Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250401006
Volume 4, Issue1 (January-April 2025), PP: 32-34.
www.indjcst.com
ISSN No: 2583-5300

Predictive Modeling for College Admission Using Machine Learning and Statistical Methods

Gopal V. Dose[1], Yuvraj M. Sanghai[2], Pranjal D. Sapkale[3], Dr. G. P. Potdar[4]
[1,2,3] Department of Computer Engineering, Pune Institute of Computer Technology (PICT), Pune, Maharashtra, India.
[4]Associate Professor, Department of Computer Engineering, Pune Institute of Computer Technology (PICT), Pune, Maharashtra, India.

To Cite this Article: *Gopal V. Dose[1], Yuvraj M. Sanghai[2], Pranjal D. Sapkale[3], Dr. G. P. Potdar[4], "Predictive Modeling for College Admission Using Machine Learning and Statistical Methods", Indian Journal of Computer Science and Technology, Volume 04, Issue 01 (January-April 2025), PP: 32-34.*

Abstract: *This study presents a predictive model for college admissions using machine learning and statistical techniques. While models such as Random Forest, XG Boost, and Multi-layer Perceptron (MLP) faced challenges due to data imbalance, a statistical model leveraging historical admission trends achieved 88% accuracy. This paper highlights the advantages of statistical methods in imbalanced datasets over conventional ML approaches.*
Keywords: *College Admission Prediction, Machine Learning, Statistical Model, Random Forest, XG Boost, MLP, Data Imbalance.*

I.INTRODUCTION

The process of securing admission to a college is a pivotal stage in the academic journey of students. This phase often induces anxiety and stress due to the factors that influence admission decisions, including academic performance, demographic information, and institutional quotas. As students apply to multiple institutions, often with little knowledge of their chances of acceptance, they face significant uncertainty, which can lead to wasted efforts and financial resources. Our group identified this challenge and sought to develop a predictive system to assist students in navigating the college admission process more effectively by forecasting their chances of acceptance.

With the recent advances in machine learning (ML) technologies, predictive models have become increasingly popular for forecasting outcomes across various domains, including education [1]. In this project, we explored how machine learning could be employed to predict college admissions based on historical student data, such as academic performance and demographic attributes. Our group initially focused on three prominent machine learning models - **Random Forest**, **XG-Boost**, and **Multi-layer Perceptron (MLP)**—to develop our predictive system. However, we quickly encountered challenges with data imbalance, which hindered the performance of these models [2].

Recognizing the limitations of the machine learning models, our group decided to implement an alternative approach: a **statistical model** based on historical admission trends. This statistical model successfully addressed the imbalance in the dataset and achieved an impressive accuracy of **88%**, compared to the **22%** accuracy of the machine learning models. This paper outlines the collaborative effort of our group to develop this predictive system, the methodologies we employed, and the broader implications of our findings for the field of college admission prediction.

The college admission process can be complex and stressful for students. Predicting admission chances based on academic performance and demographic data can help students make informed decisions. Machine learning provides a way to model and predict such outcomes based on historical data. This paper examines the use of machine learning algorithms such as Random Forest, XG Boost, and MLP to predict college admissions. However, the models struggled with the imbalanced nature of the dataset. As a result, a statistical model based on historical data was developed, which achieved significantly better performance, with an accuracy of 88%. This paper outlines the methodology, results, and implications of using this approach.

II.RELATED WORK

In educational contexts, predictive modeling has been widely used to anticipate student performance, dropout rates, and college admissions [3]. The existing body of work can be categorized into two major approaches: machine learning models and statistical models.

a) **Machine Learning in Education:** Several studies have applied machine learning algorithms like decision trees, logistic regression, and ensemble learning methods (e.g., Random Forest, Gradient Boosting) to predict academic success and college admissions [4]. These models analyze historical student data, including academic records, demographic features, and institutional data, to forecast outcomes. While machine learning models excel at identifying patterns in large datasets, they often struggle with imbalanced data where certain groups (e.g., students with higher academic scores) are overrepresented.

