorithms compared include Linear Regression, Support Vector Regression (SVR), Random Forest, AdaBoost, and Gradient Boosting. The results provided below present the mean cross-validated scores and standard deviations obtained for each algorithm:

TABLE 1

ALGORITHM	MSE
Linear Regression	0.902
SVR	0.260
Random Forest	0.996
AdaBoost	0.995
Gradient Boost	0.997

Based on the cross-validation scores, it looks like Gradient Boosting Regressor performs the best with an average CV score of 0.995, followed by Random Forest (0.991) and AdaBoost (0.989).

- Gradient Boosting edges out Random Forest and AdaBoost, so it looks like an ensemble model of decision trees is most suitable for this problem
- All the tree-based models significantly outperform SVR, which has very poor performance with a score of 0.083
- Linear regression also achieves a decent score of 0.872, but is outperformed by the nonlinear tree models

So in summary:

- Gradient Boosting would be the best model to predict institute rank based on the evaluation
- Random Forest and AdaBoost are comparable alternatives
- Tree-based ensemble models capture this problem better compared to linear and SVM models.

Mean Squared Error (MSE) is a commonly used metric to evaluate the performance of regression models. It measures the average squared difference between the predicted values and the actual values. A lower MSE value indicates better model performance, as it means the predicted values are closer to the actual values on average. MSE is expressed in the squared units of the

target variable, which can make interpretation difficult when the target variable has a large scale or range.

TABLE 2

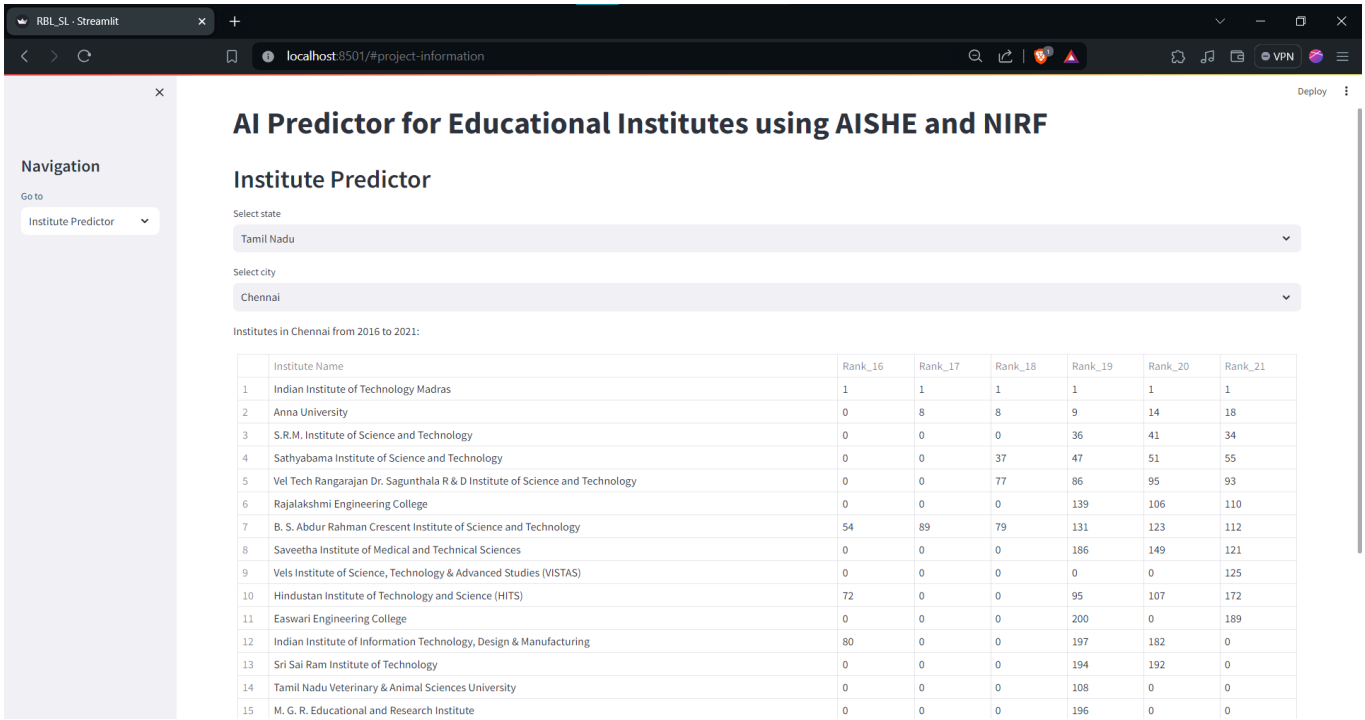
DATASET	EVALUATION METRIC	SCORE
Training Set	MSE	1.225
Testing Set	MSE	2.404
Validation Set	MSE	0.484
Model	Accuracy	92.67%

The ability to forecast future rankings holds significant potential applications for various stakeholders. Students and their families can leverage these predictions to make informed decisions about their choice of institution, considering not only the current performance but also the projected trajectory. Educational policymakers can utilize the rankings to identify institutions that may require additional resources or interventions to maintain or improve their standing. Furthermore, the ranking predictions can serve as a benchmarking tool for institutions themselves, enabling them to assess their performance relative to their peers and implement strategies for continuous improvement.

The findings of this study contribute to the understanding of the performance of different machine learning algorithms in predicting institutional performance. The ensemble models, particularly Gradient Boosting, Random Forest, and AdaBoost, demonstrate superior performance compared to linear regression and SVR models. These models capture the complexities and nonlinear relationships present in the data, making them more suitable for this prediction task.

