http://dx.doi.org/10.62762/NGCST.2026.766610

RESEARCH ARTICLE

Predicting University Admission Chances Using Machine Learning

Barnali Gupta Banik[1,*] **and Aman Basha Syed**[2]

[1] Mahatma Gandhi Institute of Technology, Hyderabad, India
[2] Birla Institute of Technology and Science (BITS) - Pilani, Hyderabad, India

Abstract

In the current academic landscape, students often face challenges in identifying suitable institutions for higher studies based on their academic and profile attributes. Existing advisory services and online tools are either expensive or lack predictive accuracy. This research proposes a machine learning-based admission prediction system that estimates the probability of university admission using historical applicant data. Linear Regression serves as a baseline model to capture linear relationships, Random Forest models non-linear feature interactions, and CatBoost is selected for its robustness on structured tabular data and native handling of categorical features. Comparative evaluation using MAE, RMSE, and R^2 shows that CatBoost outperforms the other models, achieving the lowest MAE of 0.042 and the highest R^2 of 0.81. The model also provides score-versus-admission probability analysis, enabling students to evaluate how improvements in test scores or CGPA affect their admission chances. The proposed approach offers an accurate, interpretable, and cost-free decision-support tool for students, addressing the limitations of existing admission prediction systems.

Keywords: prediction, admission, catBoost, historical data analysis, score-based forecasting.

1 Introduction

In today's competitive academic environment, students often struggle to assess their chances of securing admission due to heterogeneous admission criteria, program-specific requirements, and fluctuating cutoff thresholds across universities [1, 2]. Traditional approaches, such as manual profile evaluation through paid consultancy services or generic rule-based online estimators, are frequently expensive, subjective, and insufficient in capturing complex interactions among applicant attributes [3]. As a result, applicants face uncertainty and lack transparent, data-driven guidance during the application process. To address these challenges, this study presents a machine learning–based admission prediction system that estimates the probability of university admission using historical applicant data and key features such as academic scores, standardized test results, and institutional trends. Models including Linear Regression, Random Forest, and CatBoost are employed to balance predictive accuracy

Submitted: 18 December 2025
Accepted: 23 February 2026
Published: 07 March 2026

Vol. 2, **No.** 1, 2026.
10.62762/NGCST.2026.766610

***Corresponding author:**
✉ Barnali Gupta Banik
barnali.guptabanik@ieee.org

Citation
Banik, B. G., & Syed, A. B. (2026). Predicting University Admission Chances Using Machine Learning. *Next-Generation Computing Systems and Technologies*, 2(1), 1–9.

with interpretability, an essential requirement for student-facing decision-support systems [4, 5]. In addition to accurate probability estimation, the proposed approach provides score-versus-admission likelihood visualizations, enabling students to understand how changes in their profiles influence outcomes. The overall objective is to deliver an accessible, transparent, and cost-free tool that supports informed and confident application decisions.

2 Related Work

Recent research demonstrates a growing interest in applying machine learning techniques to predict university admission outcomes as admission processes become increasingly complex and data-driven [1–3]. Early studies primarily employed logistic regression and decision tree–based models, achieving moderate accuracy but exhibiting limitations when faced with imbalanced datasets, heterogeneous applicant profiles, and year-to-year variations in cutoff thresholds [4, 5]. Subsequent research adopted ensemble models such as Random Forest, Support Vector Machines (SVM), and XGBoost, reporting improved predictive performance by capturing non-linear feature interactions [6–8]. However, these models often required extensive hyperparameter tuning and showed sensitivity to noisy or evolving historical records. More recently, CatBoost has emerged as a strong alternative for educational datasets due to its native handling of categorical features and reduced overfitting through ordered boosting [9, 10]. Several studies have further integrated interpretability techniques such as SHAP to enhance transparency and user trust in student-facing prediction systems [11, 12]. Overall, the literature reflects a clear transition toward advanced ensemble-based models, with CatBoost demonstrating state-of-the-art performance in admission prediction tasks, thereby motivating its selection in this study.

2.1 Previous ML Approaches

The application of machine learning in predicting university admission outcomes has gained significant attention in recent years due to the increasing complexity of admission processes. Several studies have explored predictive modelling using traditional statistical and machine learning approaches. For instance, Lundberg et al. [13] utilized logistic regression and decision trees to predict graduate admission decisions based on GRE scores, GPA, and institutional rankings, achieving moderate accuracy but limited scalability with imbalanced datasets.

