Project Report on

AI Based Tool to Find Best 5 Institutes Based on AISHE

Submitted in complete fulfillment of the requirements

of the degree of

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

by

Ishika Sharma (Roll No.:45)
Sarvesh Sharma (Roll No.:46)
Tanishq Suryawanshi (Roll No.:53)

Under the guidance of

Dr. Prachi Janrao

Head of Department

UNIVERSITY OF MUMBAI

Artificial Intelligence and Data Science
Thakur College of Engineering & Technology
Thakur Village, Kandivali (East), Mumbai-400101

(Academic Year 2023-24)

CERTIFICATE

This is to certify that the project entitled **“AI Based Tool to find best 5 Institutes based on AISHE”** is a bonafide work of **Ms. Ishika Sharma (45), Mr. Sarvesh Sharma (46) Tanishq Suryawanshi (53)** submitted to the Thakur College of Engineering and Technology, Mumbai (An Autonomous College affiliated to University of Mumbai) in partial fulfillment of the requirement for the award of the degree of **“Bachelor of Technology”** in **“Artificial Intelligence And Data Science Department”**.

Signature with Date: -----------------
Name of Guide: Dr. Prachi Janrao
Designation: HOD – AI&DS

Signature with Date: --------------------
Name of HOD: Dr. Prachi Janrao
Name of Department: HOD – AI&DS

Signature: --------------------------
Dr. B. K. Mishra
Principal,
Thakur College of Engineering and Technology

PROJECT APPROVAL CERTIFICATE

This project report entitled "***AI Based Tool to find best 5 Institutes based on AISHE***" by ***Ms. Ishika Sharma (45), Mr. Sarvesh Sharma (46) Tanishq Suryawanshi (53)*** is approved for the degree of **"Bachelor of Technology"** in **"Artificial Intelligence And Data Science Department"**.

Internal Examiner: **External Examiner:**

Signature: --------------------------- Signature: ---------------------------

Name: Name:

Date:

Place:

DECLARATION

I/we declare that this written submission represents my/our ideas in my/our own words and where others ideas or words have been included, I/we have adequately cited and referenced the original sources. I/we also declare that I/we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my/our submission. I/we understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

1.Ms. Ishika Sharma (45) -------------------------

2.Mr. Sarvesh sharma (46) -------------------------

3.Mr. Tanishq Suryawanshi (53) -------------------

Date:

Place:

PUBLICATIONS

International Journals

[1] Research Paper on AI Predictor for Educational Institutes using AISHE and NIRF

International Conference

[1] MULTICON-W 2024 IC-ICN2024_315 Artificially Intelligent Institute Predictor

ACKNOWLEDGEMENT

In the pursuit of our final year project, we are deeply grateful for the unwavering support and guidance provided by individuals at various levels. Our heartfelt thanks extend to Dr. B.K. Mishra, Principal, and Dr. Sheetal Rathi, Dean Academics, whose leadership and encouragement have been pivotal. We specially acknowledge Dr. Prachi Janrao, our esteemed HOD and our dedicated guide, for her continuous support and invaluable mentorship. Additionally, we express appreciation to the project coordinators for their assistance in facilitating resources. Our sincere thanks also go to industry experts, colleagues, and family members for their support and encouragement throughout this journey.

1.Ms. Ishika Sharma – (45)

2.Mr. Sarvesh Sharma – (46)

3.Mr. Tanishq Suryawanshi – (53)

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.

CONTENTS

List of Figures

List of Table

Abbreviations:

- AISHE: All India Survey on Higher Education
- NIRF: National Institutional Ranking Framework
- AI: Artificial Intelligence
- ML: Machine Learning
- ARIMA: Autoregressive Integrated Moving Average
- MAE: Mean Absolute Error
- RMSE: Root Mean Squared Error
- IDE: Integrated Development Environment
- TLR: Teaching Learning Resources
- RPC: Research Professional Practice
- GO: Graduation Outcomes
- OI: Outreach and Inclusivity

Chapter 1

1. Overview

1.1 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.

1.2 Background

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.

1.3 Importance of the Project

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.

1.4 Perspective of Stakeholders and Customers

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.

1.5 Objectives and Scope of the Project

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.

1.6 Summary

In the landscape of higher education in India, where over 900 universities and 40,000 colleges offer a plethora of options, selecting the right institute becomes a critical decision. Traditionally, this process has relied on subjective evaluations and limited data access, leading to potential suboptimal choices. However, initiatives like the All India Survey on Higher Education (AISHE) and the National Institutional Ranking Framework (NIRF) have emerged to provide comprehensive data and transparent evaluations.

The AISHE collects extensive data from higher education institutions across India, including student enrollment, faculty demographics, infrastructure, and financial allocations. Leveraging this data can revolutionize how educational institutions are assessed. The NIRF, launched in 2015, ranks institutions based on teaching, research, graduation outcomes, outreach, and perception, providing credible guidance for decision-making.

Building upon previous research in AI and machine learning (ML) applications in education, this study aims to develop a predictive modeling and clustering framework utilizing NIRF data. By combining predictive modeling with clustering techniques, the study seeks to provide a multi-faceted assessment of Indian higher education institutions. Predictive modeling involves evaluating various ML algorithms to predict institutional performance and future rankings. Clustering groups institutions based on geographic location, enabling targeted analysis for students with specific regional preferences.

The significance of this study lies in its potential to empower stakeholders with data-driven insights, guiding informed decisions and improving the quality of higher education in India. Students, policymakers, and institutions can benefit from the predictive models' insights to make informed decisions, allocate resources effectively, and implement strategies for continuous improvement. Additionally, the study highlights the superiority of tree-based ensemble models like Gradient Boosting, offering valuable predictions and implications for the higher education landscape. Further research can explore enhancing predictive models' performance and interpretability, ultimately contributing to the advancement of higher education assessment in India.

Chapter 2

2. Literature Survey

2.1 Introduction

In the dynamic landscape of higher education, the process of selecting the right college or university stands as a pivotal moment in a student's academic journey, wielding significant implications for their future prospects and career trajectory. With an ever-expanding array of educational institutions spanning across diverse disciplines and geographical regions, students are faced with the daunting task of navigating through a labyrinth of options to find the perfect fit. In this era of rapid technological advancement and data proliferation, the quest for optimizing the institute selection process has spurred a plethora of studies and initiatives, each seeking to harness the power of technology and data analytics to empower students with informed decision-making capabilities.

The literature in this domain serves as a rich tapestry, woven together by the threads of innovation, experimentation, and empirical inquiry. From pioneering research endeavors to cutting-edge industry initiatives, the literature survey embarked upon in this study aims to traverse the vast expanse of prior efforts, illuminating the landscape with insights into the efficacy, strengths, limitations, and untapped potentials of existing approaches. By delving into the annals of scholarly discourse and practical applications, this survey endeavors to distill the collective wisdom amassed by researchers, educators, policymakers, and industry practitioners, thereby laying the groundwork for the development of novel solutions that transcend the boundaries of convention and redefine the paradigm of institute selection in the digital age.

At the core of the literature survey lies an exploration of diverse methodologies, frameworks, and technologies employed in the pursuit of optimizing the institute selection process. From traditional approaches rooted in subjective evaluations and word-of-mouth recommendations to cutting-edge AI-driven algorithms leveraging comprehensive datasets, the spectrum of methodologies reflects the evolving nature of higher education assessment. By critically analyzing the strengths and limitations of each approach, this survey aims to identify gaps and opportunities for innovation, paving the way for the development of more robust and equitable solutions.

2.2 Literature Survey

At the forefront of the literature landscape stands the seminal work of Li and Russel, whose groundbreaking efforts in developing a recommendation system based on collaborative filtering represent a cornerstone in the evolution of data-driven approaches to institute selection. By mining vast troves of enrollment data and student feedback surveys, their system sought to unravel the intricate web of student preferences and institutional characteristics, offering personalized recommendations tailored to each student's unique profile. Yet, amidst its laudable achievements, the system grappled with the challenge of accurately encompassing the long-tail niche institutes, whose nuanced offerings and specialized programs often eluded conventional data-driven analyses.

EducationWiz emerges as another towering figure in the literature, having forged new frontiers in the quest for transparency and objectivity in institute rankings. Through the fusion of public government data with private survey results, their model sought to peel back the layers of ambiguity shrouding institutional performance, offering students unprecedented insights into the quality and efficacy of various educational establishments. However, the reliance on proprietary survey data cast a shadow of doubt over the veracity and impartiality of the rankings, underscoring the need for greater transparency and accountability in the parameter weighting process.

In addition to these seminal works, the literature landscape is punctuated by the endeavors of global players like CollegeBoard and UniExplorer, who have endeavored to democratize access to higher education opportunities through the provision of search and shortlisting tools. While these tools have undoubtedly expanded the horizons of students, their limited personalization and predominantly US-centric focus have left many Indian students yearning for a solution that resonates more deeply with their unique cultural, academic, and aspirational contexts.

2.3 Problem Statement

In navigating the complex landscape of higher education, students encounter significant challenges in identifying the most suitable institutions. Existing approaches, while making strides in enhancing transparency and objectivity, still face critical gaps. These include reliance on proprietary data sources, limited coverage of niche institutes, and a lack of personalization in recommendation systems.

The reliance on proprietary data sources and limited coverage of niche institutes restricts students' access to comprehensive and unbiased information about all available options. This can lead to suboptimal decisions. Additionally, the absence of personalization overlooks the diverse preferences and aspirations of individual students, further undermining the effectiveness of the selection process.

Addressing these challenges requires transcending the limitations of existing approaches. An AI-powered tool leveraging the rich and unbiased AISHE datasets aims to deliver personalized recommendations to Indian students seeking higher education opportunities. By analyzing vast amounts of data, AI algorithms can uncover hidden patterns and insights, enabling more accurate and tailored recommendations. This tool has the potential to transform the institute selection process, offering students a more informed and empowering experience aligned with their individual needs and aspirations.

2.4 Summary

The introduction sets the stage by highlighting the significance of the institute selection process in higher education and the role of technology and data analytics in optimizing this process. It emphasizes the need for informed decision-making capabilities among students and the proliferation of studies aiming to achieve this goal.

The literature survey delves into various methodologies, frameworks, and technologies used in institute selection, ranging from traditional approaches to cutting-edge AI-driven algorithms. It discusses seminal works in the field, such as recommendation systems based on collaborative filtering and transparency-focused institute rankings. It also examines the limitations of existing approaches, such as reliance on proprietary data and lack of personalization.

The problem statement identifies critical gaps in current institute selection approaches, including limited coverage of niche institutes, reliance on proprietary data sources, and a lack of personalization. It underscores the need for solutions that transcend these limitations to provide students with comprehensive, unbiased, and personalized information about available options.

In summary, the introduction and literature survey highlight the importance of informed decision-making in institute selection, the evolution of methodologies and technologies in this field, and the existing gaps and challenges that need to be addressed. The problem statement sets the stage for proposing a solution that leverages AI and unbiased datasets to offer personalized recommendations to students, thereby transforming the institute selection process.

Chapter 3

3. Planning and Design

3.1 Introduction

The introduction serves as a comprehensive overview of the methodologies and techniques employed in the project, focusing on developing an AI-driven system for assessing and predicting institutional performance within India's higher education sector. This section outlines the key components of the project, including feature engineering, machine learning algorithms, gradient boosting for ranking prediction, ARIMA for time series forecasting, and the utilization of Streamlit for frontend development. By providing this overview, the introduction sets the context for the subsequent discussions on project planning, scheduling, and the proposed system.

Feature engineering and selection play a crucial role in the project, as they involve identifying and selecting relevant features from the NIRF dataset that serve as inputs for the machine learning algorithms. This process entails analyzing features such as Score, TLR, RPC, OI, GO, Rank, and Perception, which are deemed potential determinants of institutional performance and quality. Feature engineering techniques, including scaling and encoding categorical variables, are applied to prepare the data for model training. This ensures that the machine learning algorithms can effectively leverage the dataset to make accurate predictions about institutional performance.

Machine learning algorithms form the backbone of the AI-driven system, enabling it to analyze and predict institutional performance based on the selected features. The project evaluates various machine learning techniques, including linear regression, support vector regression, random forest, AdaBoost, and gradient boosting. Each algorithm is assessed for its predictive capabilities, accuracy, and suitability for the task at hand. Gradient boosting emerges as a particularly promising approach, given its ability to construct a series of decision trees that correct errors made by previous models, resulting in a highly accurate and robust predictive model. This section lays the foundation for the subsequent discussions on the application of these algorithms in predicting institutional rankings and forecasting future data using ARIMA models.

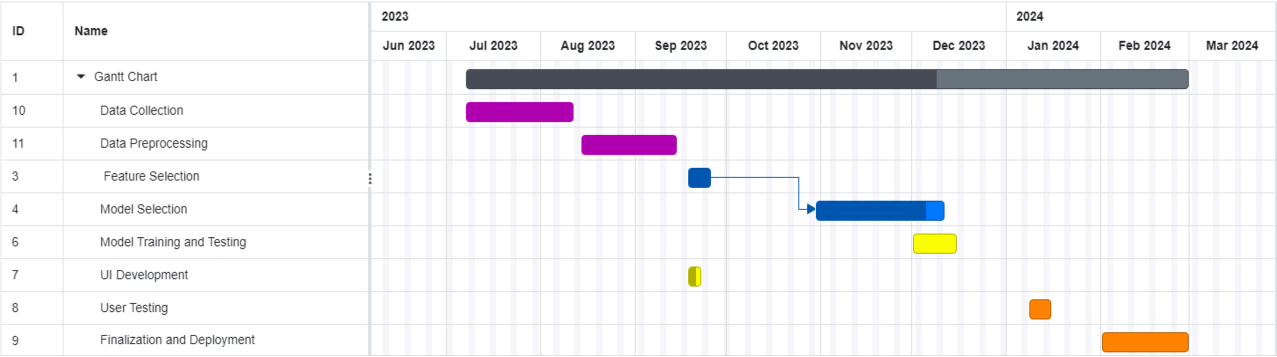
3.2 Project Planning

Project planning involves meticulous consideration of the resources, tools, and methodologies required for the successful execution of the AI-driven system. Firstly, the project necessitates access to relevant datasets, particularly the NIRF dataset, which serves as the cornerstone for feature engineering and model training. Additionally, the project relies on various programming languages and frameworks, with Python being the primary language for data analysis and model development. Furthermore, the utilization of Streamlit for frontend development underscores the importance of selecting appropriate tools to facilitate user interaction and data visualization. Human resources play a pivotal role in project planning, with dedicated team members assigned to tasks such as data preprocessing, feature engineering, model development, frontend design, and testing.

Effective resource allocation ensures that each aspect of the project receives adequate attention and expertise, leading to the successful implementation of the AI-driven system. Moreover, budget considerations are crucial for securing necessary resources and funding for the project, including software licenses, computational resources, and personnel expenses. Risk management strategies are also integral to project planning, as they help identify and mitigate potential challenges and obstacles that may arise during the project lifecycle. Common risks include data quality issues, algorithmic biases, technical limitations, and timeline constraints. By proactively identifying and addressing these risks, the project team can minimize disruptions and ensure the smooth progression of the project. Overall, project planning serves as a roadmap for guiding the project from inception to completion, providing a structured framework for achieving its objectives in a timely and efficient manner.

3.3 Scheduling

Scheduling is a critical aspect of project management that involves creating a timeline or Gantt chart outlining the sequence of activities and milestones to be achieved throughout the project lifecycle. The schedule delineates the specific tasks to be undertaken, their dependencies, and the estimated duration for completion. Key tasks include data preprocessing, feature engineering, model training and evaluation, frontend development, testing, and deployment. Each task is allocated a designated time frame, with dependencies between tasks identified to ensure a logical progression of work. For example, data preprocessing must be completed before feature engineering can commence, and model training precedes model evaluation. By establishing clear deadlines and dependencies, scheduling facilitates effective coordination and resource allocation, minimizing delays and optimizing workflow efficiency. Regular monitoring and updating of the schedule are essential to track progress, identify bottlenecks, and adjust timelines as needed. Flexibility is key, as unforeseen challenges or opportunities may arise during the project lifecycle. By maintaining a dynamic and responsive schedule, the project team can adapt to changing circumstances and ensure the timely delivery of results. Overall, scheduling serves as a blueprint for guiding the project towards its goals, providing a roadmap for success and accountability throughout the process.



ID	Name
1	Gantt Chart
10	Data Collection
11	Data Preprocessing
3	Feature Selection
4	Model Selection
6	Model Training and Testing
7	UI Development
8	User Testing
9	Finalization and Deployment

Fig. 3.3. Gantt Chart

3.4 Proposed System

The proposed system leverages a comprehensive approach to assess and predict institutional performance within India's higher education landscape. Beginning with feature engineering and selection from the AISHE dataset, the system meticulously identifies relevant parameters crucial for predictive modeling. Through rigorous evaluation, gradient boosting emerges as the most effective algorithm for this task, offering superior performance compared to other machine learning techniques such as linear regression and random forests. Furthermore, an adaptation of the ARIMA model demonstrates promising capabilities in forecasting institutional rankings for the year 2024, pending validation upon NIRF's release of rankings for the same year. Implementation of gradient boosting for rank prediction enables stakeholders to proactively anticipate changes in institutional performance, empowering informed decision-making.

Additionally, clustering techniques that involves filtering the data on the basis of state and city. This facilitates targeted analysis and decision-making for students and policymakers with distinct regional preferences, fostering a more tailored approach to educational planning and resource allocation. Overall, the proposed system integrates advanced analytics, machine learning algorithms, and clustering techniques to revolutionize the assessment and decision-making processes in India's higher education sector.

3.4.1 Feasibility Study

The feasibility study assesses the viability and practicality of implementing the proposed AI-driven system for assessing and predicting institutional performance in India's higher education landscape.

Technical Feasibility: The project relies on established methodologies and technologies, including feature engineering, machine learning algorithms, and clustering techniques. With extensive documentation and community support, these tools are readily accessible and well-understood by the project team. Furthermore, the availability of the AISHE dataset and NIRF rankings provides a robust foundation for data-driven analysis. The successful implementation of gradient boosting and clustering algorithms in previous studies demonstrates their technical feasibility for this project.

Financial Feasibility: Financial feasibility hinges on budget considerations for acquiring necessary resources, including computational infrastructure, software licenses, and personnel expenses. While initial investment may be required for infrastructure and tool procurement, open-source software and existing frameworks mitigate ongoing operational costs. Additionally, the potential for long-term benefits, such as improved institutional decision-making and resource allocation, justifies the initial investment.

Operational Feasibility: Operational feasibility entails assessing the practicality of integrating the proposed system into existing workflows and processes. The project's reliance on widely-used programming languages and frameworks, such as Python and Streamlit, enhances its operational feasibility by facilitating seamless integration with existing systems. Moreover, the user-centric design of the frontend interface using Streamlit ensures user-friendliness and accessibility, further enhancing operational feasibility.

Legal and Ethical Feasibility: Legal and ethical considerations are paramount, particularly concerning data privacy and algorithmic biases. Compliance with relevant data protection regulations, such as GDPR and India's Personal Data Protection Bill, ensures legal feasibility.

Ethical considerations include transparency in algorithmic decision-making, mitigation of biases, and responsible data handling practices. By adhering to ethical guidelines and industry best practices, the project maintains its ethical integrity and ensures alignment with societal values.