b) **Data Imbalance:** In educational datasets, data imbalance is a common issue [5]. For example, in college admission datasets,

high-performing students may dominate the data, making it difficult for machine learning models to generalize predictions for students with average or lower academic scores. Prior research has attempted to address this issue by applying techniques like oversampling and SMOTE (Synthetic Minority Over-sampling Technique). These methods aim to balance the dataset by either duplicating underrepresented instances (oversampling) or generating synthetic data points for the minority class (SMOTE). However, these approaches may introduce noise or lead to overfitting, particularly in smaller datasets [6].

c) Statistical Models: In response to the limitations of machine learning in handling imbalanced datasets, statistical models have emerged as a viable alternative. Unlike machine learning models, which rely on complex feature transformations and optimization techniques, statistical models are grounded in historical data and use simpler methods like percentile matching and group-based averages to predict outcomes [7]. In our study, the statistical model outperformed machine learning models, highlighting the utility of such models in educational prediction tasks.

III.METHODOLOGY

Developed and tested both machine learning and statistical models to predict college admissions. The dataset used in this project contains student records, including academic and demographic information.

a) Dataset Description: The dataset includes records from students applying to various colleges. The key features used in our predictive models are as follows:

Caste: The student's social category (e.g., Open, SC, ST, OBC), which is important due to caste-based reservation systems in college admissions.

Gender: The gender of the student, which may impact admission policies, particularly in gender-specific institutions or programs.

Branch: The academic discipline or branch (e.g., Computer Science, Mechanical Engineering) the student is applying for.

Location: The geographical region where the student resides. Regional quotas or preferences may influence admissions.

Percentage: The student's academic performance in terms of percentage scores. This is a key determinant of college admission eligibility.

The goal of our project was to predict whether a student would be admitted to a college based on these features.

b) Machine Learning Models: We initially focused on testing three machine learning models, each chosen for its ability to handle different types of data and capture non-linear relationships:

1. **Random Forest:** A widely used ensemble learning technique that constructs multiple decision trees and combines their outputs to improve accuracy [8].
2. **XG Boost:** An advanced gradient boosting algorithm known for its efficiency and accuracy in predictive tasks, especially in competitive environments [9].
3. **Multi-layer Perceptron (MLP):** A type of artificial neural network with multiple layers. MLPs are adept at capturing non-linear relationships in data, making them suitable for complex prediction tasks [10].

Despite their theoretical strengths, these models performed poorly due to the imbalanced nature of the dataset. High-performing students were overrepresented, leading to biased predictions and poor generalization for underrepresented groups. The highest accuracy achieved by these models was 22%.

c) Statistical Model Approach: After observing the limitations of machine learning models, we pivoted to a statistical model approach. The statistical model leverages historical data and uses a percentile-matching technique to predict admissions. The steps involved in building the statistical model are as follows:

1. **Grouping Data:** The dataset was grouped by key demographic and academic features (e.g., caste, branch, gender, location) to identify patterns in historical admission trends.
2. **Calculating Means:** The mean percentage for each group was calculated based on historical data. This allowed us to establish benchmarks for each category.
3. **Percentile Matching:** For each new student, their demographic and academic attributes were matched to the most similar historical group, and the model predicted admissions based on the average performance of that group.

The statistical model proved to be more effective than machine learning models, achieving an accuracy of **88%**. This suggests that simpler models, grounded in historical trends, can outperform more complex machine learning algorithms when the data is imbalanced.

IV.RESULTS

Table I compares the accuracy of the different models used in this study.

Table I: Model Comparison

Model	**Accuracy**
Random Forest	19%
Logistic Regression	20%
Decision Tree	22%
Statistical Model	88%

The statistical model clearly outperformed the machine learning models. This demonstrates the advantage of using a data-driven, percentile-based approach when dealing with imbalanced datasets

V.DATA ANALYSIS

The dataset revealed several imbalances that influenced the performance of the machine learning models:

Percentage Scores: Most students scored between 50% and 85%, with fewer students at the extreme low or high ends. This skew in the data made it difficult for the machine learning models to generalize predictions for students at either end of the spectrum.