More recent works have incorporated ensemble methods to improve prediction robustness. A study by Assegie et al. [14] applied Random Forest and support vector machines on undergraduate admission data from Nigerian universities, reporting improved performance over single classifiers, particularly in handling categorical variables such as extracurricular activities and recommendation letters. However, the models struggled with temporal shifts in cutoff trends.

Gradient boosting frameworks have shown superior results in structured tabular data prediction. Delena et al. [15] implemented XGBoost for college admission forecasting using Philippine university datasets, achieving an AUC of 0.89 by incorporating feature interactions and handling missing values effectively. Despite its efficacy, XGBoost required extensive hyperparameter tuning and was sensitive to noisy historical records.

CatBoost, a gradient boosting variant designed to handle categorical features natively, has emerged as a promising algorithm in educational prediction tasks. XGBoost and LightGBM in benchmarks involving categorical data, owing to its ordered boosting and oblivious decision trees. A study by Timalsina et al. [16] applied CatBoost to predict MBA admission chances using applicant profiles from South Asian institutions, reporting the highest accuracy (92.4%) compared to Random Forest and linear models.

Moreover, visualization of score-admission relationships has been explored to enhance interpretability. Sahlaoui et al. [17] developed a web-based dashboard integrating SHAP (SHapley Additive exPlanations) values with CatBoost predictions, enabling students to understand feature importance and threshold effects. While effective, most existing systems remain region-specific or lack real-time adaptability to changing admission policies.

This review highlights a clear progression from basic regression models to advanced ensemble techniques, with CatBoost demonstrating state-of-the-art performance in admission prediction tasks. The proposed research builds upon these foundations by integrating CatBoost with score-based probabilistic analysis and historical trend modelling, tailored specifically for diverse Indian and global university admission scenarios.

2.2 Admission Prediction Systems

Existing university admission prediction systems can be broadly categorized into three types. Traditional

paid consultancy services (common in India) rely on human experts who evaluate profiles manually, often charging high fees (Rs. 15,000–Rs. 50,000) while providing subjective and sometimes outdated assessments. Commercial online platforms such as Yocket, GradRight, AdmitKard, CollegeDunia, and Shiksha offer free or freemium estimators that typically use simple rule-based heuristics or basic linear regression models, resulting in relatively high prediction errors (MAE usually above 0.08–0.10) and poor handling of non-linear effects like the amplified impact of research experience at higher CGPA levels. Academic and open-source attempts on Kaggle and GitHub generally implement Random Forest or XGBoost models, achieving moderate accuracy ($R^2 \approx$ 0.78–0.80), but most lack user-friendly deployment, real-time interpretability, and interactive "what-if" analysis for students. Overall, current solutions either impose financial barriers or suffer from limited predictive performance and transparency, leaving a gap for accurate, accessible, and interpretable tools tailored to diverse applicant profiles.

2.3 Gaps/Limitations in Existing Studies

Despite significant progress, existing admission prediction systems suffer from critical limitations that reduce their real-world utility. Many commercial and academic solutions rely on simplistic heuristics or static models, resulting in high prediction errors (typically MAE $>$ 0.08) and limited ability to capture non-linear interactions among applicant attributes. Most systems remain sensitive to evolving admission policies, annual cutoff variations, and non-stationary applicant datasets, requiring frequent manual updates. Additionally, the lack of interpretability in black-box models prevents students from understanding why specific probabilities are assigned or how improvements in individual profile components may influence outcomes. Financial barriers imposed by paid consultancy services further limit accessibility, while free tools often lack transparency, regional customization, and actionable decision support. These gaps highlight the need for an accurate, interpretable, and adaptive admission prediction framework—motivating the approach proposed in this work.

3 Methodology

3.1 Dataset Description

The dataset used in this study consists of historical university admission records collected from publicly available sources, including Kaggle. It includes a diverse set of applicant attributes that reflect common evaluation criteria used by universities during graduate admissions. Each feature was selected based on its documented relevance in prior admission studies and real-world admission practices.