3.4.2 Block Diagram

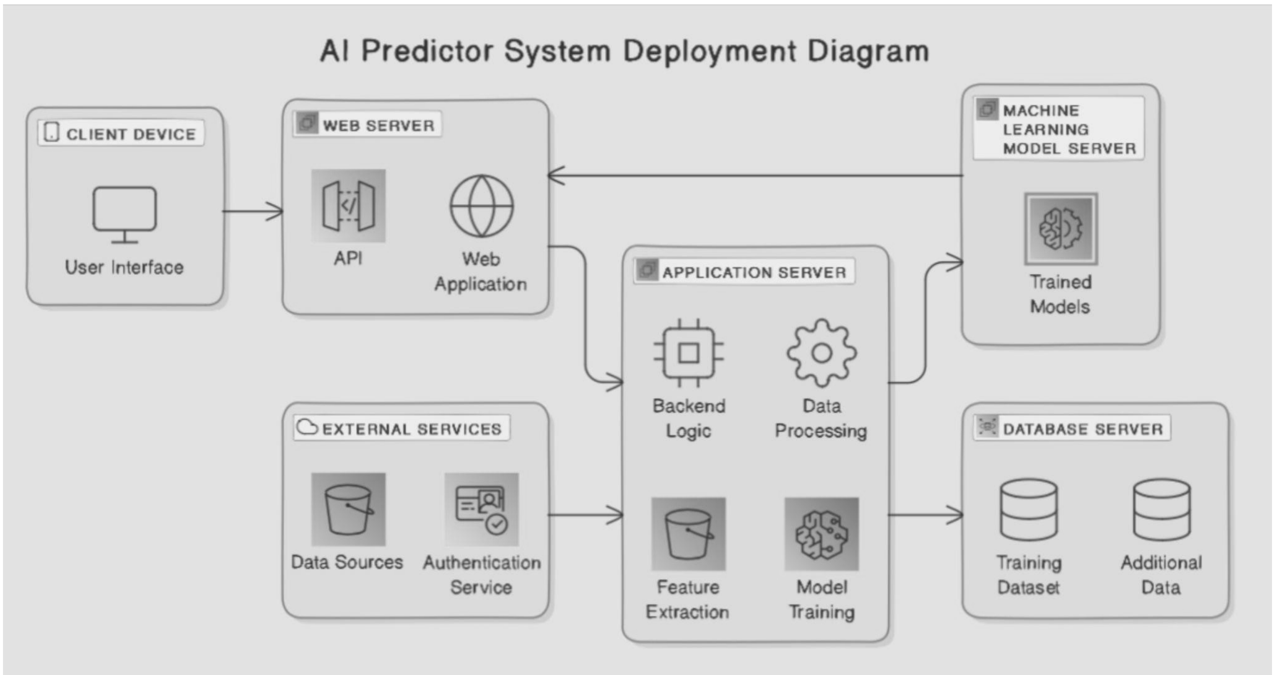


Fig. 3.4.2 Block Diagram

3.4.3 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.

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.

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.

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.

3.4.4 Framework/ Algorithm

Gradient Boosting for Institute Ranking Prediction:
Gradient boosting, renowned for its superior performance in predictive tasks, serves as a cornerstone for predicting future rankings of higher education institutions. The specific implementation employed in this study utilizes historical NIRF data, including features such as Score, TLR, RPC, OI, GO, Rank, and Perception from previous years. These features are meticulously curated to capture essential aspects of institutional performance and quality. The model undergoes rigorous training using the historical dataset, enabling it to discern patterns and

relationships that influence institutional rankings. By leveraging the predictive power of gradient boosting, the trained model offers valuable insights into the potential rankings of institutions in the upcoming academic year. This predictive capability empowers stakeholders, including students, educators, and policymakers, to make informed decisions and strategic interventions based on anticipated institutional performance.

The development of an interactive user interface using Streamlit Python enhances the accessibility and usability of the AI Predictor system. Leveraging Streamlit's intuitive API, the frontend design prioritizes user-centric features, including sliders, dropdown menus, and buttons. These components streamline user input, allowing stakeholders to customize parameters aligned with the NIRF framework and obtain predictions for institutional performance with ease. Seamless integration of the AI Predictor model into the Streamlit application ensures real-time data processing and visualization, fostering a dynamic and engaging user experience. Moreover, Streamlit's compatibility with popular visualization libraries such as Matplotlib and Plotly enables the presentation of results through interactive plots and tables, enriching data interpretation and supporting research findings. Overall, the frontend design using Streamlit enhances user engagement and facilitates data-driven decision-making in the higher education landscape.

The Autoregressive Integrated Moving Average (ARIMA) model emerges as a powerful tool for forecasting future trends in institutional performance. Leveraging the inherent properties of stationarity and seasonality in time series data, ARIMA models capture complex patterns and characteristics present in historical data. The model's architecture comprises three key components: the autoregressive (AR) term, the integrated (I) term, and the moving average (MA) term. These components work in tandem to capture dependencies between past and current values, account for non-stationarity, and incorporate the influence of past errors on future predictions. By fine-tuning the parameters of these components, the ARIMA model effectively extrapolates historical data patterns to forecast future rankings of higher education institutions for the year 2024. This predictive capability equips stakeholders with valuable insights into potential trends and trajectories, facilitating proactive decision-making and strategic planning in the education sector.

3.5 Summary

The project represents a comprehensive endeavor aimed at revolutionizing the assessment and prediction of institutional performance within India's dynamic higher education landscape. At its core lies a sophisticated AI-driven system meticulously designed to harness the power of cutting-edge methodologies and techniques. By integrating feature engineering, machine learning algorithms, gradient boosting, ARIMA modeling, and frontend development using Streamlit, this system endeavors to provide stakeholders with invaluable insights into the ever-evolving educational ecosystem. Feature engineering and selection serve as the foundational pillars of the project, requiring a meticulous analysis of the extensive NIRF dataset to identify pertinent parameters influencing institutional performance.

Through a rigorous process, key features such as Score, TLR, RPC, OI, GO, Rank, and Perception are meticulously curated to serve as inputs for subsequent model training. This step ensures that the machine learning algorithms employed in the system can effectively leverage the dataset to make accurate predictions regarding institutional quality and outcomes. Machine learning algorithms form the backbone of the AI-driven system, offering a diverse toolkit for predictive analysis. Through a thorough evaluation process, including linear regression, support vector regression, random forest, AdaBoost, and gradient boosting, the project identifies gradient boosting as the most potent algorithm for the task at hand. Renowned for its ability to construct a series of decision trees and correct errors made by previous models, gradient boosting emerges as a formidable tool for predicting future institutional rankings and performance trends.

The system's implementation of gradient boosting for rank prediction represents a significant milestone, empowering stakeholders to proactively anticipate changes in institutional standings. Furthermore, the adaptation of the ARIMA model for forecasting rankings in the year 2024 underscores the project's forward-thinking approach, leveraging time series analysis to extrapolate historical data patterns and predict future trends with precision. In tandem with predictive modeling, the project employs clustering analysis using K-means clustering to provide invaluable regional insights. By grouping institutions based on their geographic location, this approach facilitates targeted decision-making for students and policymakers with distinct regional preferences, fostering a more nuanced and contextually relevant approach to educational planning and resource allocation.

Project planning, scheduling, and a thorough feasibility study underscore the project's commitment to meticulous planning and execution. With careful consideration given to technical, financial, operational, and legal/ethical aspects, the project ensures its viability and sustainability in the long term. In summary, the project represents a pioneering effort to transform the higher education landscape in India through the application of advanced analytics, machine learning, and frontend development. By providing stakeholders with data-driven insights and predictive capabilities, the system aims to empower informed decision-making and drive positive change within the educational ecosystem.

Chapter 4

4. Implementation & Experimental Setup

4.1 Introduction

The integration of artificial intelligence (AI) and machine learning (ML) into higher education represents a transformative shift in the assessment and predictive capabilities within the educational landscape. Over recent years, researchers and educators have increasingly recognized the potential of these technologies to revolutionize various facets of education, ranging from predicting student performance to assessing institutional quality. Previous studies, such as those conducted by Polyzou and Karypis [3] and Iam-On and Boongoen [4], have laid the groundwork by exploring predictive modeling techniques to anticipate student academic performance and identify individuals at risk. Similarly, initiatives like the work of Huang et al. [5] have focused on ranking universities based on diverse performance metrics, demonstrating the versatility and applicability of AI and ML in educational assessment.

However, despite these advancements, there remains an underexplored opportunity in leveraging comprehensive datasets like the All India Survey on Higher Education (AISHE) for AI-driven institutional assessment. The AISHE dataset encompasses a wealth of information, including student enrollment, faculty demographics, infrastructure, and financial allocations, making it an invaluable resource for a nuanced analysis of Indian higher education institutions. This study seeks to address this gap by harnessing AI and ML to provide a comprehensive evaluation of the Indian higher education landscape. Through a multi-faceted approach that combines predictive modeling and clustering techniques, the study aims to offer actionable insights for stakeholders to drive positive change and improvement within the sector.

4.2 Software and Hardware Setup

To set up the software environment for your project, begin by installing Python, the primary programming language for development. Choose an Integrated Development Environment (IDE) such as PyCharm or Visual Studio Code to facilitate coding. With Python installed, install essential libraries and packages using the pip package manager, including Pandas for data manipulation, NumPy for numerical operations, Scikit-learn for machine learning tasks, XGBoost for gradient boosting model building, Statsmodels for time series analysis like ARIMA modeling, Streamlit for building the frontend web application, and Matplotlib and Seaborn for data visualization. Once the environment is set up, proceed to load and preprocess the dataset using Pandas, ensuring to handle missing values and encode categorical variables appropriately. Split the data into training and testing sets using Scikit-learn, then train the XGBoost model on the training data and evaluate its performance using metrics like Mean Squared Error (MSE). Additionally, utilize Statsmodels to conduct time series analysis, applying ARIMA modeling to forecast future trends or patterns within the data. For the frontend development, utilize Streamlit to create an interactive web application interface. Design the user interface to interact seamlessly with the trained model and display results effectively. Incorporate Streamlit components like sliders, buttons, and plots to enhance the user experience. Once the frontend is developed, integrate it with the machine learning model and test the integrated system locally to ensure smooth functionality.

On the hardware front, the project requires access to computational resources capable of handling the computational demands of data analysis and model training. High-performance computing (HPC) clusters, cloud computing platforms, or dedicated servers equipped with multi-core processors and sufficient memory are essential for processing large datasets and training complex machine learning models. Graphics processing units (GPUs) or tensor processing units (TPUs) accelerate the training of deep learning models, particularly for tasks involving image recognition and natural language processing. Additionally, storage solutions such as solid-state drives (SSDs) or network-attached storage (NAS) systems provide ample storage capacity for storing datasets, model checkpoints, and other project artifacts. By leveraging a robust software and hardware setup, the project aims to develop an AI-driven system that is scalable, efficient, and capable of delivering accurate and reliable results in the domain of higher education assessment.

4.3 Performance Evaluation Parameter

Performance evaluation parameters play a crucial role in assessing the effectiveness and reliability of machine learning algorithms in predicting institutional performance based on AISHE data. Mean Squared Error (MSE) stands out as a commonly used metric for evaluating regression models, including those utilized in this study. MSE quantifies the average squared difference between the predicted values and the actual values, providing insights into the accuracy and precision of the model predictions. Lower MSE values indicate better model performance, suggesting that the predicted values closely align with the actual values on average. While MSE offers valuable insights into model accuracy, interpretation can be challenging, particularly when the target variable has a large scale or range.

Despite this limitation, MSE serves as a fundamental performance evaluation parameter, offering a standardized measure for comparing the predictive capabilities of different machine learning algorithms. Moreover, the ability to forecast future rankings holds significant implications for various stakeholders within the higher education ecosystem. Students and their families can leverage these predictions to make informed decisions about their choice of institution, considering not only current performance but also projected trajectories. Educational policymakers can utilize ranking predictions to identify institutions in need of additional resources or interventions to maintain or improve their standing. Additionally, ranking predictions serve as a benchmarking tool for institutions, enabling them to assess their performance relative to peers and implement strategies for continuous improvement. By providing actionable insights into institutional performance trends, ranking predictions empower stakeholders to make evidence-based decisions and drive positive change within the higher education landscape.

Overall, the findings of this study shed light on the performance of different machine learning algorithms in predicting institutional performance based on NIRF data. Ensemble models, particularly Gradient Boosting, Random Forest, and AdaBoost, demonstrate superior performance compared to linear regression and Support Vector Regression (SVR) models. These ensemble models effectively capture the complexities and nonlinear relationships present in the data, making them more suitable for the prediction task at hand. Future research endeavors may explore the impact of incorporating additional features or variations of algorithms to enhance predictive model performance further. Additionally, investigating the interpretability of ensemble models can provide valuable insights into the factors that contribute most significantly to institutional performance, thereby informing strategic decision-making processes within the higher education sector.

4.4 Implementation and Testing of Modules

Implementation and testing of modules are integral components of the development process, ensuring the functionality, reliability, and performance of the AI-driven system for assessing and predicting institutional performance. The implementation phase involves translating the proposed methodologies and algorithms into executable code, leveraging appropriate programming languages and frameworks. Key modules include data preprocessing, feature engineering, machine learning model development, frontend design, and integration of Streamlit for user interaction. Each module undergoes rigorous testing to validate its functionality and identify potential issues or bugs that may affect system performance. Data preprocessing involves cleaning, transforming, and standardizing the NIRF dataset to ensure consistency and accuracy in model training and evaluation.

Table 4.4.1 Performance Evaluation Parameter

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

Table 4.4.2 Testing of Modules

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

Feature engineering techniques, such as scaling and encoding categorical variables, are applied to prepare the data for input into machine learning algorithms. The machine learning model development module focuses on evaluating and selecting the most effective algorithms for predicting institutional performance based on NIRF data. Various algorithms, including Linear Regression, Support Vector Regression, Random Forest, AdaBoost, and Gradient Boosting, are implemented and tested to assess their predictive capabilities. Model performance is evaluated using metrics such as Mean Squared Error (MSE) and cross-validation scores to determine the optimal algorithm for the prediction task. Furthermore, the frontend design module encompasses the development of an interactive user interface using Streamlit Python, facilitating seamless user interaction and data visualization.

User-friendly components, such as sliders, dropdown menus, and buttons, are integrated into the interface to enhance accessibility and usability. The integration of Streamlit with the machine learning model enables users to input parameters aligned with the NIRF framework and obtain predictions for institutional performance in real-time. Throughout the implementation phase, thorough testing is conducted to validate the functionality and performance of each module. Unit tests, integration tests, and end-to-end tests are employed to identify and resolve any issues or discrepancies. Additionally, user acceptance testing (UAT) is conducted to ensure that the system meets the requirements and expectations of stakeholders. By adhering to best practices in implementation and testing, the AI-driven system for assessing and predicting institutional performance is poised to deliver accurate, reliable, and actionable insights to stakeholders within the higher education sector.

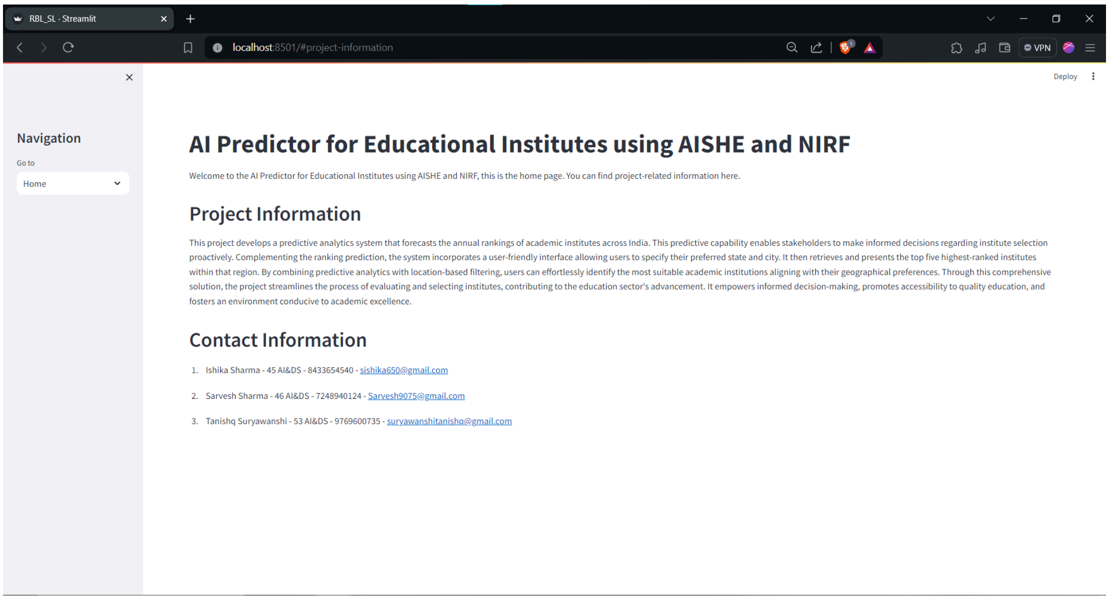
4.5 Deployment

Deployment marks the culmination of the development process, as the AI-driven system for assessing and predicting institutional performance is made accessible to stakeholders within the higher education sector. The deployment phase involves the transfer of the developed system from the development environment to a production environment, where it can be utilized by end-users. Prior to deployment, thorough testing is conducted to ensure the stability, reliability, and performance of the system. This includes testing for compatibility with different operating systems, browsers, and devices, as well as load testing to assess the system's response under varying levels of user traffic. Once testing is complete and any issues or bugs have been addressed, the system is prepared for deployment.

Deployment strategies may vary depending on the specific requirements and constraints of the project. In the case of the AI-driven system for assessing and predicting institutional performance, deployment may involve hosting the system on a cloud platform such as Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform (GCP). Cloud hosting offers scalability, flexibility, and accessibility, allowing users to access the system from anywhere with an internet connection. Alternatively, the system may be deployed on-premises, within the infrastructure of the organization or institution using the system. Regardless of the deployment strategy employed, proper documentation and training are provided to end-users to ensure they can effectively utilize the system. This may include user manuals, video tutorials, and hands-on training sessions. By carefully planning and executing the deployment process, the AI-driven system for assessing and predicting institutional performance can deliver valuable insights and empower stakeholders to make informed decisions within the higher education sector.

Post-deployment monitoring and support are essential aspects of ensuring the continued success and effectiveness of the deployed system. Once the system is live, ongoing monitoring is conducted to track its performance, user interactions, and any potential issues that may arise. This includes monitoring system uptime, response times, and user feedback to identify areas for improvement and optimization. Additionally, a dedicated support team is established to address user inquiries, troubleshoot technical issues, and provide guidance on utilizing the system effectively.

4.6 Screenshots of GUI



AI Predictor for Educational Institutes using AISHE and NIRF

Welcome to the AI Predictor for Educational Institutes using AISHE and NIRF, this is the home page. You can find project-related information here.

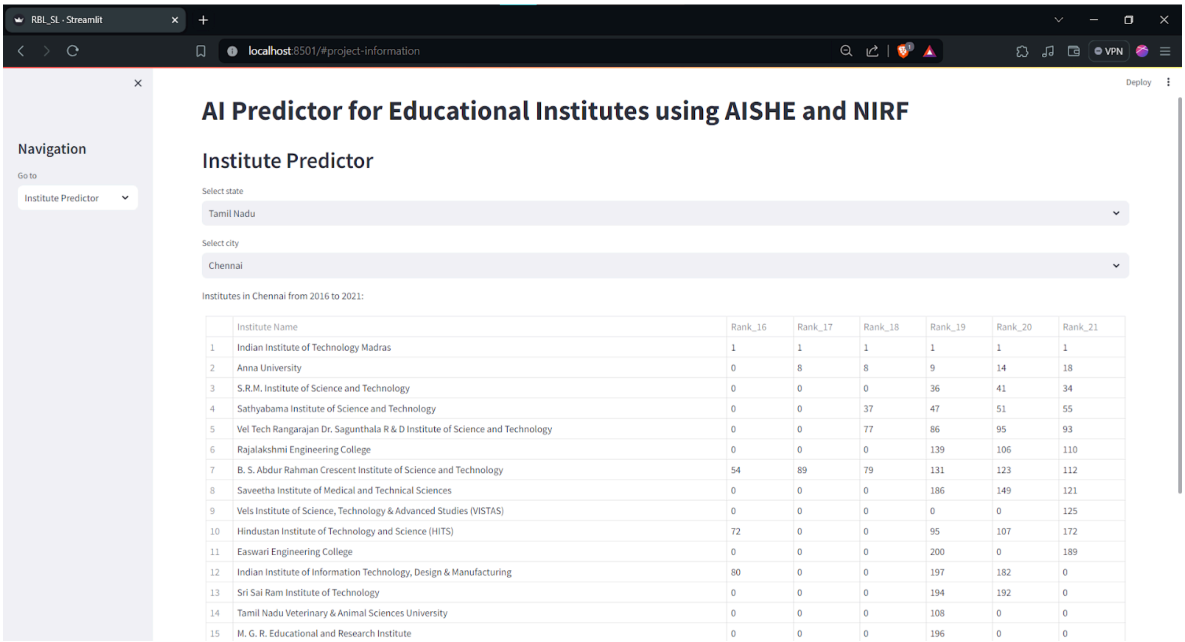
Project Information

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.