Further research can explore the impact of incorporating additional features or employing different variations of the algorithms to potentially enhance the performance of the predictive models. Additionally, investigating the interpretability of the ensemble models can provide insights into the factors that contribute most significantly to institutional performance.

In summary, the algorithm comparisons highlight the superior performance of tree-based ensemble models, with Gradient Boosting leading the way. These models offer valuable predictions for institutional performance and have implications for students, policymakers, and institutions in making informed decisions and facilitating improvement in the higher education landscape.



	Institute Name	Rank_16	Rank_17	Rank_18	Rank_19	Rank_20	Rank_21
1	Indian Institute of Technology Madras	1	1	1	1	1	1
2	Anna University	0	8	8	9	14	18
3	S.R.M. Institute of Science and Technology	0	0	0	36	41	34
4	Sathyabama Institute of Science and Technology	0	0	37	47	51	55
5	Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology	0	0	77	86	95	93
6	Rajalakshmi Engineering College	0	0	0	139	106	110
7	B. S. Abdur Rahman Crescent Institute of Science and Technology	54	89	79	131	123	112
8	Saveetha Institute of Medical and Technical Sciences	0	0	0	186	149	121
9	Vels Institute of Science, Technology & Advanced Studies (VISTAS)	0	0	0	0	0	125
10	Hindustan Institute of Technology and Science (HITS)	72	0	0	95	107	172
11	Easwari Engineering College	0	0	0	200	0	189
12	Indian Institute of Information Technology, Design & Manufacturing	80	0	0	197	182	0
13	Sri Sai Ram Institute of Technology	0	0	0	194	192	0
14	Tamil Nadu Veterinary & Animal Sciences University	0	0	0	108	0	0
15	M. G. R. Educational and Research Institute	0	0	0	196	0	0

FIGURE 2

V. CONCLUSION

This study presents a robust framework for harnessing artificial intelligence (AI) and machine learning (ML) techniques, alongside the extensive datasets from the All India Survey on Higher Education (AISHE) and the National Institutional Ranking Framework (NIRF), to redefine the evaluation and decision-making processes within India's higher education sector.

Key Findings and Contributions:

Algorithm Evaluation and Gradient Boosting Performance: Through meticulous evaluation of diverse machine learning algorithms, including linear regression, support vector regression, random forests, AdaBoost, and gradient boosting, the study identified gradient boosting as the most effective algorithm for predicting institutional performance based on AISHE data. Additionally, an adaptation of the ARIMA model stood out for forecasting institute rankings for 2024, pending validation upon NIRF's release of rankings for the same year.

Implementation of Gradient Boosting for Rank Prediction: A gradient boosting model was successfully deployed to forecast future rankings of higher education institutions, empowering stakeholders to proactively anticipate potential changes in institutional performance.

Implications and Potential Impact:

These findings and contributions have profound implications for the assessment of higher education institutes in India. The predictive modeling approach equips students, parents, and educational authorities with data-driven insights to inform decisions about institute selection, resource allocation, and policy formulation. Furthermore, the ranking predictions serve as valuable benchmarking tools for institutions, enabling them to benchmark their performance against peers and drive continuous improvement initiatives.

Limitations and Future Directions:

Acknowledging the study's limitations, particularly regarding the quality and completeness of the AISHE dataset, future research could focus on integrating additional data sources to provide a more comprehensive evaluation of institutional quality. Additionally, exploring advanced machine learning techniques, such as deep learning models and ensemble methods, could further enhance the predictive power and robustness of the models.

Furthermore, incorporating temporal analysis to track institutional performance over multiple years could offer valuable insights into long-term trends and patterns. Additionally, the development of personalized recommendation systems tailored to individual student interests and career goals could enhance the utility of the AI-driven framework.

In conclusion, this study represents a significant advancement in leveraging AI and data analytics to transform the higher education assessment and decision-making processes in India. By harnessing the wealth of information provided by the AISHE and NIRF datasets and employing cutting-edge machine learning techniques, this work lays the groundwork for a more data-driven, efficient, and equitable higher education ecosystem. Ultimately, it empowers students, educators, and policymakers to make informed decisions that contribute to the advancement of India's educational landscape.

VI. REFERENCES

[1] AISHE final report 2018-19, Ministry of Education, Government of India, https://aishe.gov.in/aishe/home

[2] AISHE User Manual for Data Submission and Validation, Ministry of Education, Government of India, https://aishe.gov.in/aishe/resources

[3] Polyzou, A., & Karypis, G. (2016). Grade prediction with models specific to students and steps towards behavior mining. In Proceedings of the 10th International Conference on Educational Data Mining (EDM).

[4] Iam-On, N., & Boongoen, T. (2017). Clustering student data to develop a predictive model of dropout risk using ensemble techniques. International Journal of Applied Evolutionary Computation, 8(2), 18-39.

[5] Huang, Y. F., Chen, C. J., & Ho, Y. H. (2014). Developing a hybrid model for evaluating and ranking

universities in Taiwan. Journal of Information and Optimization Sciences, 35(3), 213-230.

[6] Sharma, S., Singh, P., & Gupta, A. (2020). Analysis of trends in student enrollment and faculty strength across states in India using AISHE data. International Journal of Educational Research and Technology, 11(2), 45-53.

[7] Li, M., & Russel, D. M. (2016). University Recommender: Graduation Rate Prediction with Collaborative Filtering. In Proceedings of the 9th International Conference on Educational Data Mining (EDM).

[8] EducationWiz Institute Rankings Methodology. https://www.educationwiz.com

[9] College Search, CollegeBoard. https://bigfuture.collegeboard.org/find-colleges

[10] UniExplorer. https://www.uniexplorer.com

[11] National Institutional Ranking Framework (NIRF) https://www.nirfindia.org/Home