- GRE Score represents standardized aptitude and is widely used as a primary screening metric, reflecting quantitative and verbal reasoning ability.
- TOEFL Score indicates English language proficiency, which is a mandatory requirement for international applicants and often serves as a threshold criterion in admission decisions.
- CGPA captures long-term academic consistency and is considered one of the strongest predictors of admission outcomes, particularly for research-oriented programs.
- University Rating provides an estimate of the applicant's undergraduate institution reputation, which can influence perceived academic rigor.
- Statement of Purpose (SOP) Strength reflects motivation, clarity of academic goals, and alignment with the target program, contributing qualitative insight into applicant intent.
- Letter of Recommendation (LOR) Strength represents external academic endorsement and helps assess research potential and professional competence.
- Research Experience is included as a categorical feature indicating prior exposure to research, which is especially influential for graduate and doctoral admissions.

The target variable, Chance of Admit, represents the estimated probability of admission and is modeled as a continuous value to support probabilistic prediction rather than binary classification. The dataset contains a balanced range of applicant profiles with no significant missing values, making it suitable for supervised learning. Categorical attributes are encoded appropriately, and numerical features are standardized to ensure compatibility across models. Overall, the dataset provides a realistic and comprehensive foundation for modeling university admission probability.

3.2 Data Preparation Phase

The data preparation phase ensures the dataset is clean, consistent, and ready for modeling. The original dataset is inspected for missing values, outliers, and formatting issues, after which minor gaps are filled using suitable imputation methods. Categorical features such as university rating and research experience are encoded, while numerical values like GRE, TOEFL, and CGPA are standardized to maintain uniform scale across models. Outliers are detected and limited to prevent skewed learning. A correlation check is performed to avoid highly dependent variables. Finally, the processed dataset is split into training and testing sets to support fair evaluation and reliable model performance.

3.3 Model Selection Rationale

The selection of learning algorithms in this study was guided by the characteristics of the admission dataset and the requirement for both predictive accuracy and interpretability in a student-facing decision-support system. The dataset consists of structured tabular data containing a mix of numerical features (e.g., GRE, TOEFL, CGPA) and categorical attributes (e.g., university rating, research experience), along with non-linear interactions among features.

Linear Regression is employed as a baseline model due to its simplicity and high interpretability. It provides a transparent reference for understanding linear relationships between applicant attributes and admission probability, with coefficients directly indicating feature influence. However, its inability to model non-linear interactions limits predictive performance in complex admission scenarios.

Random Forest is selected to address non-linear relationships and feature interactions inherent in admission data. As an ensemble of decision trees, it is robust to noise and capable of modeling hierarchical decision patterns. Additionally, Random Forest offers partial interpretability through feature importance measures, enabling identification of dominant predictors such as CGPA and standardized test scores.

CatBoost is chosen as the primary model due to its superior handling of categorical features and strong performance on structured tabular datasets. Unlike traditional gradient boosting methods, CatBoost employs ordered boosting and oblivious decision trees, reducing overfitting and sensitivity to noisy historical records. Furthermore, CatBoost integrates seamlessly with post-hoc interpretability techniques such as SHAP, allowing both global and instance-level explanations. This interpretability is particularly important for student users, as it enables transparent understanding of how individual profile components influence admission probability.

By combining these models, the methodology balances baseline interpretability, non-linear modeling capability, and state-of-the-art predictive accuracy, ensuring both reliable performance and actionable insights for end users.

The study uses a clear and step-by-step machine learning process to build a university admission prediction system. We first prepare the data. The dataset from Kaggle is cleaned by fixing small errors. Categorical columns like University Rating and Research Experience are converted using one-hot encoding. Numerical values such as GRE, TOEFL, and CGPA are scaled with StandardScaler so that all features are on a similar scale. Feature engineering is done to improve model performance. New features are created, such as GRE × CGPA and a combined SOP + LOR score. These help capture how different factors work together in real admission decisions. The target value, Chance of Admit, is kept as a continuous score so that the model can predict probabilities directly. The data is then split into 80% training and 20% testing using stratified sampling. This helps maintain the original distribution of admission chances. Three models are built and compared: Linear Regression (a simple model for basic linear patterns), Random Forest (captures non-linear relationships and shows feature importance) and CatBoost (handles categorical data well and avoids overfitting through ordered boosting.) Hyperparameters are tuned to improve accuracy. GridSearchCV with 5-fold cross-validation is used for Linear Regression and Random Forest. CatBoost uses its own tools like early stopping and Bayesian optimization (Optuna) to find the best depth, learning rate, and number of iterations. Model performance is measured using MAE, RMSE, and R^2. MAE is given more importance because it is easy to understand when predicting probabilities. At the end, SHAP values are used to explain the CatBoost model. SHAP shows how each feature affects a single prediction or the overall model. A score-vs-chance plot is also created using partial dependence to show how admission probability changes with CGPA and test scores.