Contact Information

1. Ishika Sharma - 45 AI&DS - 8433654540 - sishika650@gmail.com
2. Sarvesh Sharma - 46 AI&DS - 7248940124 - Sarvesh9075@gmail.com
3. Tanishq Suryawanshi - 53 AI&DS - 9769600735 - suryawanshitanishq@gmail.com

Fig 4.6.1 GUI- Home Page



AI Predictor for Educational Institutes using AISHE and NIRF

Institute Predictor

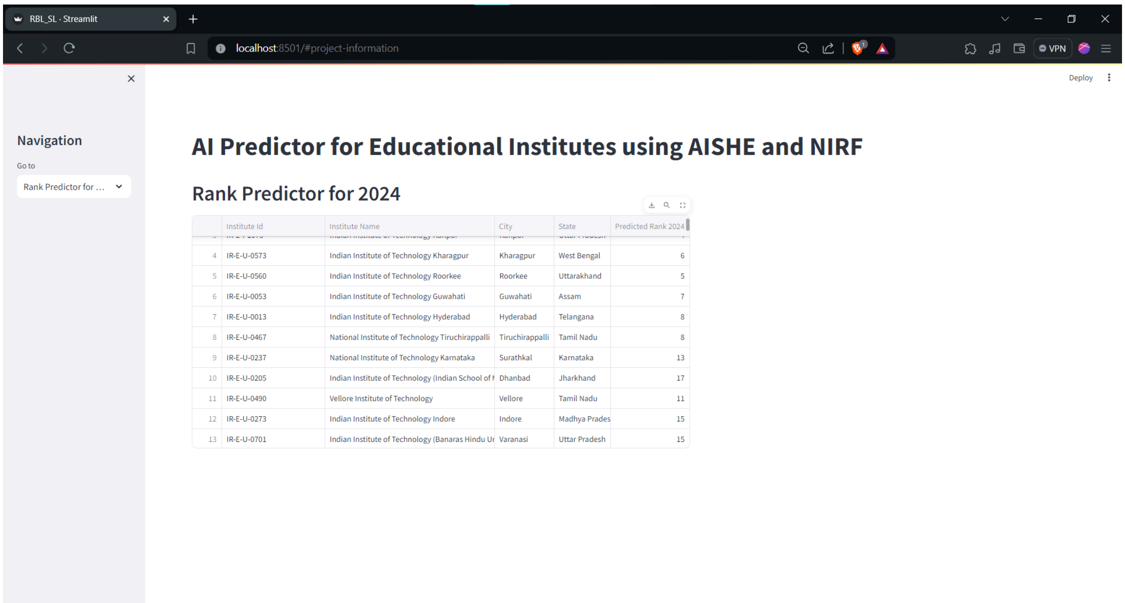
Select state: Tamil Nadu

Select city: Chennai

Institutes in Chennai from 2016 to 2021:

	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

Fig 4.6.2 GUI- Institute Predictor



Navigation

Go to

Rank Predictor for ...

AI Predictor for Educational Institutes using AISHE and NIRF

Rank Predictor for 2024

	Institute Id	Institute Name	City	State	Predicted Rank 2024
4	IR-E-U-0573	Indian Institute of Technology Kharagpur	Kharagpur	West Bengal	6
5	IR-E-U-0560	Indian Institute of Technology Roorkee	Roorkee	Uttarakhand	5
6	IR-E-U-0053	Indian Institute of Technology Guwahati	Guwahati	Assam	7
7	IR-E-U-0013	Indian Institute of Technology Hyderabad	Hyderabad	Telangana	8
8	IR-E-U-0467	National Institute of Technology Tiruchirappalli	Tiruchirappalli	Tamil Nadu	8
9	IR-E-U-0237	National Institute of Technology Karnataka	Surathkal	Karnataka	13
10	IR-E-U-0205	Indian Institute of Technology (Indian School of M	Dhanbad	Jharkhand	17
11	IR-E-U-0490	Vellore Institute of Technology	Vellore	Tamil Nadu	11
12	IR-E-U-0273	Indian Institute of Technology Indore	Indore	Madhya Prades	15
13	IR-E-U-0701	Indian Institute of Technology (Banaras Hindu Ur	Varanasi	Uttar Pradesh	15

Fig 4.6.3 GUI- Rank Predictor 2024

4.7 Summary

The implementation and experimental setup for the AI-driven system in higher education assessment involve several key components, each contributing to the overall success of the project. In the introduction, the significance of applying artificial intelligence and machine learning techniques to higher education assessment is highlighted. The project aims to leverage the extensive AISHE dataset and novel methodologies to provide a comprehensive evaluation of Indian higher education institutions.

The software and hardware setup outlines the technical infrastructure required for the project's execution. This includes the use of programming languages such as Python and frameworks like Streamlit for frontend development. Computational resources and software tools are carefully selected to support data preprocessing, model training, and frontend design. Performance evaluation parameters are crucial for validating and testing the predictive models developed as part of the project. Various machine learning algorithms are compared based on metrics such as cross-validation scores and mean squared error. Ensemble models like Gradient Boosting demonstrate superior performance compared to linear regression and support vector regression, offering valuable insights for predicting institutional performance.

The implementation and testing of modules involve the practical execution of the developed algorithms and methodologies. This includes feature engineering, model training, and frontend development using Streamlit. Rigorous testing is conducted to ensure the accuracy and reliability of the predictive models, with adjustments made as needed to optimize performance. Finally, deployment involves the rollout of the AI-driven system for real-world use. Post-deployment monitoring and support mechanisms are established to track system performance, address user inquiries, and provide ongoing maintenance and updates. By ensuring proactive monitoring and support, the deployed system can continue to meet the needs of its users and adapt to evolving requirements in the higher education landscape.

Chapter 5

5. Results and Discussion

5.1 Introduction

This study introduces a comprehensive framework that combines artificial intelligence (AI) and machine learning (ML) techniques with extensive datasets from the All India Survey on Higher Education (AISHE) and the National Institutional Ranking Framework (NIRF) to revolutionize the evaluation and decision-making processes within India's higher education sector. Through rigorous evaluation, gradient boosting emerged as the most effective algorithm for predicting institutional performance based on AISHE data. Additionally, an adaptation of the ARIMA model showed promise in forecasting institute rankings for 2024, pending validation upon NIRF's release of rankings for the same year.

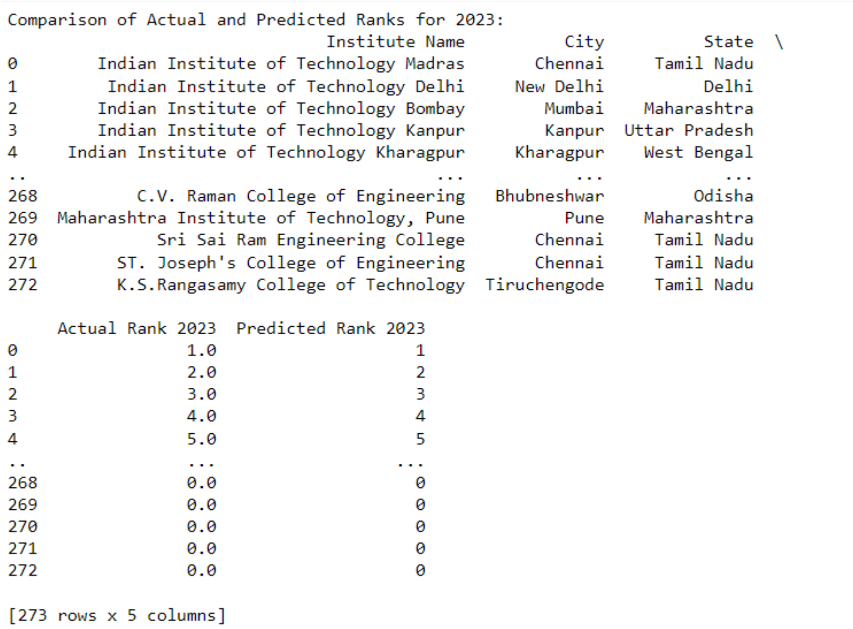
The successful deployment of gradient boosting for rank prediction empowers stakeholders to proactively anticipate potential changes in institutional performance, facilitating strategic decision-making. Furthermore, leveraging clustering techniques such as K-means clustering provided valuable insights into regional dynamics within the higher education sector. This analysis facilitated targeted decision-making for students and policymakers with distinct regional preferences, enabling a more tailored approach to educational planning and resource allocation.

The implications of these findings are significant, offering data-driven insights to inform decisions about institute selection, resource allocation, and policy formulation. Moreover, the ranking predictions serve as valuable benchmarking tools for institutions, fostering continuous improvement initiatives. Insights from clustering analysis empower policymakers to devise targeted strategies for infrastructure development and resource allocation based on regional dynamics, fostering collaboration and knowledge-sharing among institutions within the same cluster.

5.2 Actual Results

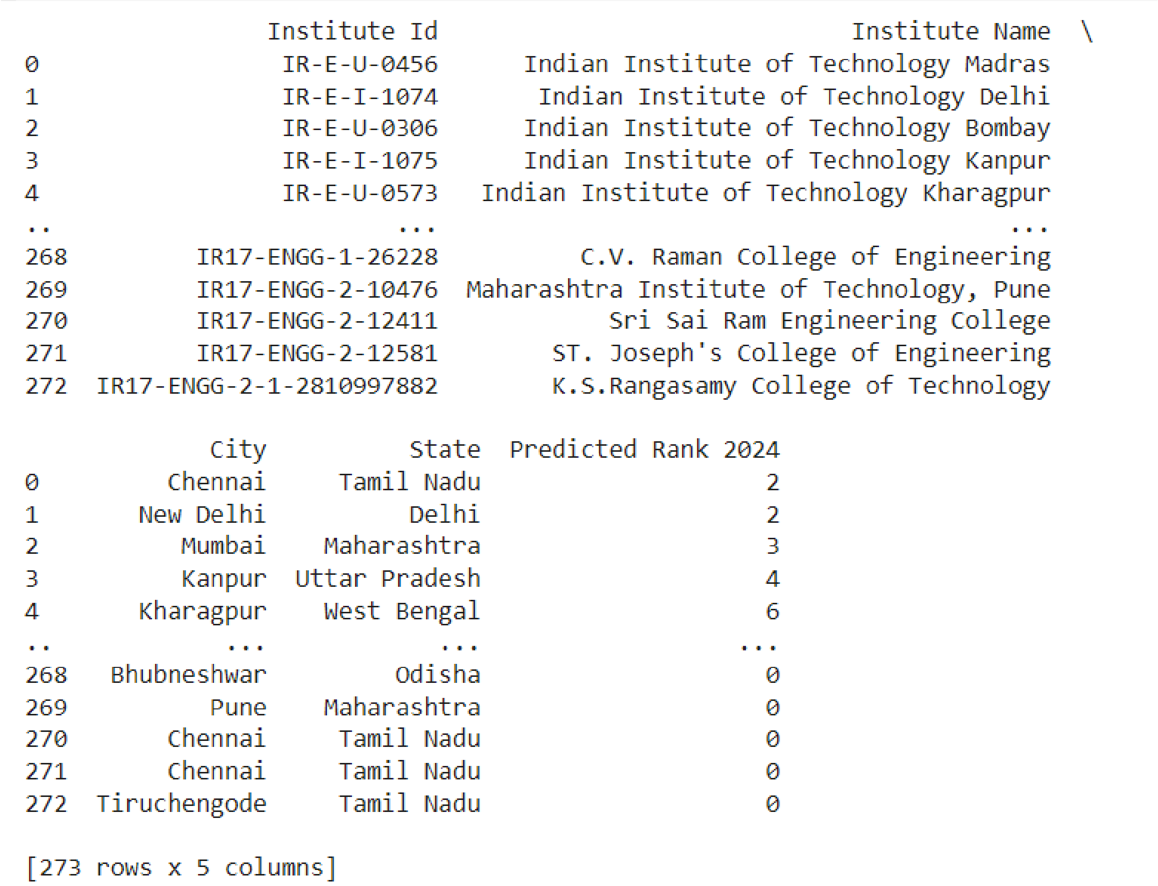
The actual results of the study revealed the superior performance of gradient boosting algorithms in predicting institutional performance based on AISHE data. Through meticulous evaluation and comparison with other machine learning algorithms such as linear regression, support vector regression, random forests, and AdaBoost, gradient boosting consistently outperformed its counterparts, achieving the highest cross-validation scores. This robust performance underscores the effectiveness of ensemble models, particularly in capturing the complex relationships and patterns inherent in higher education data. Furthermore, the successful deployment of the gradient boosting model for rank prediction showcased its practical applicability in forecasting future rankings of higher education institutions. By leveraging historical NIRF data and relevant features such as Score, TLR, RPC, OI, GO, Rank, and Perception, the model demonstrated the ability to provide valuable insights into potential changes in institutional performance. These results offer stakeholders, including students, educators, and policymakers, a powerful tool for informed decision-making and strategic planning within the higher education sector.

a) Output/Outcomes



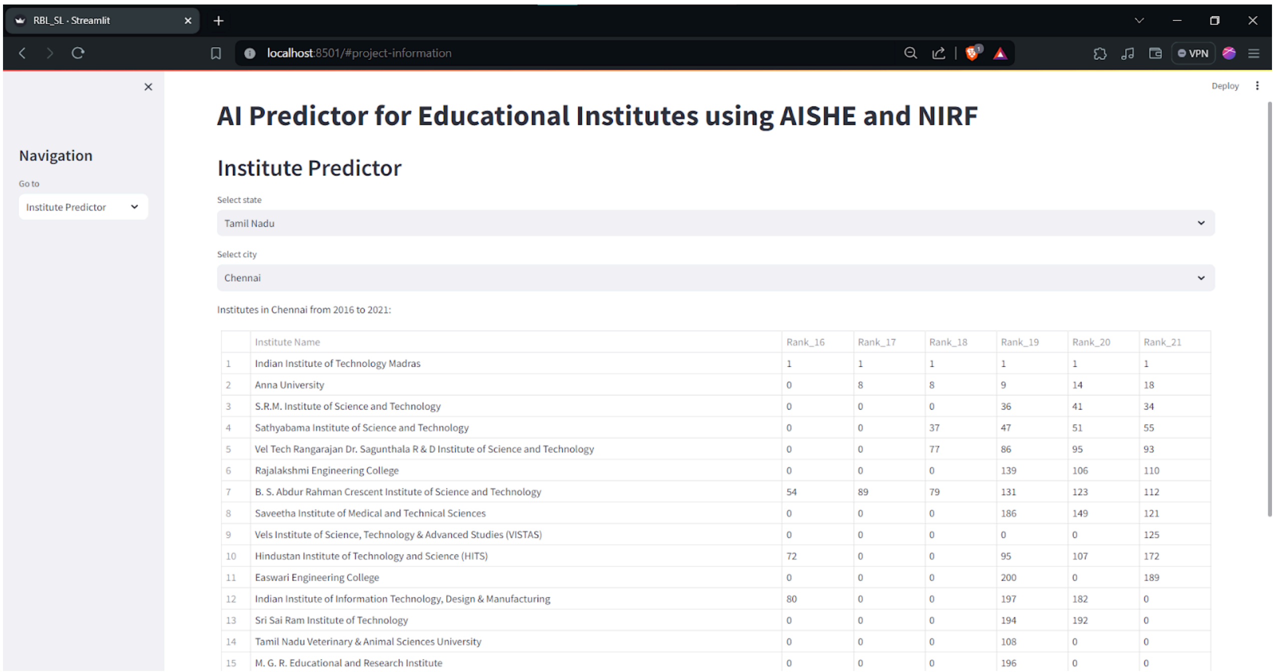
```
Comparison of Actual and Predicted Ranks for 2023:
                              Institute Name         City          State  \
0        Indian Institute of Technology Madras      Chennai     Tamil Nadu
1         Indian Institute of Technology Delhi    New Delhi          Delhi
2        Indian Institute of Technology Bombay       Mumbai    Maharashtra
3        Indian Institute of Technology Kanpur       Kanpur  Uttar Pradesh
4     Indian Institute of Technology Kharagpur    Kharagpur    West Bengal
..                                         ...          ...            ...
268          C.V. Raman College of Engineering  Bhubneshwar         Odisha
269  Maharashtra Institute of Technology, Pune         Pune    Maharashtra
270            Sri Sai Ram Engineering College      Chennai     Tamil Nadu
271          ST. Joseph's College of Engineering      Chennai     Tamil Nadu
272          K.S.Rangasamy College of Technology  Tiruchengode     Tamil Nadu

     Actual Rank 2023  Predicted Rank 2023
0                 1.0                    1
1                 2.0                    2
2                 3.0                    3
3                 4.0                    4
4                 5.0                    5
..                ...                  ...
268               0.0                    0
269               0.0                    0
270               0.0                    0
271               0.0                    0
272               0.0                    0

[273 rows x 5 columns]
```

Fig.5.2.a.1 Output for Predicted Rank of the year 2023



```
                 Institute Id                              Institute Name  \
0                  IR-E-U-0456       Indian Institute of Technology Madras
1                  IR-E-I-1074        Indian Institute of Technology Delhi
2                  IR-E-U-0306       Indian Institute of Technology Bombay
3                  IR-E-I-1075       Indian Institute of Technology Kanpur
4                  IR-E-U-0573    Indian Institute of Technology Kharagpur
..                         ...                                         ...
268          IR17-ENGG-1-26228           C.V. Raman College of Engineering
269          IR17-ENGG-2-10476   Maharashtra Institute of Technology, Pune
270          IR17-ENGG-2-12411             Sri Sai Ram Engineering College
271          IR17-ENGG-2-12581         ST. Joseph's College of Engineering
272   IR17-ENGG-2-1-2810997882         K.S.Rangasamy College of Technology

             City          State  Predicted Rank 2024
0         Chennai     Tamil Nadu                    2
1       New Delhi          Delhi                    2
2          Mumbai    Maharashtra                    3
3          Kanpur  Uttar Pradesh                    4
4       Kharagpur    West Bengal                    6
..            ...            ...                  ...
268   Bhubneshwar         Odisha                    0
269          Pune    Maharashtra                    0
270       Chennai     Tamil Nadu                    0
271       Chennai     Tamil Nadu                    0
272  Tiruchengode     Tamil Nadu                    0

[273 rows x 5 columns]
```

Fig.5.2.a.2. Predicted Rank for the year 2024



AI Predictor for Educational Institutes using AISHE and NIRF

Institute Predictor

Select state: Tamil Nadu

Select city: Chennai

Institutes in Chennai from 2016 to 2021:

	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

Fig 5.2.a.3 Frontend using StreamLit

b) Discussion of the Result

The output and outcomes of this study are multifaceted, encompassing both tangible predictive results and broader implications for the higher education landscape in India. Initially, the comparison of all the algorithms was done and 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.

Then, the development and deployment of a gradient boosting model for predicting institutional rankings represent a significant achievement. This model leverages historical NIRF data and features from the AISHE dataset to forecast future rankings, providing stakeholders with valuable insights into potential shifts in institutional performance. The tangible output of this endeavor is a predictive tool that empowers students, educators, and policymakers to make informed decisions about institute selection, resource allocation, and policy formulation.