Caste Distribution: The "Open" caste category accounted for 32.09% of the dataset, representing a significant overrepresentation compared to other caste categories. This imbalance contributed to biased predictions, as the models favoured Open-category students.

Branch Preferences: The data revealed that 51.74% of students applied to just three branches: Computer Science, Mechanical Engineering, and Civil Engineering, indicating a skew in branch preferences.

Geographic Distribution: Certain regions, such as Amravati, contributed a disproportionate number of students (3.10%), while smaller regions were underrepresented. This geographic skew impacted the predictive power of the machine learning models.

VI.DISCUSSION

The failure of machine learning models highlights the challenge of working with imbalanced datasets.

Techniques such as SMOTE and oversampling could be explored further. However, the statistical model's success indicates that simpler, data-driven approaches can sometimes outperform complex algorithms, especially when the data is skewed [11].

VII.FUTURE WORK

Improving the Statistical Model: While the statistical approach achieved better results than the machine learning models, adding more real-time data and refining the model's logic could further improve accuracy.

Real-Time Predictions: In future versions of this system, live updates on college cut-offs and quotas could be integrated to provide real-time predictions for students.

Advanced Machine Learning Models: Once more balanced and rich data becomes available, the project could revisit advanced machine learning models like Random Forest and XG Boost to refine the predictions.

VIII.CONCLUSION

In this study, we explored machine learning and statistical models for college admission prediction. The machine learning models, due to the imbalanced dataset, achieved a maximum accuracy of 22. In contrast, the statistical model, which used historical admission data, achieved an accuracy of 88. Future work could involve enhancing the dataset with additional features such as extracurricular activities, personal statements, and financial background, as well as exploring methods to balance the data before applying machine learning models.

References

[1] *S. Bhoite, K. Patil, and A. Sharma, "Predictive analytics of engineering and technology admissions," in Proc. IEEE Int. Conf. Computational Intelligence in Data Mining (ICCIDM), Dec. 2022, pp. 123-128.*

[2] *K. Kumari, A. Gupta, and P. Joshi, "CAPSLG: College Admission Predictor," in Proc. Int. Conf. on Automation and Systems Technology (ICAST), 2019, pp. 56-60.*

[3] *V. Sastry, P. Rao, and N. Gupta, "An automated prediction model for college admission system," Elem. Educ. Online, vol. 20, no. 7,pp. 234-239, 2021.*

[4] *A. Sivasangari, P. Suresh, and R. Jagan, "Prediction probability of university admission," in Proc. Int. Conf. on Computational Modeling and Communication (ICCMC), 2021, pp. 142-148.*

[5] *H. Mengash, "Using data mining techniques to predict student performance," IEEE Access, vol. 8, pp. 12415-12422, 2020.*

[6] *Z. Wang, L. Zhang, and X. Zhang, "Prediction of admission lines using machine learning," in Proc. Int. Conf. on Computer and Communication Technologies (ICCC), 2016, pp. 345-350.*

[7] *R. K. Sripath, M. Sharma, and V. K. Gupta, "Student career prediction using advanced ML techniques," Int. J. Eng. Technol. (IJET), vol. 15, no. 6, pp. 123-130, 2018.*

[8] *M. Nie, Y. Luo, and T. Zhang, "Advanced forecasting of career choices," Front. Comput. Sci., vol. 12, no. 3, pp. 567-574, 2018.*

[9] *A. Slim, R. Raj, and S. Kumar, "Predicting student enrollment based on characteristics," in Proc. Educational Data Mining (EDM), 2018, pp. 111-116.*

[10] *D. Kabakchieva, N. Georgiev, and D. Ivanova, "Analyzing university data for student profiles," in Proc. Educational Data Mining (EDM), 2011, pp. 87-92.*

[11] *A. Ragab, M. Ali, and S. Khaleel, "HRSPCA: Hybrid recommender for predicting admission," in Proc. Int. Symp. on Data Analysis (ISDA), 2012, pp. 203-208.*