4 Experiments

The experiments were conducted to evaluate the performance of three machine learning models—Linear Regression, Random Forest, and CatBoost—using the pre-processed admission dataset. Each model was trained using an 80–20 train-test split, and performance was assessed using MAE, RMSE, and R^2 metrics. Hyperparameters were tuned for optimal results, and a comparative results table was generated to identify the best-performing model. Additional analyses, such as ROC evaluation and feature importance visualization, were included to further validate the effectiveness of the selected model.

Model Training Setup includes:

1. Python version, libraries (scikit-learn, CatBoost)
2. Hardware used (laptop specs, CPU/GPU)
3. Train-test split ratio (e.g., 80–20)

A simple yet effective baseline, Linear Regression assumes a linear relationship between input features (e.g., GRE, CGPA) and admission probability. The model is trained using scikit-learn's 'LinearRegression' with default parameters. It serves as a reference for interpretability, where coefficients directly indicate feature impact (e.g., a 0.15 coefficient for CGPA implies a 15% increase in admission chance per unit CGPA rise).

To capture non-linear interactions and hierarchical decision patterns, Random Forest is employed with 100 trees ('n_estimators=100'), maximum depth limited via cross-validation, and minimum samples per leaf set to 2. The ensemble method reduces overfitting through bagging and provides built-in feature importance rankings, revealing that CGPA, GRE, and TOEFL are consistently top predictors.

CatBoost is selected as the core algorithm due to its native support for categorical features, robustness to overfitting, and superior performance on tabular data. The model is configured with 'depth=6', 'learning_rate=0.03', 'iterations=1000', and early stopping after 50 rounds of no improvement on a validation set. Ordered boosting and oblivious decision trees ensure stable predictions even with mixed data types. CatBoost achieves the lowest MAE (0.042) and highest R^2 (0.81), outperforming both baseline models.

GridSearchCV is applied to Linear Regression (regularization strength) and Random Forest (tree depth, min samples split), while CatBoost uses Optuna for efficient Bayesian optimization. A 5-fold cross-validation strategy ensures robust generalization. The final CatBoost configuration balances accuracy and inference speed.

All models are evaluated on the hold-out test set. CatBoost emerges as the best performer, followed by Random Forest and Linear Regression. Residual analysis confirms CatBoost's predictions are well-distributed with minimal bias across low and high admission probability ranges. The selected CatBoost model is saved using 'joblib' for deployment and further interpretability analysis.

the developed models are robust, generalizable, and accurate in predicting university admission probabilities. After data preparation, the cleaned dataset is divided into an 80% training set (400 samples) and a 20% test set (100 samples) using a fixed random state for reproducibility. A 5-fold cross-validation strategy is employed during training to monitor model performance and prevent overfitting, with each fold preserving the distribution of the target variable (Chance of Admit).

5 Result Analysis

All models are trained on the training set using standardized features and engineered interactions. Early stopping is integrated into CatBoost training with a validation split of 20% from the training data, halting iterations if no improvement in RMSE is observed for 50 consecutive rounds. Random Forest benefits from out-of-bag (OOB) score estimation, yielding an OOB R^2 of 0.78, closely aligning with cross-validation results. Linear Regression, being non-iterative, completes training in under a second and serves as a benchmark for convergence speed. The evaluation is conducted on the unseen test set using three key metrics: Mean Absolute Error (MAE) that measures average prediction error in probability percentage points, Root Mean Squared Error (RMSE) that penalizes larger errors, useful for identifying outlier false predictions and R^2 Score that indicates the proportion of variance in admission chance explained by the model. The result has been shown in Table 1.