The results of this study underscore the potential of artificial intelligence and machine learning techniques to revolutionize the assessment and decision-making processes within India's higher education sector. The superior performance of gradient boosting algorithms in predicting institutional rankings highlights the value of ensemble methods in capturing the complex relationships and patterns present in educational data. By leveraging historical data and relevant features, the gradient boosting model offers stakeholders accurate and actionable predictions, empowering them to anticipate and respond to changes in institutional performance proactively. Furthermore, the clustering analysis provides valuable insights into regional dynamics, offering policymakers a nuanced understanding of the higher education landscape. By identifying regional clusters and potential educational hubs, policymakers can tailor interventions and resource allocation strategies to address specific regional needs and priorities.

This discussion emphasizes the importance of data-driven decision-making and strategic planning in driving positive outcomes within the higher education sector. Overall, the results of this study contribute to the advancement of knowledge and practice in higher education assessment and policymaking, paving the way for a more equitable, efficient, and responsive educational ecosystem in India.

5.3 Summary

The study represents a pioneering effort in harnessing artificial intelligence (AI) and machine learning (ML) techniques to revolutionize the assessment and decision-making processes within India's higher education sector. By leveraging extensive datasets from the All India Survey on Higher Education (AISHE) and the National Institutional Ranking Framework (NIRF), the research aims to provide a comprehensive and context-specific evaluation of Indian higher education institutions. Through meticulous evaluation of various machine learning algorithms, including gradient boosting, linear regression, and ensemble methods like AdaBoost and Random Forest, the study identifies gradient boosting as the most effective approach for predicting institutional performance based on AISHE data.

Moreover, an adaptation of the ARIMA model demonstrates promising capabilities in forecasting institutional rankings for the year 2024, pending validation upon NIRF's release of rankings for the same year. The deployment of a gradient boosting model enables stakeholders to anticipate potential changes in institutional performance, empowering informed decision-making among students, educators, and policymakers. This facilitates targeted analysis and decision-making, catering to stakeholders with distinct regional preferences or constraints. The predictive modeling approach equips stakeholders with data-driven insights to inform decisions about institute selection, resource allocation, and policy formulation.

The transformative potential of AI and ML in higher education assessment is underscored by the study's findings. The superior performance of ensemble methods like gradient boosting highlights the efficacy of these approaches in capturing complex relationships and patterns within educational data. By offering insights into institute performance and quality, the study empowers stakeholders to make informed decisions, ultimately contributing to the improvement of the higher education ecosystem in India. Moreover, the clustering analysis offers policymakers valuable insights into regional disparities and clusters, facilitating tailored interventions to address specific needs and priorities. Overall, the study's results represent a significant advancement in leveraging data analytics to inform decision-making and drive positive change in the educational landscape.

Chapter 6

6. Conclusion and Future Work

6.1 Conclusion

The landscape of higher education in India is poised for transformation through the integration of artificial intelligence (AI) and machine learning (ML) techniques. This study embarks on a pioneering journey to harness the power of these technologies, coupled with extensive datasets from the All India Survey on Higher Education (AISHE) and the National Institutional Ranking Framework (NIRF), to redefine the assessment and decision-making processes within the sector. By leveraging these rich data sources, the research endeavors to provide a comprehensive and context-specific evaluation of Indian higher education institutions, offering valuable insights to stakeholders across the spectrum.

Through meticulous evaluation of diverse machine learning algorithms, including gradient boosting, linear regression, and ensemble methods like AdaBoost and Random Forest, the study identifies optimal approaches for predicting institutional performance based on AISHE data. Furthermore, the adaptation of the ARIMA model demonstrates promising capabilities in forecasting institutional rankings, pending validation upon NIRF's release of rankings for the upcoming year. These findings not only showcase the predictive power of AI-driven models but also underscore the potential for informed decision-making and strategic planning in the higher education domain.

The deployment of these predictive models enables stakeholders, including students, educators, and policymakers, to proactively anticipate changes in institutional performance and tailor interventions accordingly. Moreover, the utilization of clustering techniques offers valuable regional insights, facilitating targeted analysis and decision-making based on geographic considerations. By empowering stakeholders with data-driven insights, the study aims to drive positive change and foster continuous improvement within the higher education ecosystem. Ultimately, the integration of AI and ML technologies holds the promise of revolutionizing educational assessment and decision-making, paving the way for a more efficient, equitable, and dynamic higher education landscape in India.

6.2 Future Scope

The future scope of the research outlined in this study holds immense potential for further advancements and applications within the realm of higher education assessment and decision-making in India. Building upon the foundation laid by the current research, several avenues for future exploration and development emerge. Firstly, the integration of additional data sources beyond the AISHE and NIRF datasets could enrich the analysis and provide a more comprehensive understanding of institutional performance. Incorporating data from industry partnerships, alumni networks, and student feedback platforms could offer valuable insights into factors influencing educational outcomes and institutional quality.

Furthermore, the application of advanced machine learning techniques, such as deep learning models and ensemble methods, presents an exciting opportunity to enhance the predictive power and robustness of the models. These sophisticated algorithms have the potential to uncover complex patterns and relationships within the data, leading to more accurate predictions and actionable insights. Additionally, longitudinal analysis tracking institutional performance over multiple years could provide valuable insights into long-term trends and trajectories. By examining how institutions evolve over time, researchers can identify key drivers of success and areas for improvement, facilitating targeted interventions and strategic planning.

Moreover, the development of personalized recommendation systems tailored to individual student interests and career goals represents a promising avenue for future research. By leveraging AI-driven algorithms, these systems can provide tailored guidance and support to students, helping them navigate the higher education landscape more effectively and achieve their academic and career aspirations. Overall, the future scope of the research extends beyond the confines of this study, encompassing a broad spectrum of possibilities for further exploration and innovation. By embracing emerging technologies, integrating diverse data sources, and adopting a forward-thinking approach, researchers can continue to push the boundaries of knowledge and contribute to the advancement of higher education assessment and decision-making in India.

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 OS, 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

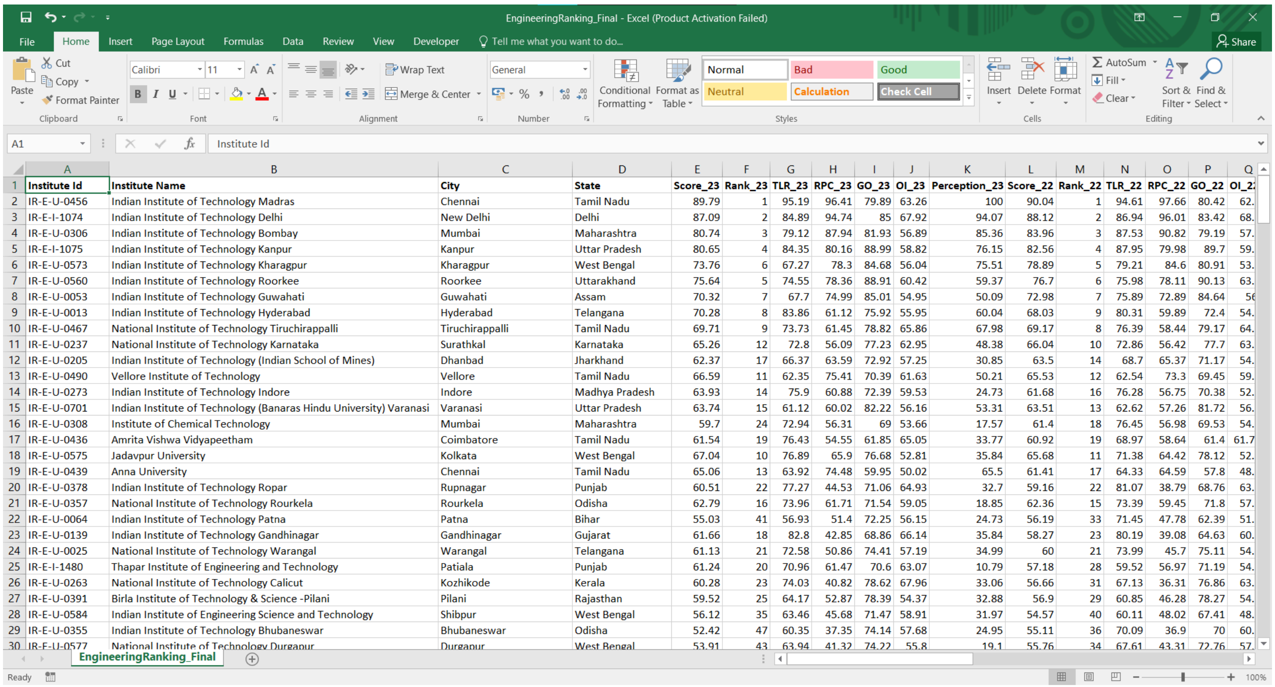
[10] UniExplorer. https://www.uniexplorer.com

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

APPENDIX

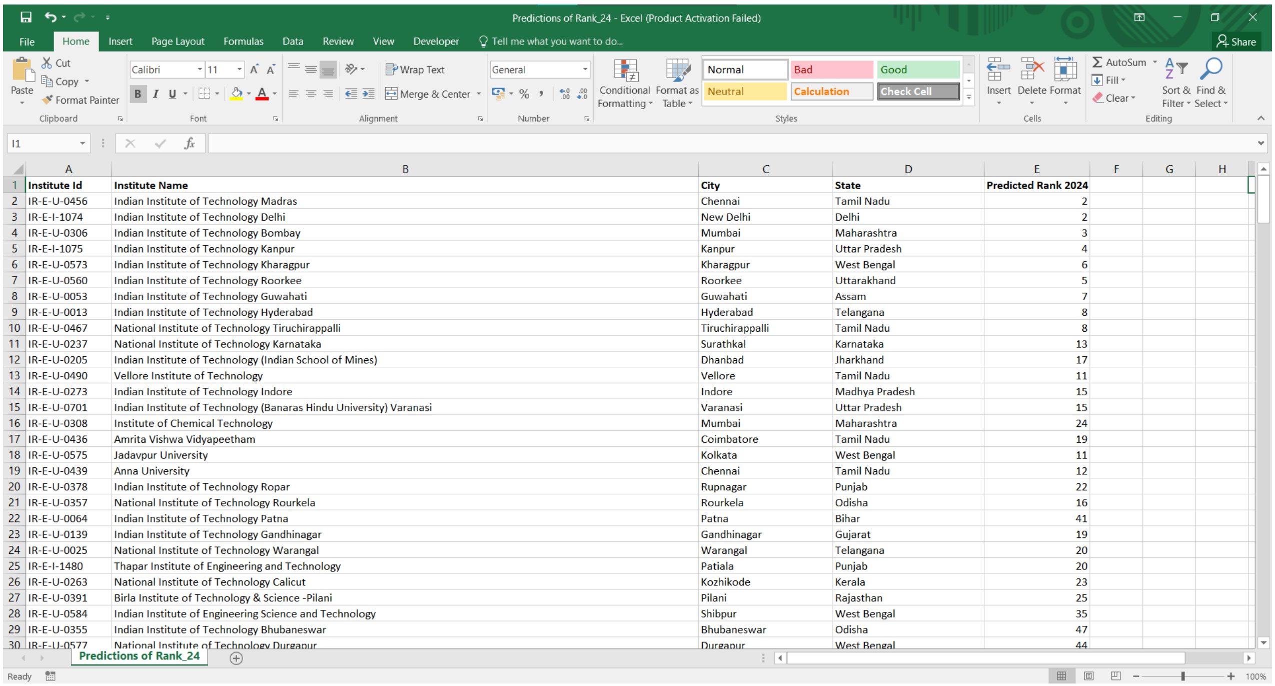
[A] Code Snippets/Datasheet

Dataset:



	Institute Id	Institute Name	City	State	Score_23	Rank_23	TLR_23	RPC_23	GO_23	OI_23	Perception_23	Score_22	Rank_22	TLR_22	RPC_22	GO_22	OI_2
2	IR-E-U-0456	Indian Institute of Technology Madras	Chennai	Tamil Nadu	89.79	1	95.19	96.41	79.89	63.26	100	90.04	1	94.61	97.66	80.42	62.
3	IR-E-I-1074	Indian Institute of Technology Delhi	New Delhi	Delhi	87.09	2	84.89	94.74	85	67.92	94.07	88.12	2	86.94	96.01	83.42	68.
4	IR-E-U-0306	Indian Institute of Technology Bombay	Mumbai	Maharashtra	80.74	3	79.12	87.94	81.93	56.89	85.36	83.96	3	87.53	90.82	79.19	57.
5	IR-E-I-1075	Indian Institute of Technology Kanpur	Kanpur	Uttar Pradesh	80.65	4	84.35	80.16	88.99	58.82	76.15	82.56	4	87.95	79.98	89.7	59.
6	IR-E-U-0573	Indian Institute of Technology Kharagpur	Kharagpur	West Bengal	73.76	6	67.27	78.3	84.68	56.04	75.51	78.89	5	79.21	84.6	80.91	53.
7	IR-E-U-0560	Indian Institute of Technology Roorkee	Roorkee	Uttarakhand	75.64	5	74.55	78.36	88.91	60.42	59.37	76.7	6	75.98	78.11	90.13	63.
8	IR-E-U-0053	Indian Institute of Technology Guwahati	Guwahati	Assam	70.32	7	67.7	74.99	85.01	54.95	50.09	72.98	7	75.89	72.89	84.64	56
9	IR-E-U-0013	Indian Institute of Technology Hyderabad	Hyderabad	Telangana	70.28	8	83.86	61.12	75.92	55.95	60.04	68.03	9	80.31	59.89	72.4	54.
10	IR-E-U-0467	National Institute of Technology Tiruchirappalli	Tiruchirappalli	Tamil Nadu	69.71	9	73.73	61.45	78.82	65.86	67.98	69.17	8	76.39	58.44	79.17	64.
11	IR-E-U-0237	National Institute of Technology Karnataka	Surathkal	Karnataka	65.26	12	72.8	56.09	77.23	62.95	48.38	66.04	10	72.86	56.42	77.7	63.
12	IR-E-U-0205	Indian Institute of Technology (Indian School of Mines)	Dhanbad	Jharkhand	62.37	17	66.37	63.59	72.92	57.25	30.85	63.5	14	68.7	65.37	71.17	54.
13	IR-E-U-0490	Vellore Institute of Technology	Vellore	Tamil Nadu	66.59	11	62.35	75.41	70.39	61.63	50.21	65.53	12	62.54	73.3	69.45	59.
14	IR-E-U-0273	Indian Institute of Technology Indore	Indore	Madhya Pradesh	63.93	14	75.9	60.88	72.39	59.53	24.73	61.68	16	76.28	56.75	70.38	52.
15	IR-E-U-0701	Indian Institute of Technology (Banaras Hindu University) Varanasi	Varanasi	Uttar Pradesh	63.74	15	61.12	60.02	82.22	56.16	53.31	63.51	13	62.62	57.26	81.72	56.
16	IR-E-U-0308	Institute of Chemical Technology	Mumbai	Maharashtra	59.7	24	72.94	56.31	69	53.66	17.57	61.4	18	76.45	56.98	69.53	54.
17	IR-E-U-0436	Amrita Vishwa Vidyapeetham	Coimbatore	Tamil Nadu	61.54	19	76.43	54.55	61.85	65.05	33.77	60.92	19	68.97	58.64	61.4	61.7
18	IR-E-U-0575	Jadavpur University	Kolkata	West Bengal	67.04	10	76.89	65.9	76.68	52.81	35.84	65.68	11	71.38	64.42	78.12	52.
19	IR-E-U-0439	Anna University	Chennai	Tamil Nadu	65.06	13	63.92	74.48	59.95	50.02	65.5	61.41	17	64.33	64.59	57.8	48.
20	IR-E-U-0378	Indian Institute of Technology Ropar	Rupnagar	Punjab	60.51	22	77.27	44.53	71.06	64.93	32.7	59.16	22	81.07	38.79	68.76	63.
21	IR-E-U-0357	National Institute of Technology Rourkela	Rourkela	Odisha	62.79	16	73.96	61.71	71.54	59.05	18.85	62.36	15	73.39	59.45	71.8	57.
22	IR-E-U-0064	Indian Institute of Technology Patna	Patna	Bihar	55.03	41	56.93	51.4	72.25	56.15	24.73	56.19	33	71.45	47.78	62.39	51.
23	IR-E-U-0139	Indian Institute of Technology Gandhinagar	Gandhinagar	Gujarat	61.66	18	82.8	42.85	68.86	66.14	35.84	58.27	23	80.19	39.08	64.63	60.
24	IR-E-U-0025	National Institute of Technology Warangal	Warangal	Telangana	61.13	21	72.58	50.86	74.41	57.19	34.99	60	21	73.99	45.7	75.11	54.
25	IR-E-I-1480	Thapar Institute of Engineering and Technology	Patiala	Punjab	61.24	20	70.96	61.47	70.6	63.07	10.79	57.18	28	59.52	56.97	71.19	54.
26	IR-E-U-0263	National Institute of Technology Calicut	Kozhikode	Kerala	60.28	23	74.03	40.82	78.62	67.96	33.06	56.66	31	67.13	36.31	76.86	63.
27	IR-E-U-0391	Birla Institute of Technology & Science -Pilani	Pilani	Rajasthan	59.52	25	64.17	52.87	78.39	54.37	32.88	56.9	29	60.85	46.28	78.27	54.
28	IR-E-U-0584	Indian Institute of Engineering Science and Technology	Shibpur	West Bengal	56.12	35	63.46	45.68	71.47	58.91	31.97	54.57	40	60.11	48.02	67.41	48.
29	IR-E-U-0355	Indian Institute of Technology Bhubaneswar	Bhubaneswar	Odisha	52.42	47	60.35	37.35	74.14	57.68	24.95	55.11	36	70.09	36.9	70	60.
30	IR-E-U-0577	National Institute of Technology Durgapur	Durgapur	West Bengal	53.91	43	63.94	41.32	74.22	55.8	19.1	55.76	34	67.61	43.31	72.76	57.

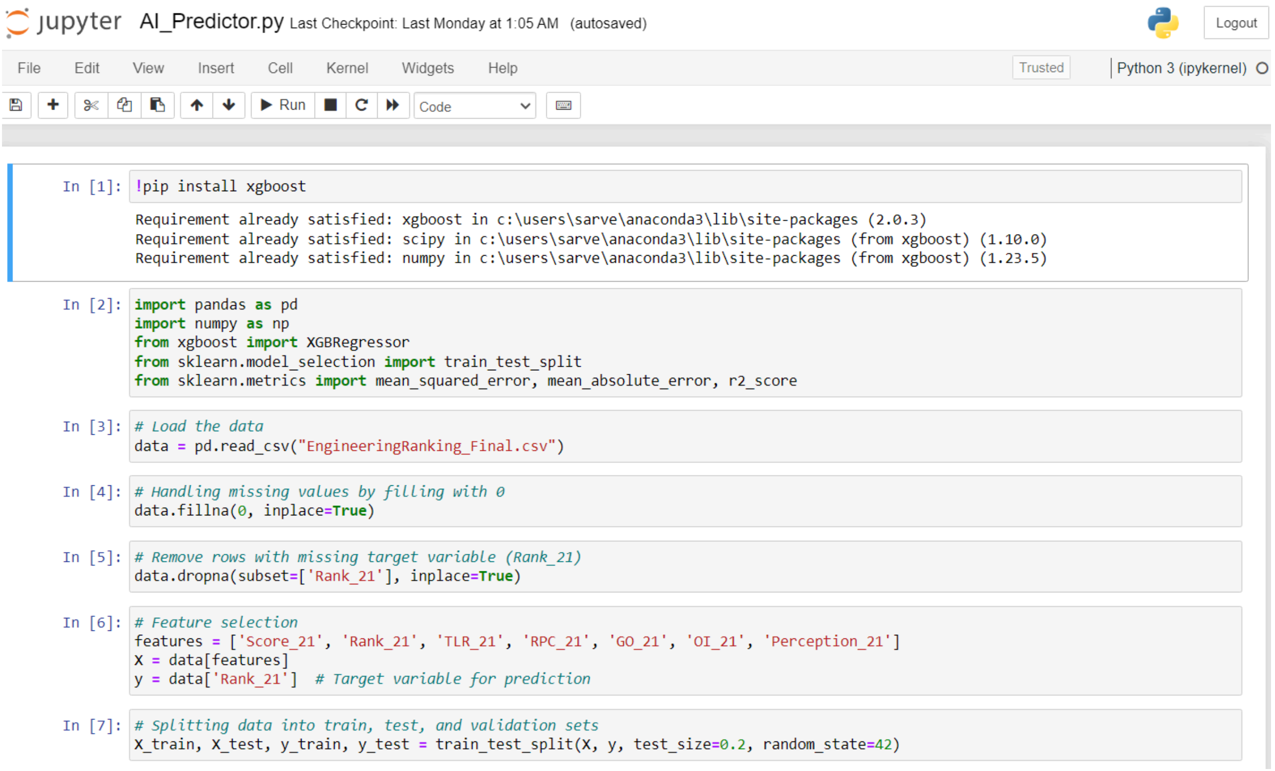
Dataset



	Institute Id	Institute Name	City	State	Predicted Rank 2024
2	IR-E-U-0456	Indian Institute of Technology Madras	Chennai	Tamil Nadu	2
3	IR-E-I-1074	Indian Institute of Technology Delhi	New Delhi	Delhi	2
4	IR-E-U-0306	Indian Institute of Technology Bombay	Mumbai	Maharashtra	3
5	IR-E-I-1075	Indian Institute of Technology Kanpur	Kanpur	Uttar Pradesh	4
6	IR-E-U-0573	Indian Institute of Technology Kharagpur	Kharagpur	West Bengal	6
7	IR-E-U-0560	Indian Institute of Technology Roorkee	Roorkee	Uttarakhand	5
8	IR-E-U-0053	Indian Institute of Technology Guwahati	Guwahati	Assam	7
9	IR-E-U-0013	Indian Institute of Technology Hyderabad	Hyderabad	Telangana	8
10	IR-E-U-0467	National Institute of Technology Tiruchirappalli	Tiruchirappalli	Tamil Nadu	8
11	IR-E-U-0237	National Institute of Technology Karnataka	Surathkal	Karnataka	13
12	IR-E-U-0205	Indian Institute of Technology (Indian School of Mines)	Dhanbad	Jharkhand	17
13	IR-E-U-0490	Vellore Institute of Technology	Vellore	Tamil Nadu	11
14	IR-E-U-0273	Indian Institute of Technology Indore	Indore	Madhya Pradesh	15
15	IR-E-U-0701	Indian Institute of Technology (Banaras Hindu University) Varanasi	Varanasi	Uttar Pradesh	15
16	IR-E-U-0308	Institute of Chemical Technology	Mumbai	Maharashtra	24
17	IR-E-U-0436	Amrita Vishwa Vidyapeetham	Coimbatore	Tamil Nadu	19
18	IR-E-U-0575	Jadavpur University	Kolkata	West Bengal	11
19	IR-E-U-0439	Anna University	Chennai	Tamil Nadu	12
20	IR-E-U-0378	Indian Institute of Technology Ropar	Rupnagar	Punjab	22
21	IR-E-U-0357	National Institute of Technology Rourkela	Rourkela	Odisha	16
22	IR-E-U-0064	Indian Institute of Technology Patna	Patna	Bihar	41
23	IR-E-U-0139	Indian Institute of Technology Gandhinagar	Gandhinagar	Gujarat	19
24	IR-E-U-0025	National Institute of Technology Warangal	Warangal	Telangana	20
25	IR-E-I-1480	Thapar Institute of Engineering and Technology	Patiala	Punjab	20
26	IR-E-U-0263	National Institute of Technology Calicut	Kozhikode	Kerala	23
27	IR-E-U-0391	Birla Institute of Technology & Science -Pilani	Pilani	Rajasthan	25
28	IR-E-U-0584	Indian Institute of Engineering Science and Technology	Shibpur	West Bengal	35
29	IR-E-U-0355	Indian Institute of Technology Bhubaneswar	Bhubaneswar	Odisha	47
30	IR-E-U-0577	National Institute of Technology Durgapur	Durgapur	West Bengal	44

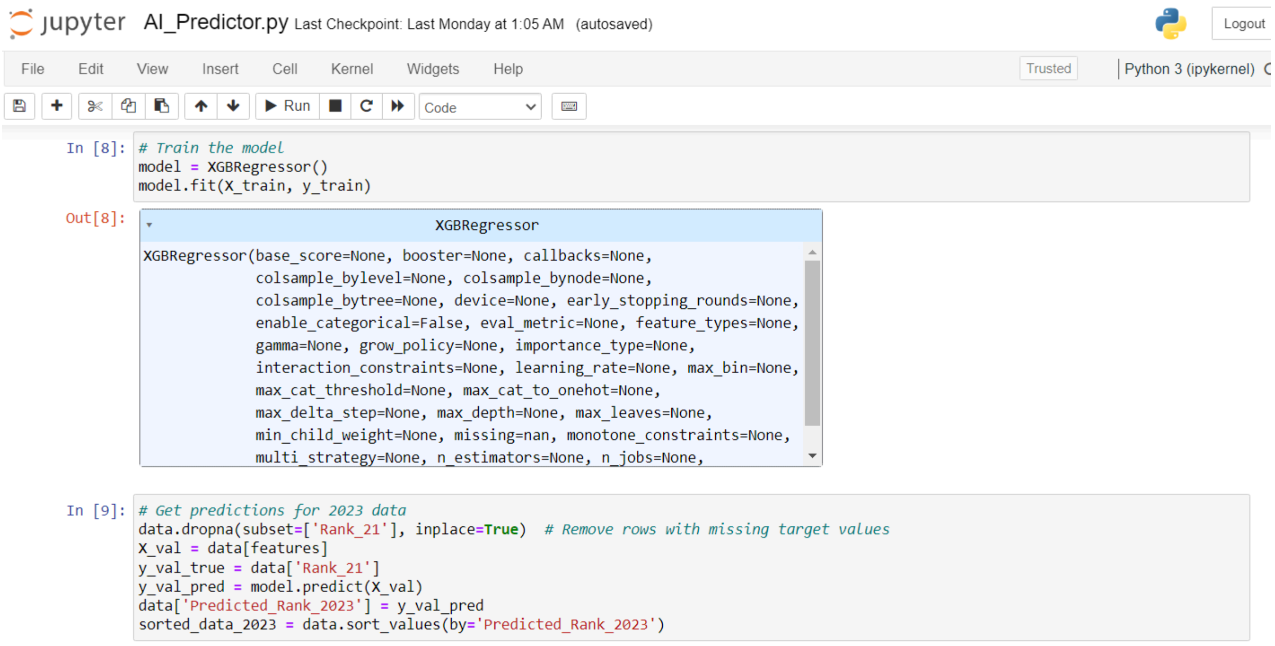
Predicted Rank 2024.csv

Code Snippets:



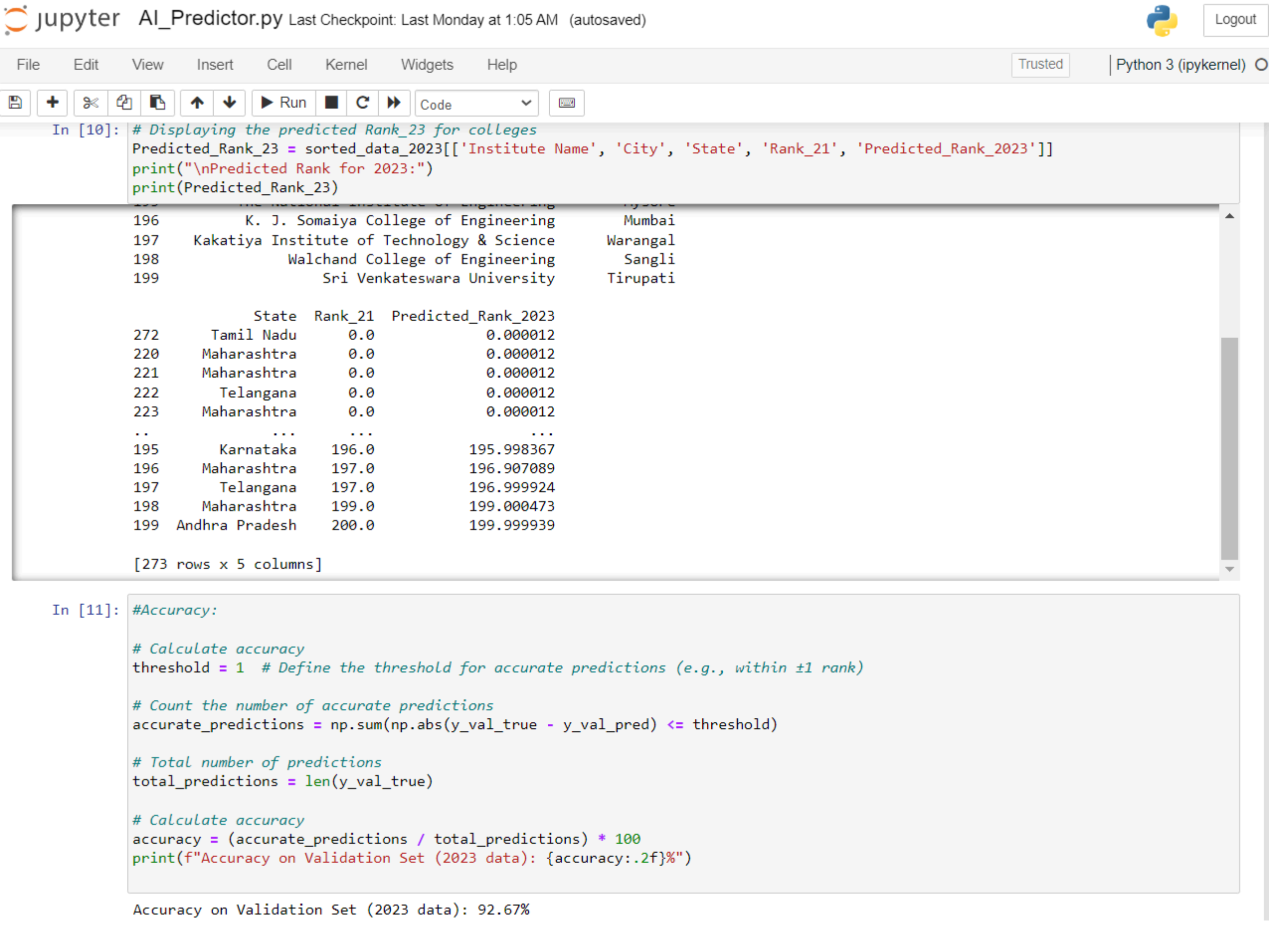
```
In [1]: !pip install xgboost

        Requirement already satisfied: xgboost in c:\users\sarve\anaconda3\lib\site-packages (2.0.3)
        Requirement already satisfied: scipy in c:\users\sarve\anaconda3\lib\site-packages (from xgboost) (1.10.0)
        Requirement already satisfied: numpy in c:\users\sarve\anaconda3\lib\site-packages (from xgboost) (1.23.5)

In [2]: import pandas as pd
        import numpy as np
        from xgboost import XGBRegressor
        from sklearn.model_selection import train_test_split
        from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

In [3]: # Load the data
        data = pd.read_csv("EngineeringRanking_Final.csv")

In [4]: # Handling missing values by filling with 0
        data.fillna(0, inplace=True)

In [5]: # Remove rows with missing target variable (Rank_21)
        data.dropna(subset=['Rank_21'], inplace=True)

In [6]: # Feature selection
        features = ['Score_21', 'Rank_21', 'TLR_21', 'RPC_21', 'GO_21', 'OI_21', 'Perception_21']
        X = data[features]
        y = data['Rank_21']  # Target variable for prediction

In [7]: # Splitting data into train, test, and validation sets
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Fig. Code Snippet 1



```
In [8]: # Train the model
        model = XGBRegressor()
        model.fit(X_train, y_train)

Out[8]:                      XGBRegressor
        XGBRegressor(base_score=None, booster=None, callbacks=None,
                     colsample_bylevel=None, colsample_bynode=None,
                     colsample_bytree=None, device=None, early_stopping_rounds=None,
                     enable_categorical=False, eval_metric=None, feature_types=None,
                     gamma=None, grow_policy=None, importance_type=None,
                     interaction_constraints=None, learning_rate=None, max_bin=None,
                     max_cat_threshold=None, max_cat_to_onehot=None,
                     max_delta_step=None, max_depth=None, max_leaves=None,
                     min_child_weight=None, missing=nan, monotone_constraints=None,
                     multi_strategy=None, n_estimators=None, n_jobs=None,

In [9]: # Get predictions for 2023 data
        data.dropna(subset=['Rank_21'], inplace=True)  # Remove rows with missing target values
        X_val = data[features]
        y_val_true = data['Rank_21']
        y_val_pred = model.predict(X_val)
        data['Predicted_Rank_2023'] = y_val_pred
        sorted_data_2023 = data.sort_values(by='Predicted_Rank_2023')
```

Fig. Code Snippet 2


jupyter AI_Predictor.py Last Checkpoint: Last Monday at 1:05 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

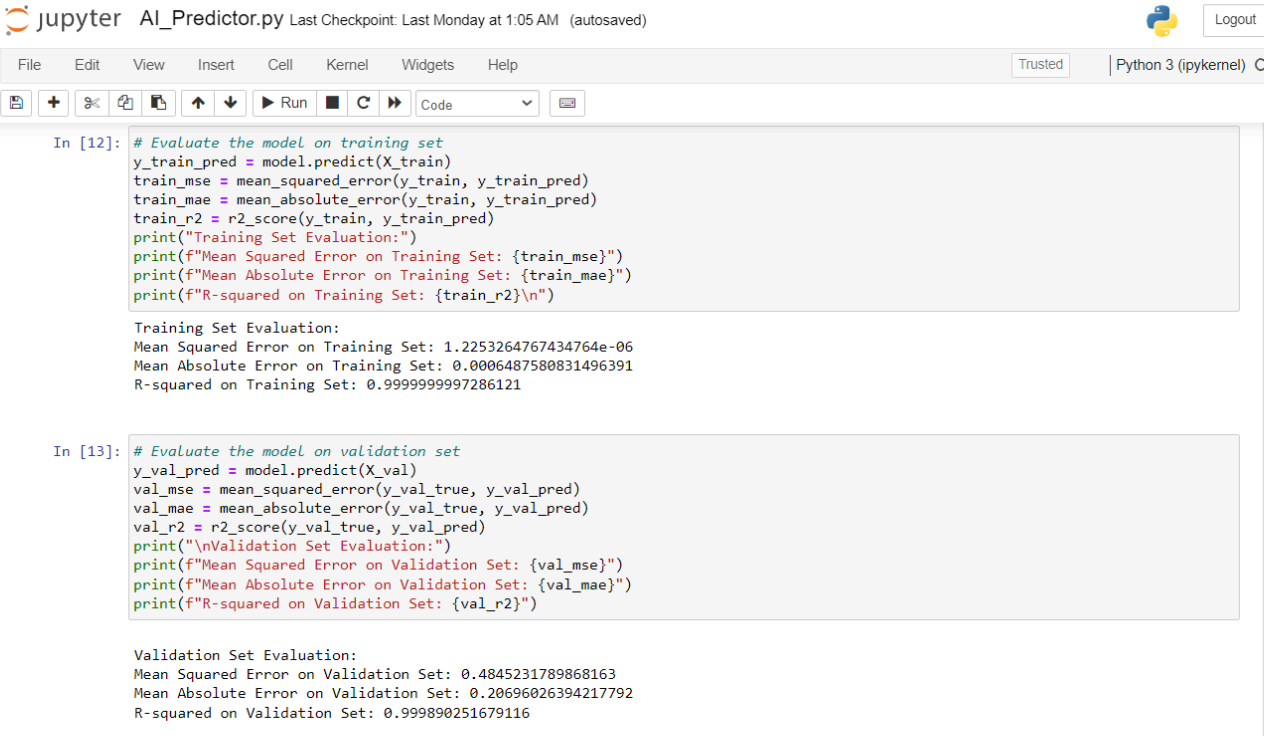
```
In [10]: # Displaying the predicted Rank_23 for colleges
         Predicted_Rank_23 = sorted_data_2023[['Institute Name', 'City', 'State', 'Rank_21', 'Predicted_Rank_2023']]
         print("\nPredicted Rank for 2023:")
         print(Predicted_Rank_23)
```

```
196          K. J. Somaiya College of Engineering       Mumbai
197    Kakatiya Institute of Technology & Science     Warangal
198               Walchand College of Engineering       Sangli
199                   Sri Venkateswara University     Tirupati

              State  Rank_21  Predicted_Rank_2023
272      Tamil Nadu      0.0             0.000012
220     Maharashtra      0.0             0.000012
221     Maharashtra      0.0             0.000012
222       Telangana      0.0             0.000012
223     Maharashtra      0.0             0.000012
..              ...      ...                  ...
195       Karnataka    196.0           195.998367
196     Maharashtra    197.0           196.907089
197       Telangana    197.0           196.999924
198     Maharashtra    199.0           199.000473
199  Andhra Pradesh    200.0           199.999939

[273 rows x 5 columns]
```

```
In [11]: #Accuracy:

         # Calculate accuracy
         threshold = 1  # Define the threshold for accurate predictions (e.g., within ±1 rank)

         # Count the number of accurate predictions
         accurate_predictions = np.sum(np.abs(y_val_true - y_val_pred) <= threshold)

         # Total number of predictions
         total_predictions = len(y_val_true)

         # Calculate accuracy
         accuracy = (accurate_predictions / total_predictions) * 100
         print(f"Accuracy on Validation Set (2023 data): {accuracy:.2f}%")
```

```
Accuracy on Validation Set (2023 data): 92.67%
```

Fig. Code Snippet 3


jupyter AI_Predictor.py Last Checkpoint: Last Monday at 1:05 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

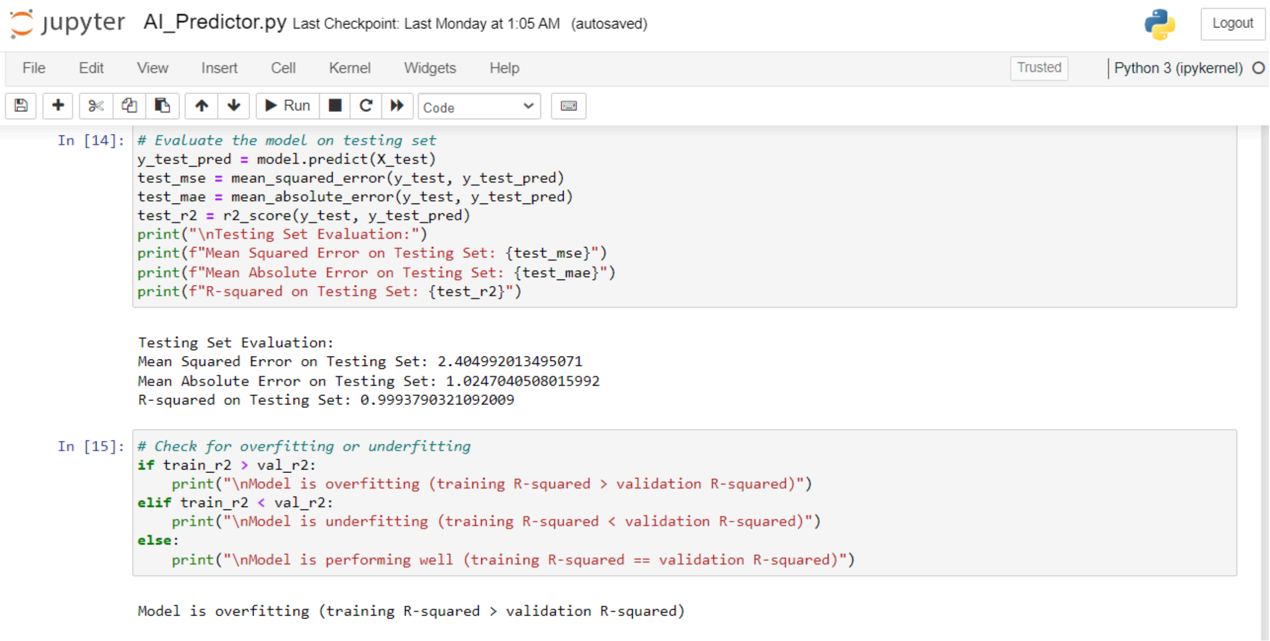
```
In [12]: # Evaluate the model on training set
         y_train_pred = model.predict(X_train)
         train_mse = mean_squared_error(y_train, y_train_pred)
         train_mae = mean_absolute_error(y_train, y_train_pred)
         train_r2 = r2_score(y_train, y_train_pred)
         print("Training Set Evaluation:")
         print(f"Mean Squared Error on Training Set: {train_mse}")
         print(f"Mean Absolute Error on Training Set: {train_mae}")
         print(f"R-squared on Training Set: {train_r2}\n")
```