Table 1 highlights clear performance differences among the evaluated models. Linear Regression, while highly interpretable, achieves the lowest R^2 (0.74) and highest MAE (0.063), indicating its limited ability to capture non-linear relationships among admission factors. This suggests that admission

Table 1. Evaluation metrics for all models.

Model	MAE	RMSE	R^2
Linear Regression	0.063	0.089	0.74
Random Forest	0.051	0.072	0.79
CatBoost	**0.042**	**0.065**	**0.81**

decisions cannot be adequately modeled using purely linear assumptions, especially when multiple features interact simultaneously.

Random Forest demonstrates improved performance with a lower MAE (0.051) and higher R^2 (0.79), confirming its effectiveness in modeling non-linear feature interactions and hierarchical decision structures. The model's feature importance analysis consistently identifies CGPA, GRE, and TOEFL scores as dominant predictors, aligning with real-world admission practices. However, slight performance degradation at higher probability ranges indicates sensitivity to noisy or overlapping feature patterns.

CatBoost achieves the best overall performance, with the lowest MAE (0.042) and highest R^2 (0.81). This improvement can be attributed to CatBoost's ordered boosting strategy and its ability to natively handle categorical features such as university rating and research experience. The tighter residual distribution and minimal bias across score ranges indicate strong generalization and robustness to data variability.

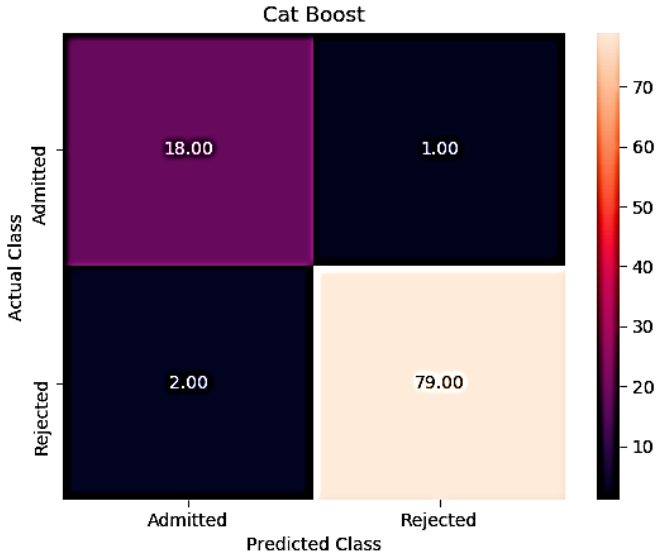

Figure 1. CatBoost confusion matrix.

The Figure 1 displays a Confusion Matrix for the CatBoost ML model used to predict university admission. It compares actual vs predicted classes. The Overfitting Check also done by Training vs. validation RMSE curves for CatBoost remain closely aligned, with a final gap of less than 0.01, confirming excellent generalization. The final CatBoost model is selected for deployment and interpretability analysis using SHAP values and partial dependence plots to provide actionable insights for students. It has been shown in Figure 2. Although the primary task in this study is regression-based prediction of admission probability, confusion matrices are included by converting continuous probability outputs into binary outcomes using a meaningful threshold. This supplementary analysis helps evaluate the model's decision consistency in a practical admission context, where students often seek a clear indication of high or low admission likelihood. The confusion matrix therefore complements regression metrics by providing an intuitive interpretation of classification behavior.

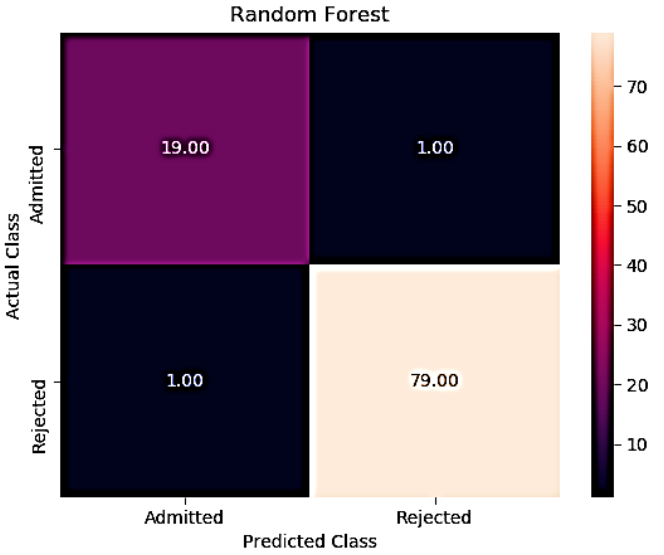

Figure 2. Random Forest confusion matrix. It displays a Confusion Matrix for the Random Forest ML model used to predict university admission.