```
Training Set Evaluation:
Mean Squared Error on Training Set: 1.2253264767434764e-06
Mean Absolute Error on Training Set: 0.0006487580831496391
R-squared on Training Set: 0.9999999997286121
```

```
In [13]: # Evaluate the model on validation set
         y_val_pred = model.predict(X_val)
         val_mse = mean_squared_error(y_val_true, y_val_pred)
         val_mae = mean_absolute_error(y_val_true, y_val_pred)
         val_r2 = r2_score(y_val_true, y_val_pred)
         print("\nValidation Set Evaluation:")
         print(f"Mean Squared Error on Validation Set: {val_mse}")
         print(f"Mean Absolute Error on Validation Set: {val_mae}")
         print(f"R-squared on Validation Set: {val_r2}")
```

```
Validation Set Evaluation:
Mean Squared Error on Validation Set: 0.4845231789868163
Mean Absolute Error on Validation Set: 0.20696026394217792
R-squared on Validation Set: 0.999890251679116
```

Fig. Code Snippet 4



In [14]:

```
# Evaluate the model on testing set
y_test_pred = model.predict(X_test)
test_mse = mean_squared_error(y_test, y_test_pred)
test_mae = mean_absolute_error(y_test, y_test_pred)
test_r2 = r2_score(y_test, y_test_pred)
print("\nTesting Set Evaluation:")
print(f"Mean Squared Error on Testing Set: {test_mse}")
print(f"Mean Absolute Error on Testing Set: {test_mae}")
print(f"R-squared on Testing Set: {test_r2}")
```

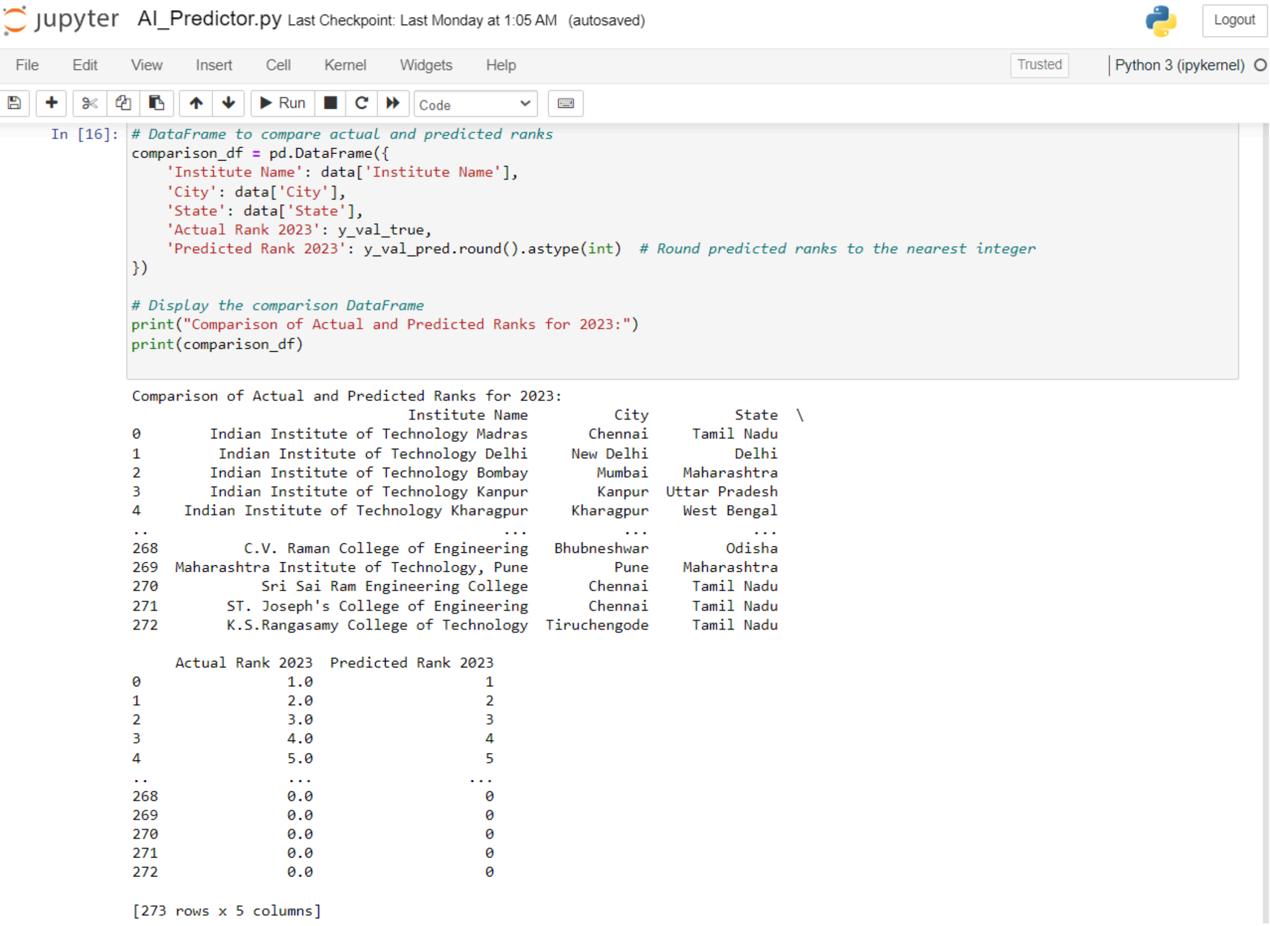
```
Testing Set Evaluation:
Mean Squared Error on Testing Set: 2.404992013495071
Mean Absolute Error on Testing Set: 1.0247040508015992
R-squared on Testing Set: 0.9993790321092009
```

In [15]:

```
# Check for overfitting or underfitting
if train_r2 > val_r2:
    print("\nModel is overfitting (training R-squared > validation R-squared)")
elif train_r2 < val_r2:
    print("\nModel is underfitting (training R-squared < validation R-squared)")
else:
    print("\nModel is performing well (training R-squared == validation R-squared)")
```

```
Model is overfitting (training R-squared > validation R-squared)
```

Fig. Code Snippet 5

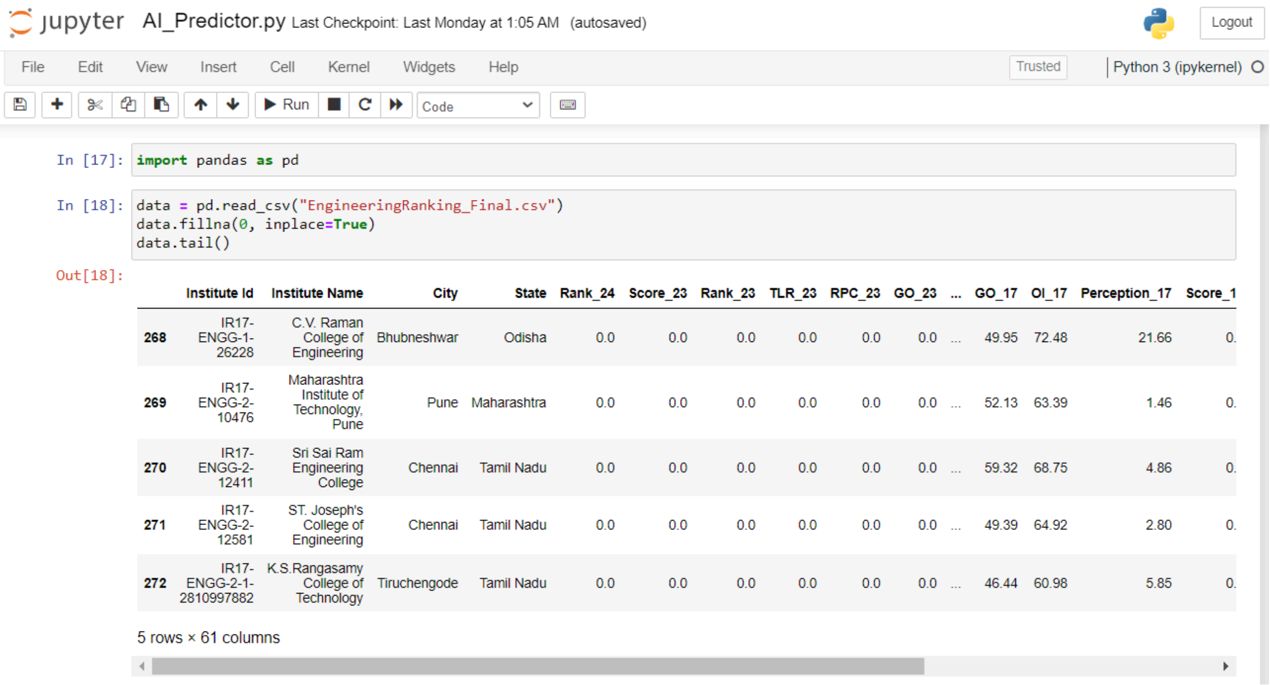


In [16]:

```
# DataFrame to compare actual and predicted ranks
comparison_df = pd.DataFrame({
    'Institute Name': data['Institute Name'],
    'City': data['City'],
    'State': data['State'],
    'Actual Rank 2023': y_val_true,
    'Predicted Rank 2023': y_val_pred.round().astype(int)  # Round predicted ranks to the nearest integer
})

# Display the comparison DataFrame
print("Comparison of Actual and Predicted Ranks for 2023:")
print(comparison_df)
```

```
Comparison of Actual and Predicted Ranks for 2023:
                                Institute Name          City          State  \
0        Indian Institute of Technology Madras       Chennai     Tamil Nadu   
1         Indian Institute of Technology Delhi     New Delhi          Delhi   
2        Indian Institute of Technology Bombay        Mumbai    Maharashtra   
3        Indian Institute of Technology Kanpur        Kanpur  Uttar Pradesh   
4     Indian Institute of Technology Kharagpur     Kharagpur    West Bengal   
..                                         ...           ...            ...   
268          C.V. Raman College of Engineering   Bhubneshwar         Odisha   
269  Maharashtra Institute of Technology, Pune          Pune    Maharashtra   
270            Sri Sai Ram Engineering College       Chennai     Tamil Nadu   
271        ST. Joseph's College of Engineering       Chennai     Tamil Nadu   
272        K.S.Rangasamy College of Technology  Tiruchengode     Tamil Nadu   

     Actual Rank 2023  Predicted Rank 2023  
0                 1.0                    1  
1                 2.0                    2  
2                 3.0                    3  
3                 4.0                    4  
4                 5.0                    5  
..                ...                  ...  
268               0.0                    0  
269               0.0                    0  
270               0.0                    0  
271               0.0                    0  
272               0.0                    0  

[273 rows x 5 columns]
```

Fig. Code Snippet 6


jupyter AI_Predictor.py Last Checkpoint: Last Monday at 1:05 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [17]: import pandas as pd
```

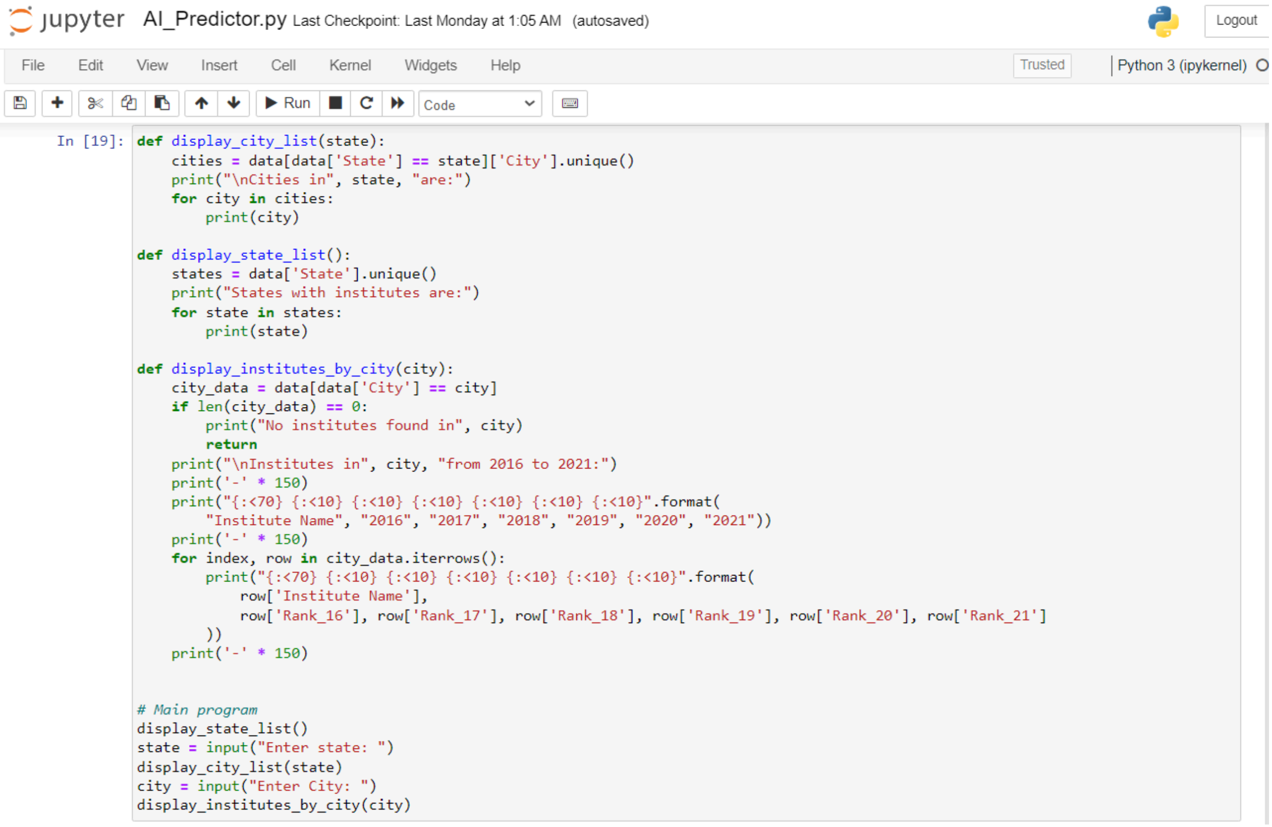
```
In [18]: data = pd.read_csv("EngineeringRanking_Final.csv")
         data.fillna(0, inplace=True)
         data.tail()
```

Out[18]:

	Institute Id	Institute Name	City	State	Rank_24	Score_23	Rank_23	TLR_23	RPC_23	GO_23	...	GO_17	OI_17	Perception_17	Score_1
268	IR17-ENGG-1-26228	C.V. Raman College of Engineering	Bhubneshwar	Odisha	0.0	0.0	0.0	0.0	0.0	0.0	...	49.95	72.48	21.66	0.
269	IR17-ENGG-2-10476	Maharashtra Institute of Technology, Pune	Pune	Maharashtra	0.0	0.0	0.0	0.0	0.0	0.0	...	52.13	63.39	1.46	0.
270	IR17-ENGG-2-12411	Sri Sai Ram Engineering College	Chennai	Tamil Nadu	0.0	0.0	0.0	0.0	0.0	0.0	...	59.32	68.75	4.86	0.
271	IR17-ENGG-2-12581	ST. Joseph's College of Engineering	Chennai	Tamil Nadu	0.0	0.0	0.0	0.0	0.0	0.0	...	49.39	64.92	2.80	0.
272	IR17-ENGG-2-1-2810997882	K.S.Rangasamy College of Technology	Tiruchengode	Tamil Nadu	0.0	0.0	0.0	0.0	0.0	0.0	...	46.44	60.98	5.85	0.

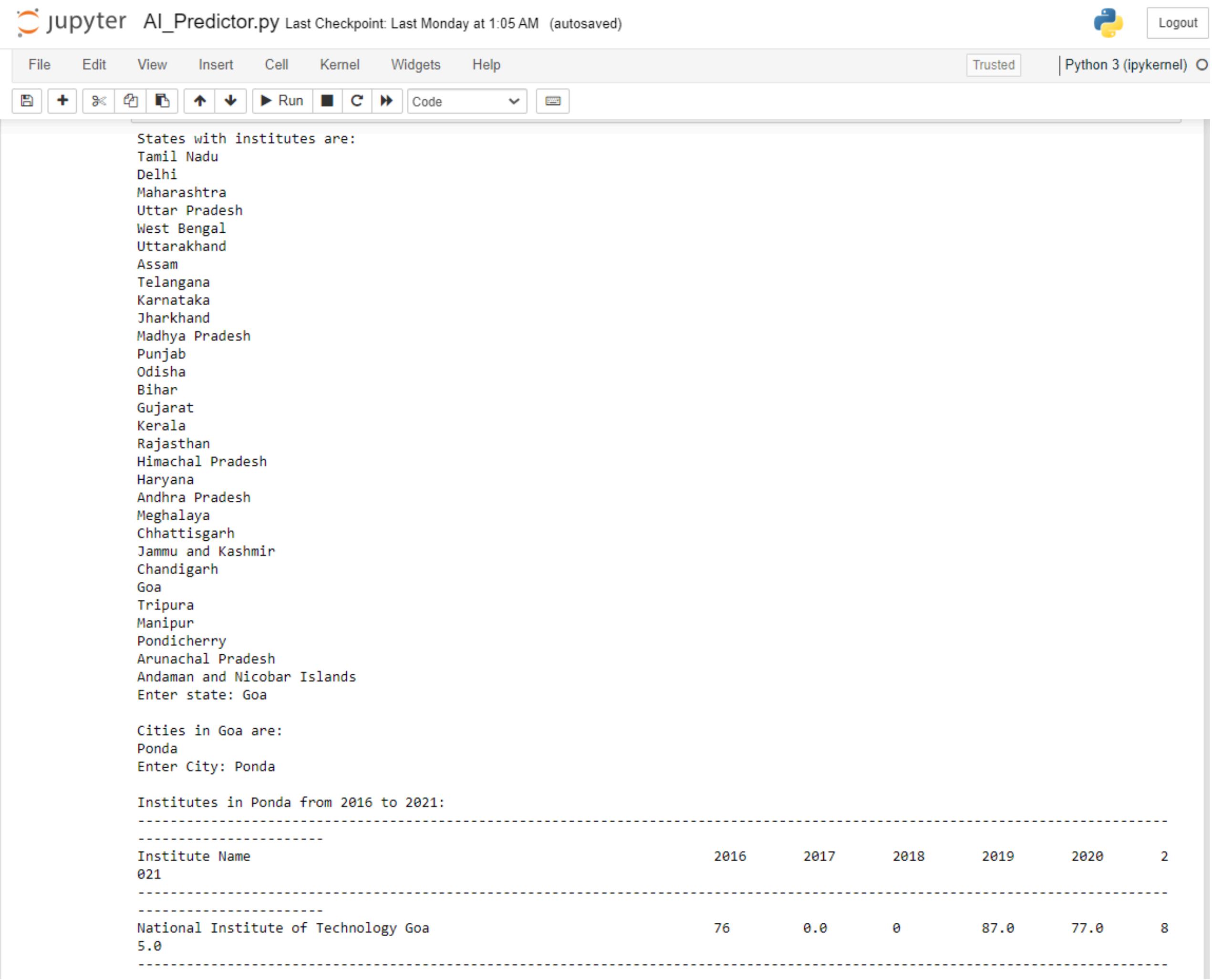
5 rows × 61 columns

Fig. Code Snippet 7


jupyter AI_Predictor.py Last Checkpoint: Last Monday at 1:05 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