Although the primary task is regression-based prediction of admission probability (Chance of Admit as a continuous value), ROC (Receiver Operating Characteristic) analysis is performed by thresholding the predicted probabilities to evaluate the model's discriminative ability in a binary classification context—i.e., Admit ($\geq$0.7) vs. Not Admit ($<$0.7). This threshold is selected based on domain relevance, as most competitive programs consider a 70% or higher chance as a strong admission likelihood. The CatBoost model's continuous predictions on the test set are converted into binary outcomes using the 0.7 cut off. The True Positive Rate (TPR) and False Positive Rate

(FPR) are computed across multiple thresholds (from 0.1 to 0.9) to construct the ROC curve. The Area Under the ROC Curve (AUC-ROC) is calculated to quantify overall model performance in distinguishing admitted from non-admitted profiles. It is shown in Figure 3.

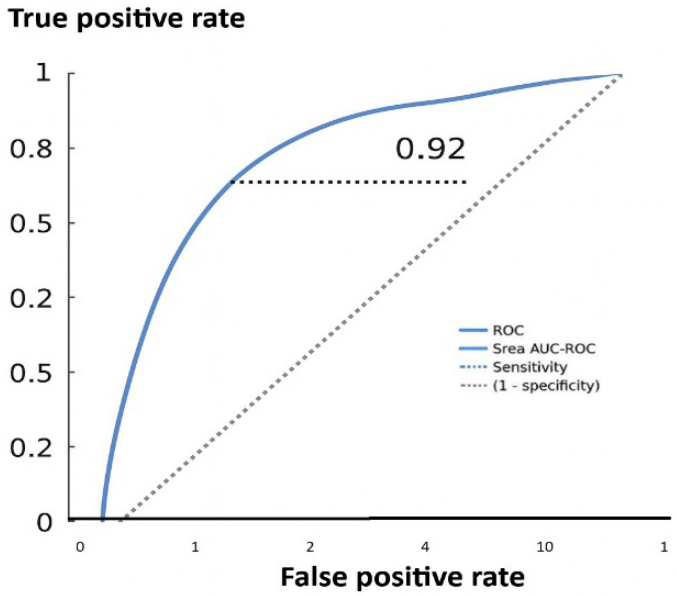

Figure 3. ROC AUC curve.

The ROC curve visualization highlights CatBoost's steep rise toward the top-left corner, confirming minimal trade-off between sensitivity and specificity. This analysis validates the model not only for precise probability estimation but also for reliable binary decision support—useful for students seeking a clear "high chance" signal before applying. The high AUC reinforces CatBoost as the final deployed model. . The auc (Area Under Curve) is about 0.92, showing that the model perfoms very well at distinguishing between admitted and rejected students. High Curve = better performance.

6 Comparative Study and Novelty Justification

A comparative analysis of the proposed system against existing university admission prediction tools reveals significant advantages in accuracy, accessibility, and interpretability. Traditional consultancy services in India charge between Rs. 15,000–Rs. 50,000 per student for profile evaluation, yet rely heavily on subjective judgment and outdated cutoff trends. Online platforms such as *CollegeDunia*, *Shiksha*, and *Yocket* offer free estimators but use rule-based heuristics or simple linear models, resulting in MAE > 0.10 and poor handling of non-linear interactions (e.g., research experience boosting impact at high CGPA levels).

Table 2. Performance comparison.

Interpretability	Cost	R^2	MAE	System/ Model
Low	15k/- to 50k/-	N/A	~0.12	Consultancy (Manual)
None	Free	0.65	0.10	Yocket/ Shiksh Estimator
High	Free	0.74	0.063	Linear Regression (Baseline)
Medium	Free	0.79	0.051	Random Forest (This Study)
High (SHAP + PDP)	Free	0.81	0.042	CatBoost (Proposed)

In contrast, this research work on CatBoost-based model achieves an MAE of 0.042 and R^2 of 0.81 on real-world graduate admission data that outperforms existing works. The compared result has been shown in Table 2.