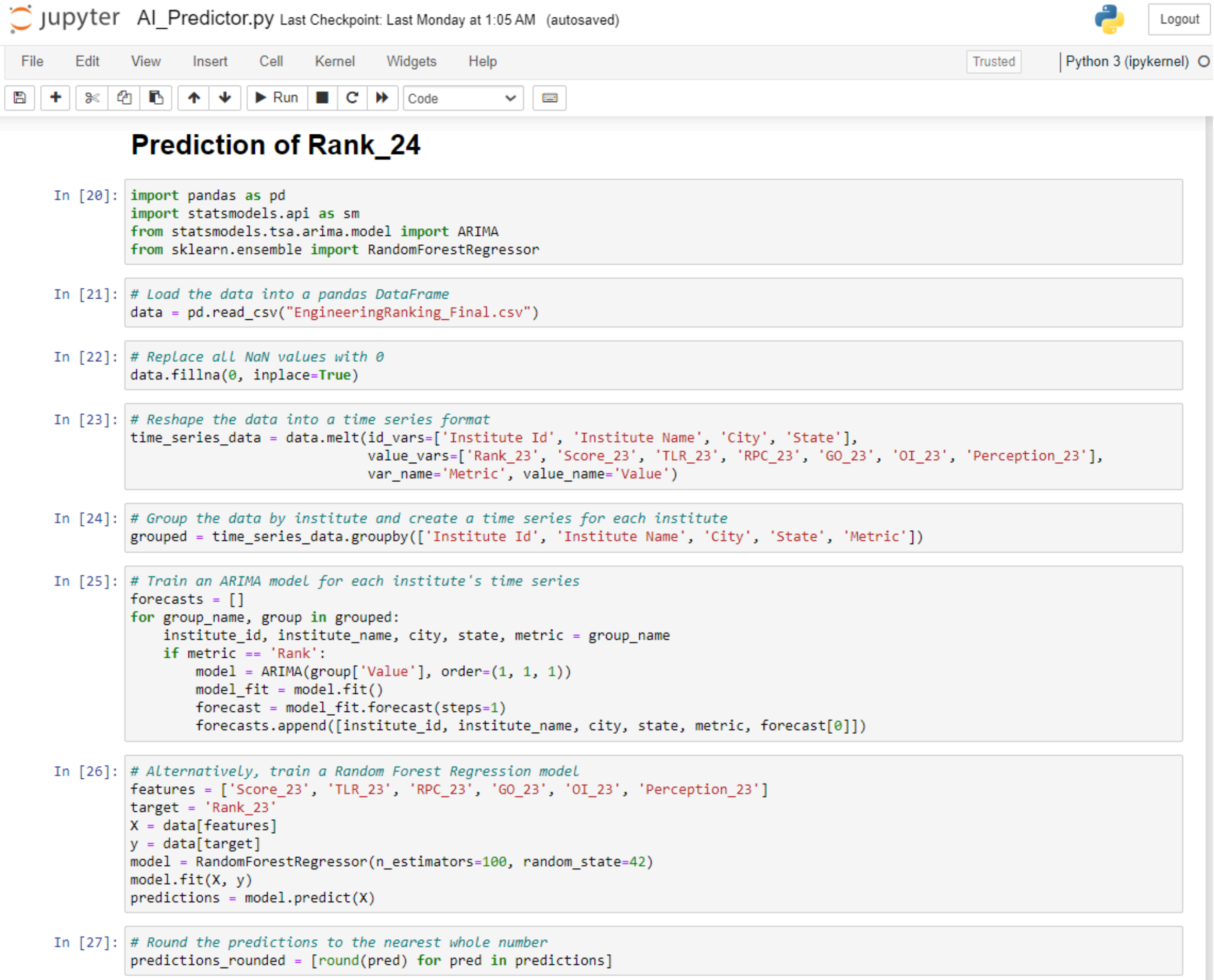
```
In [19]: def display_city_list(state):
             cities = data[data['State'] == state]['City'].unique()
             print("\nCities in", state, "are:")
             for city in cities:
                 print(city)

         def display_state_list():
             states = data['State'].unique()
             print("States with institutes are:")
             for state in states:
                 print(state)

         def display_institutes_by_city(city):
             city_data = data[data['City'] == city]
             if len(city_data) == 0:
                 print("No institutes found in", city)
                 return
             print("\nInstitutes in", city, "from 2016 to 2021:")
             print('-' * 150)
             print("{:<70} {:<10} {:<10} {:<10} {:<10} {:<10} {:<10}".format(
                 "Institute Name", "2016", "2017", "2018", "2019", "2020", "2021"))
             print('-' * 150)
             for index, row in city_data.iterrows():
                 print("{:<70} {:<10} {:<10} {:<10} {:<10} {:<10} {:<10}".format(
                     row['Institute Name'],
                     row['Rank_16'], row['Rank_17'], row['Rank_18'], row['Rank_19'], row['Rank_20'], row['Rank_21']
                 ))
             print('-' * 150)


         # Main program
         display_state_list()
         state = input("Enter state: ")
         display_city_list(state)
         city = input("Enter City: ")
         display_institutes_by_city(city)
```

Fig. Code Snippet 8



```
States with institutes are:
Tamil Nadu
Delhi
Maharashtra
Uttar Pradesh
West Bengal
Uttarakhand
Assam
Telangana
Karnataka
Jharkhand
Madhya Pradesh
Punjab
Odisha
Bihar
Gujarat
Kerala
Rajasthan
Himachal Pradesh
Haryana
Andhra Pradesh
Meghalaya
Chhattisgarh
Jammu and Kashmir
Chandigarh
Goa
Tripura
Manipur
Pondicherry
Arunachal Pradesh
Andaman and Nicobar Islands
Enter state: Goa

Cities in Goa are:
Ponda
Enter City: Ponda

Institutes in Ponda from 2016 to 2021:
------------------------------------------------------------------------------------------------------------------------------
------------------------
Institute Name                                                        2016       2017       2018       2019       2020       2
021
------------------------------------------------------------------------------------------------------------------------------
------------------------
National Institute of Technology Goa                                  76         0.0        0          87.0       77.0       8
5.0
------------------------------------------------------------------------------------------------------------------------------
```

Fig. Code Snippet 9



Prediction of Rank_24

In [20]:
```
import pandas as pd
import statsmodels.api as sm
from statsmodels.tsa.arima.model import ARIMA
from sklearn.ensemble import RandomForestRegressor
```

In [21]:
```
# Load the data into a pandas DataFrame
data = pd.read_csv("EngineeringRanking_Final.csv")
```

In [22]:
```
# Replace all NaN values with 0
data.fillna(0, inplace=True)
```

In [23]:
```
# Reshape the data into a time series format
time_series_data = data.melt(id_vars=['Institute Id', 'Institute Name', 'City', 'State'],
                             value_vars=['Rank_23', 'Score_23', 'TLR_23', 'RPC_23', 'GO_23', 'OI_23', 'Perception_23'],
                             var_name='Metric', value_name='Value')
```

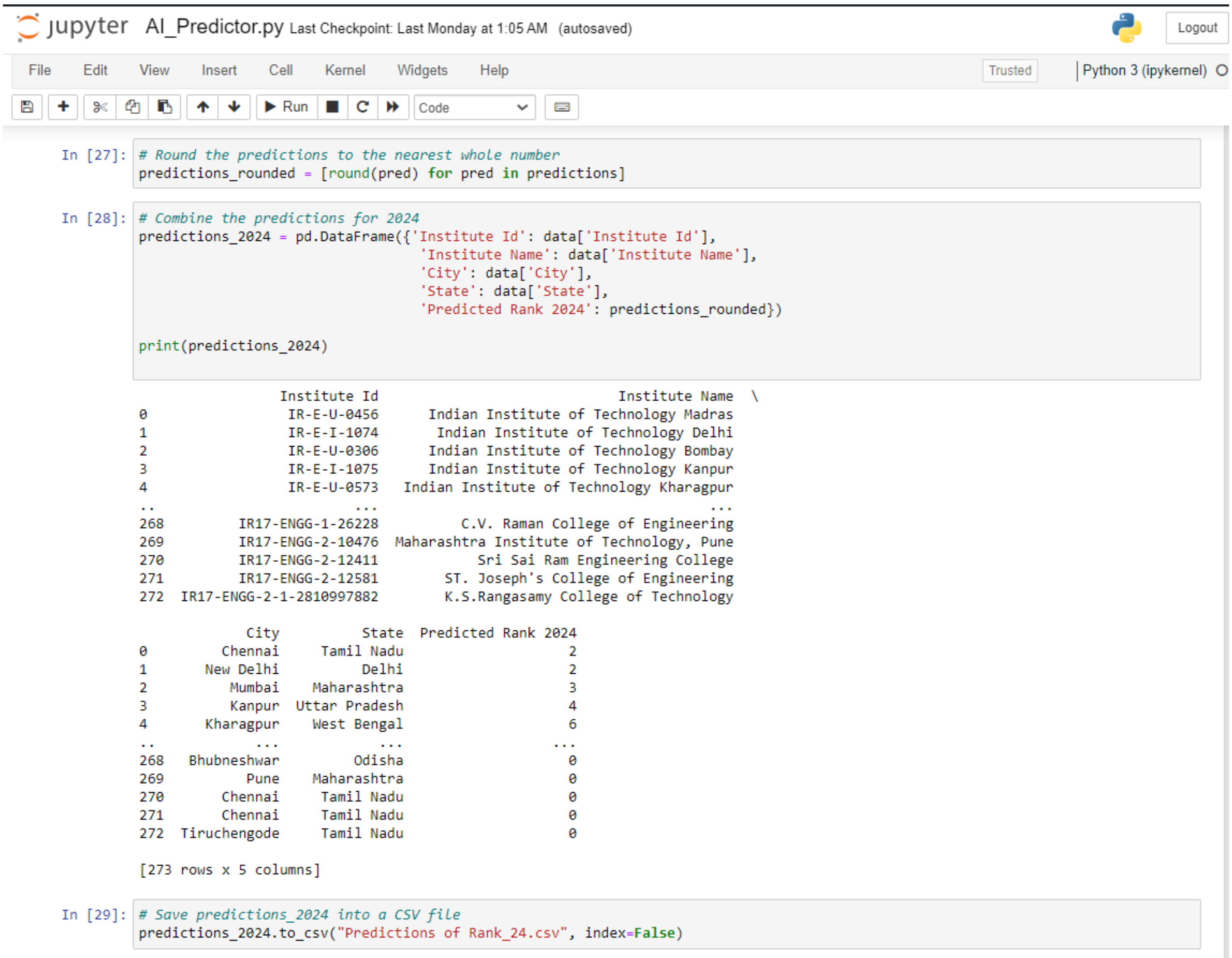
In [24]:
```
# Group the data by institute and create a time series for each institute
grouped = time_series_data.groupby(['Institute Id', 'Institute Name', 'City', 'State', 'Metric'])
```

In [25]:
```
# Train an ARIMA model for each institute's time series
forecasts = []
for group_name, group in grouped:
    institute_id, institute_name, city, state, metric = group_name
    if metric == 'Rank':
        model = ARIMA(group['Value'], order=(1, 1, 1))
        model_fit = model.fit()
        forecast = model_fit.forecast(steps=1)
        forecasts.append([institute_id, institute_name, city, state, metric, forecast[0]])
```

In [26]:
```
# Alternatively, train a Random Forest Regression model
features = ['Score_23', 'TLR_23', 'RPC_23', 'GO_23', 'OI_23', 'Perception_23']
target = 'Rank_23'
X = data[features]
y = data[target]
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X, y)
predictions = model.predict(X)
```

In [27]:
```
# Round the predictions to the nearest whole number
predictions_rounded = [round(pred) for pred in predictions]
```

Fig. Code Snippet 10



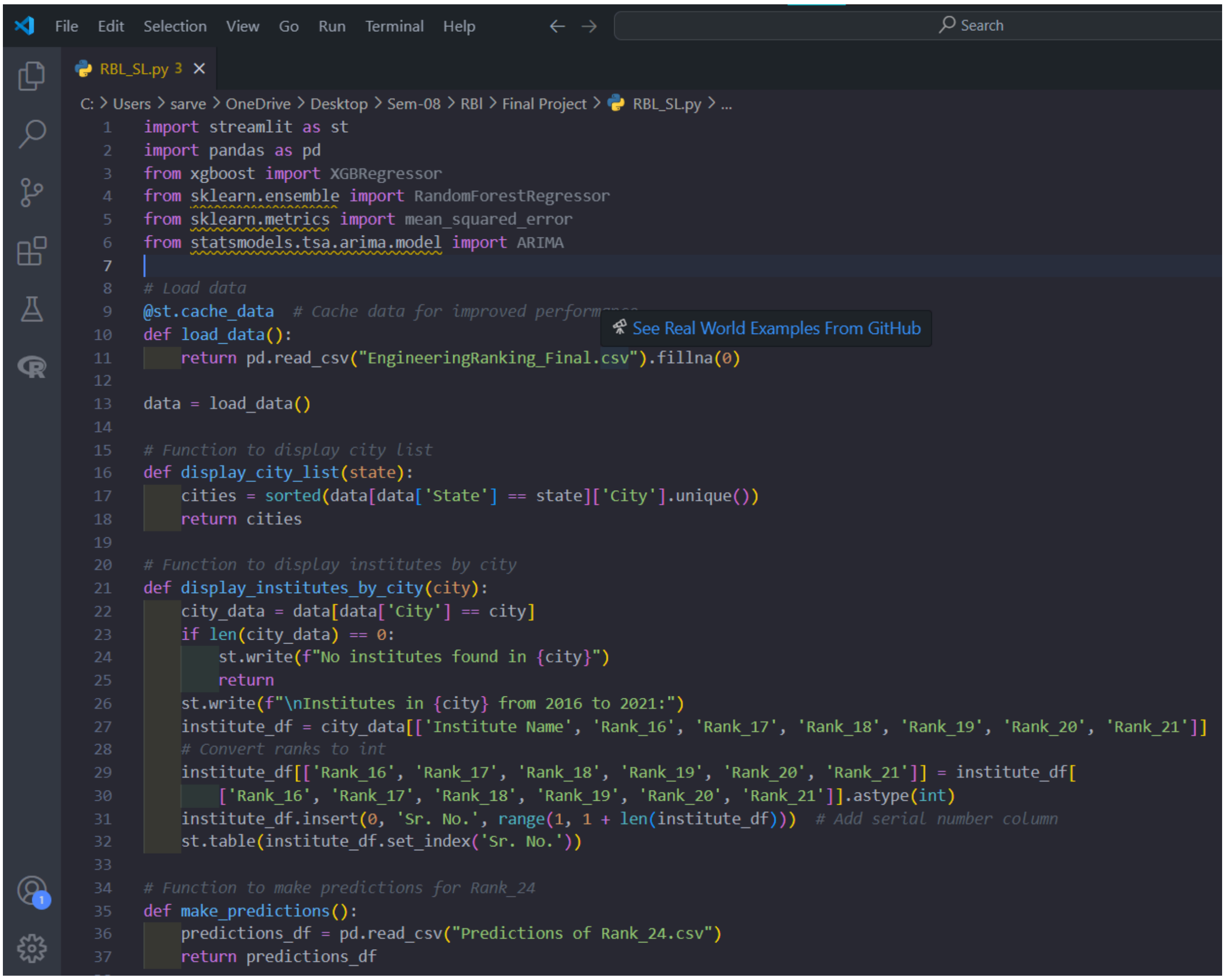
```
In [27]: # Round the predictions to the nearest whole number
         predictions_rounded = [round(pred) for pred in predictions]
```

```
In [28]: # Combine the predictions for 2024
         predictions_2024 = pd.DataFrame({'Institute Id': data['Institute Id'],
                                          'Institute Name': data['Institute Name'],
                                          'City': data['City'],
                                          'State': data['State'],
                                          'Predicted Rank 2024': predictions_rounded})

         print(predictions_2024)
```

```
                  Institute Id                               Institute Name  \
0                  IR-E-U-0456      Indian Institute of Technology Madras
1                  IR-E-I-1074       Indian Institute of Technology Delhi
2                  IR-E-U-0306      Indian Institute of Technology Bombay
3                  IR-E-I-1075      Indian Institute of Technology Kanpur
4                  IR-E-U-0573   Indian Institute of Technology Kharagpur
..                         ...                                        ...
268            IR17-ENGG-1-26228          C.V. Raman College of Engineering
269            IR17-ENGG-2-10476  Maharashtra Institute of Technology, Pune
270            IR17-ENGG-2-12411            Sri Sai Ram Engineering College
271            IR17-ENGG-2-12581          ST. Joseph's College of Engineering
272   IR17-ENGG-2-1-2810997882          K.S.Rangasamy College of Technology

            City          State  Predicted Rank 2024
0        Chennai     Tamil Nadu                    2
1      New Delhi          Delhi                    2
2         Mumbai    Maharashtra                    3
3         Kanpur  Uttar Pradesh                    4
4      Kharagpur    West Bengal                    6
..           ...            ...                  ...
268  Bhubneshwar         Odisha                    0
269         Pune    Maharashtra                    0
270      Chennai     Tamil Nadu                    0
271      Chennai     Tamil Nadu                    0
272  Tiruchengode     Tamil Nadu                    0

[273 rows x 5 columns]
```

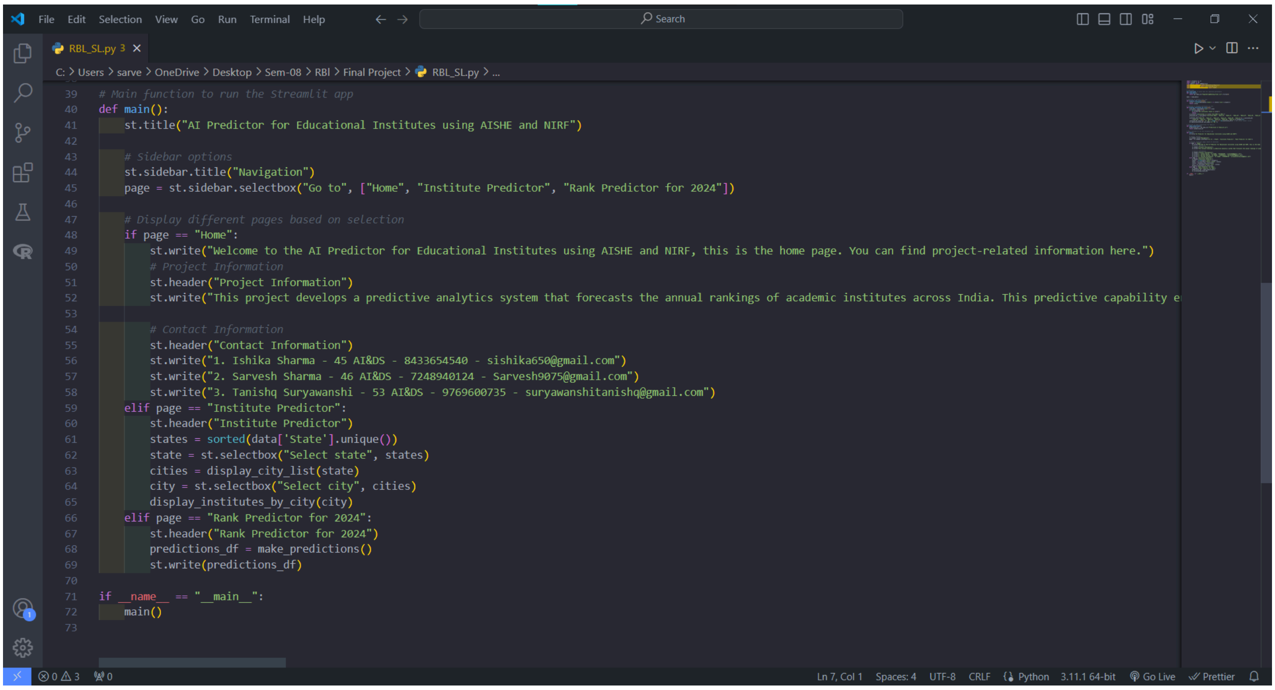
```
In [29]: # Save predictions_2024 into a CSV file
         predictions_2024.to_csv("Predictions of Rank_24.csv", index=False)
```

Fig. Code Snippet 11


File Edit Selection View Go Run Terminal Help
RBL_SL.py 3
C: > Users > sarve > OneDrive > Desktop > Sem-08 > RBI > Final Project > RBL_SL.py > ...
See Real World Examples From GitHub

```python
import streamlit as st
import pandas as pd
from xgboost import XGBRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
from statsmodels.tsa.arima.model import ARIMA

# Load data
@st.cache_data  # Cache data for improved performance
def load_data():
    return pd.read_csv("EngineeringRanking_Final.csv").fillna(0)

data = load_data()

# Function to display city list
def display_city_list(state):
    cities = sorted(data[data['State'] == state]['City'].unique())
    return cities

# Function to display institutes by city
def display_institutes_by_city(city):
    city_data = data[data['City'] == city]
    if len(city_data) == 0:
        st.write(f"No institutes found in {city}")
        return
    st.write(f"\nInstitutes in {city} from 2016 to 2021:")
    institute_df = city_data[['Institute Name', 'Rank_16', 'Rank_17', 'Rank_18', 'Rank_19', 'Rank_20', 'Rank_21']]
    # Convert ranks to int
    institute_df[['Rank_16', 'Rank_17', 'Rank_18', 'Rank_19', 'Rank_20', 'Rank_21']] = institute_df[
        ['Rank_16', 'Rank_17', 'Rank_18', 'Rank_19', 'Rank_20', 'Rank_21']].astype(int)
    institute_df.insert(0, 'Sr. No.', range(1, 1 + len(institute_df)))  # Add serial number column
    st.table(institute_df.set_index('Sr. No.'))

# Function to make predictions for Rank_24
def make_predictions():
    predictions_df = pd.read_csv("Predictions of Rank_24.csv")
    return predictions_df
```

Fig. Code Snippet 12


File Edit Selection View Go Run Terminal Help
Search
RBL_SL.py 3
C: > Users > sarve > OneDrive > Desktop > Sem-08 > RBI > Final Project > RBL_SL.py > ...
⊗ 0 ⚠ 3 0
Ln 7, Col 1 Spaces: 4 UTF-8 CRLF Python 3.11.1 64-bit Go Live Prettier

```python
# Main function to run the Streamlit app
def main():
    st.title("AI Predictor for Educational Institutes using AISHE and NIRF")

    # Sidebar options
    st.sidebar.title("Navigation")
    page = st.sidebar.selectbox("Go to", ["Home", "Institute Predictor", "Rank Predictor for 2024"])

    # Display different pages based on selection
    if page == "Home":
        st.write("Welcome to the AI Predictor for Educational Institutes using AISHE and NIRF, this is the home page. You can find project-related information here.")
        # Project Information
        st.header("Project Information")
        st.write("This project develops a predictive analytics system that forecasts the annual rankings of academic institutes across India. This predictive capability e

        # Contact Information
        st.header("Contact Information")
        st.write("1. Ishika Sharma - 45 AI&DS - 8433654540 - sishika650@gmail.com")
        st.write("2. Sarvesh Sharma - 46 AI&DS - 7248940124 - Sarvesh9075@gmail.com")
        st.write("3. Tanishq Suryawanshi - 53 AI&DS - 9769600735 - suryawanshitanishq@gmail.com")
    elif page == "Institute Predictor":
        st.header("Institute Predictor")
        states = sorted(data['State'].unique())
        state = st.selectbox("Select state", states)
        cities = display_city_list(state)
        city = st.selectbox("Select city", cities)
        display_institutes_by_city(city)
    elif page == "Rank Predictor for 2024":
        st.header("Rank Predictor for 2024")
        predictions_df = make_predictions()
        st.write(predictions_df)

if __name__ == "__main__":
    main()
```

Fig. Code Snippet 13

[B] Copy of published research paper and certificate

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.

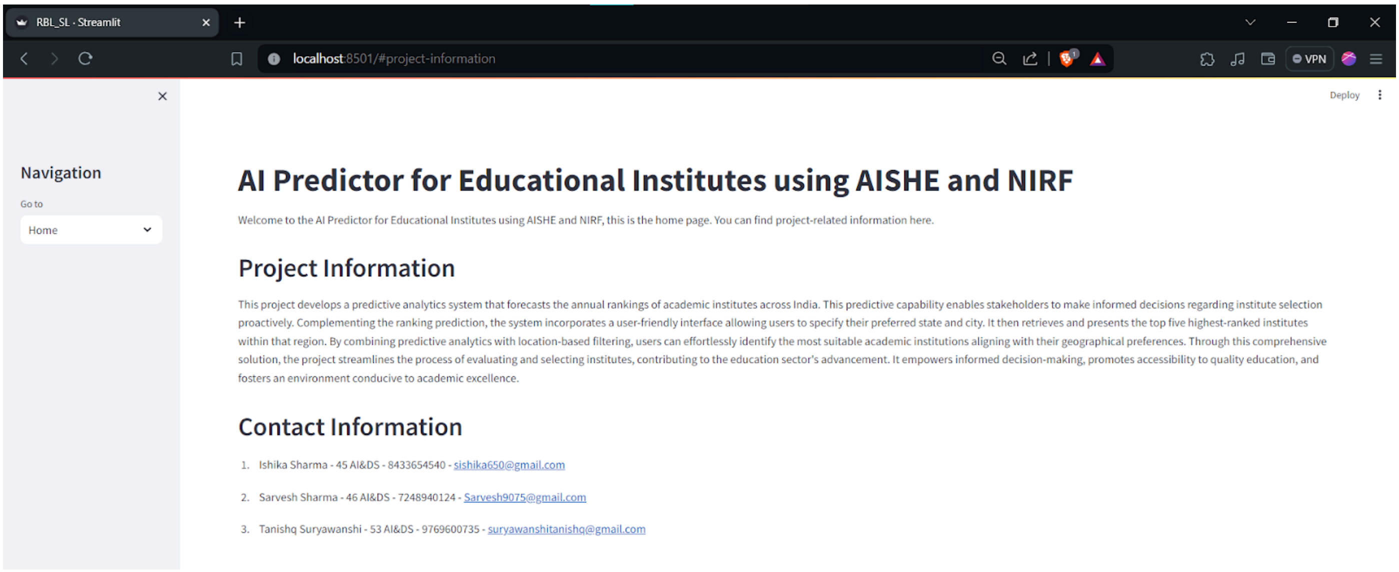
ARIMA for predicting 2024 data:

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.



AI Predictor for Educational Institutes using AISHE and NIRF

Welcome to the AI Predictor for Educational Institutes using AISHE and NIRF, this is the home page. You can find project-related information here.

Project Information

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.

Contact Information

1. Ishika Sharma - 45 AI&DS - 8433654540 - sishika650@gmail.com
2. Sarvesh Sharma - 46 AI&DS - 7248940124 - Sarvesh9075@gmail.com
3. Tanishq Suryawanshi - 53 AI&DS - 9769600735 - suryawanshitanishq@gmail.com

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

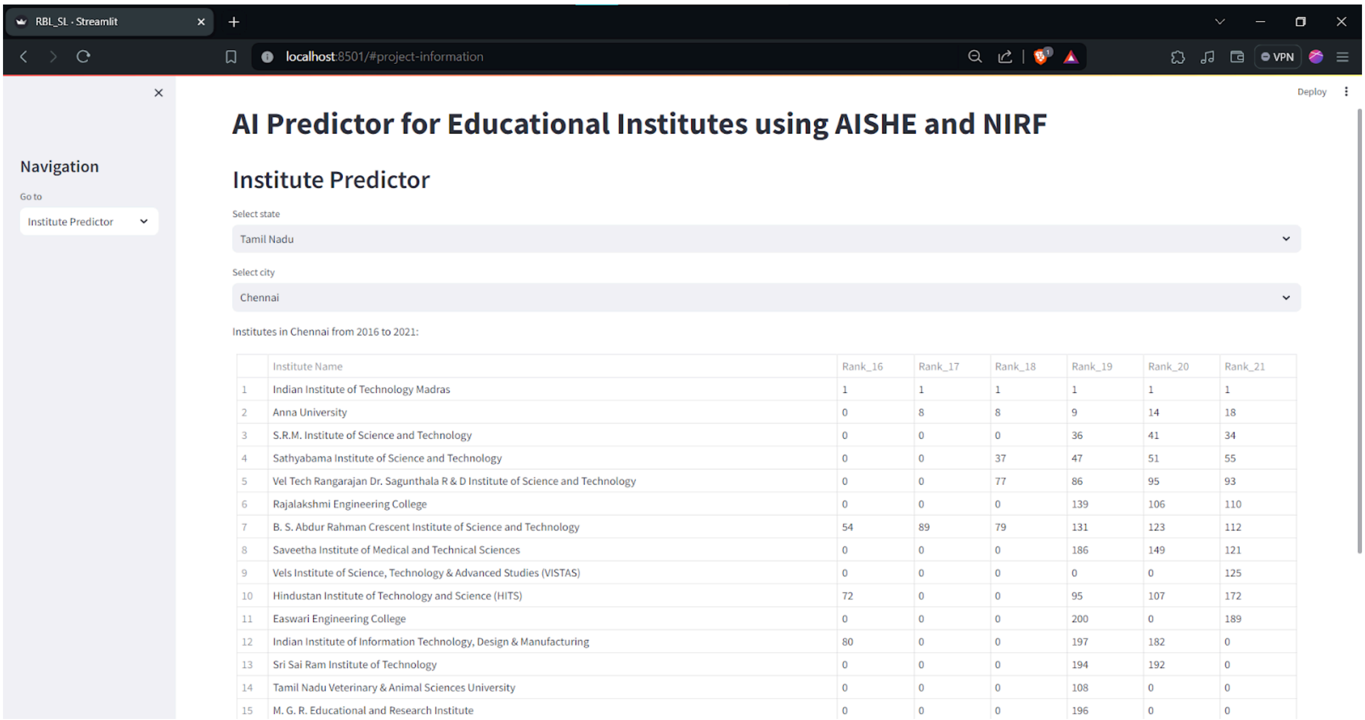
Testing Set	MSE	2.404
Validation Set	MSE	0.484
Model	Accuracy	92.67%

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.



AI Predictor for Educational Institutes using AISHE and NIRF

Institute Predictor

Go to: Institute Predictor

Select state: Tamil Nadu

Select city: Chennai

Institutes in Chennai from 2016 to 2021:

	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).

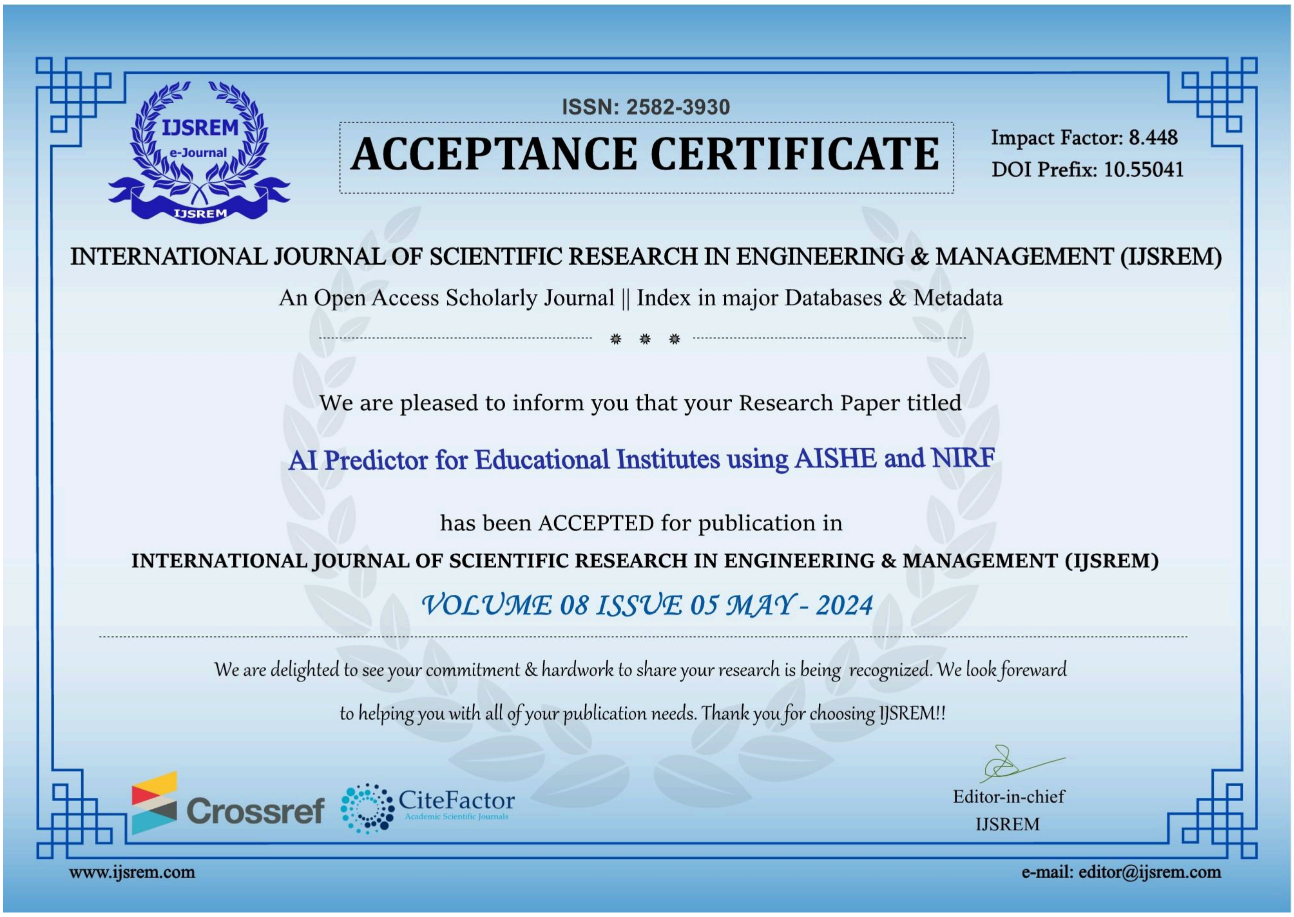
[8] EducationWiz Institute Rankings Methodology. https://www.educationwiz.com

[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

Certificate:



IJSREM e-Journal IJSREM

ISSN: 2582-3930

ACCEPTANCE CERTIFICATE

Impact Factor: 8.448
DOI Prefix: 10.55041

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING & MANAGEMENT (IJSREM)

An Open Access Scholarly Journal || Index in major Databases & Metadata

We are pleased to inform you that your Research Paper titled

AI Predictor for Educational Institutes using AISHE and NIRF

has been ACCEPTED for publication in

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING & MANAGEMENT (IJSREM)

VOLUME 08 ISSUE 05 MAY - 2024

We are delighted to see your commitment & hardwork to share your research is being recognized. We look foreward to helping you with all of your publication needs. Thank you for choosing IJSREM!!

Crossref CiteFactor Academic Scientific Journals

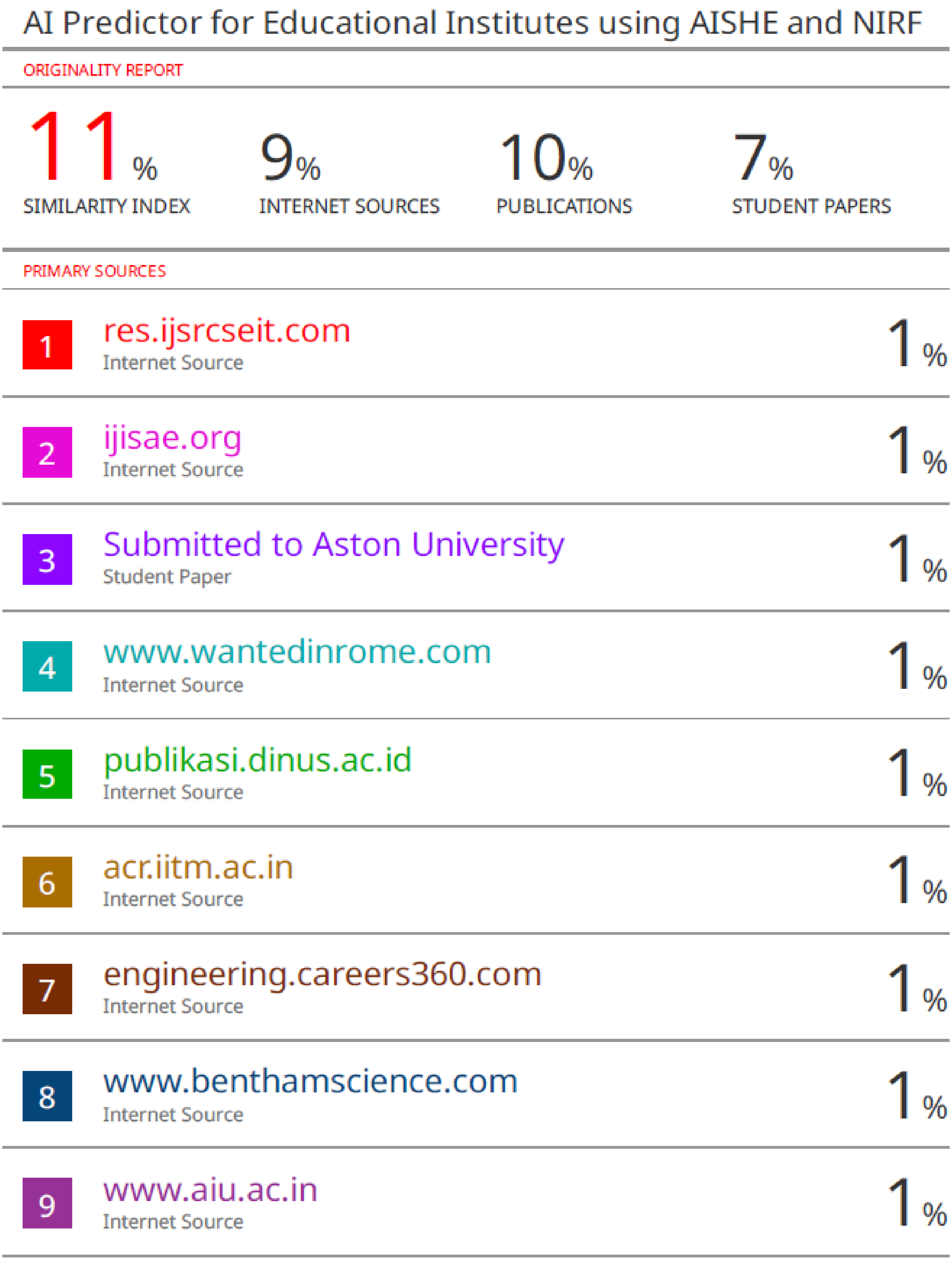
Editor-in-chief
IJSREM

www.ijsrem.com

e-mail: editor@ijsrem.com

Fig. Certificate

[C] Plagiarism check report:



AI Predictor for Educational Institutes using AISHE and NIRF

ORIGINALITY REPORT

11%	9%	10%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	res.ijsrcseit.com Internet Source	1%
2	ijisae.org Internet Source	1%
3	Submitted to Aston University Student Paper	1%
4	www.wantedinrome.com Internet Source	1%
5	publikasi.dinus.ac.id Internet Source	1%
6	acr.iitm.ac.in Internet Source	1%
7	engineering.careers360.com Internet Source	1%
8	www.benthamscience.com Internet Source	1%
9	www.aiu.ac.in Internet Source	1%

Fig. Plagiarism Report