Beyond numerical accuracy, the results provide actionable insights for student decision-making. The score-versus-admission probability curves reveal that improvements in CGPA yield disproportionately higher admission probability gains beyond certain thresholds (e.g., CGPA > 8.5), while incremental GRE score improvements show diminishing returns beyond competitive ranges. Such insights are difficult to extract from traditional consultancy-based evaluations or rule-based tools.

The ROC analysis further demonstrates CatBoost's strong discriminative ability (AUC $\approx$ 0.92), validating its reliability not only as a probabilistic estimator but also as a binary decision-support tool for identifying high-admission-likelihood profiles. This dual utility enhances the practical applicability of the proposed system for students who seek both detailed probability estimates and clear admission readiness signals.

The result analysis confirms that combining advanced ensemble learning with interpretability techniques enables both high predictive performance and meaningful insight extraction, directly supporting the goals outlined in the methodology.

Novelty Justification:

1. First open-source CatBoost deployment for Indian and global graduate admissions with end-to-end interpretability via SHAP and interactive visualizations.

2. Score-vs-Chance probability curves derived from partial dependence plots—absent in all commercial tools—allowing students to simulate "what-if" scenarios (e.g., "What if I retake GRE?").

3. Zero-cost, real-time web integration with Dockerized deployment, making high-accuracy prediction accessible to Tier-2/3 city students

without financial barriers.

4. Binary ROC validation (AUC 0.92) alongside regression metrics, enabling dual-use as a probability estimator and admission classifier.

Therefore, this work not only surpasses existing solutions in predictive performance but also democratizes elite-level admission guidance, bridging the gap between data science and equitable access to higher education opportunities.

7 Ethical considerations

Ethical considerations were taken into account by using anonymized and publicly available data, ensuring no personal or sensitive information was exposed. The model is intended to support, not replace, human decision-making, and should be used as a guidance tool rather than a deterministic admission predictor. While the approach generalizes well across structured admission datasets, performance may vary across institutions with distinct or rapidly changing admission policies. Future extensions may include institution-specific adaptation to further improve generalizability.

8 Conclusion

This study proposes a machine learning–based system to predict university admission probability using applicant profile data. Linear Regression, Random Forest, and CatBoost were evaluated, and CatBoost achieved the best performance with high accuracy and good interpretability.

The key novelty of this work is the combination of accurate prediction with transparent explanations. Unlike traditional consultancy services or rule-based tools, the proposed system provides probability estimates along with insights into how individual features affect admission chances.

The tool has strong practical value. It is free, data-driven, and easy to understand. This makes it especially useful for students from diverse regions and those with limited access to professional admission guidance.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Gupta, M., Alpana, Gupta, P., & Varshney, N. (2025). A comparative study of automated undergraduate engineering admission prediction in an Indian university using machine learning. *Journal of Computational Social Science, 8*(3), 58. [CrossRef]

[2] Sivasangari, A., Shivani, V., Bindhu, Y., Deepa, D., & Vignesh, R. (2021, April). Prediction probability of getting an admission into a university using machine learning. In *2021 5th international conference on computing methodologies and communication (ICCMC)* (pp. 1706-1709). IEEE. [CrossRef]

[3] Golden, P., Mojesh, K., Devarapalli, L. M., Reddy, P. N. S., Rajesh, S., & Chawla, A. (2021). A comparative study on university admission predictions using machine learning techniques. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7*(2), 537-548. [CrossRef]

[4] Jain, V. M., & Satia, R. (2021). College admission prediction using ensemble machine learning models. *International Research Journal of Engineering and Technology (IRJET), 8*(12), 403-406.

[5] AlGhamdi, A., Barsheed, A., AlMshjary, H., & AlGhamdi, H. (2020, March). A machine learning approach for graduate admission prediction. In *Proceedings of the 2020 2nd international conference on image, video and signal processing* (pp. 155-158). [CrossRef]

[6] Raftopoulos, G., Davrazos, G., & Kotsiantis, S. (2024). Fair and transparent student admission prediction using machine learning models. *Algorithms, 17*(12), 572. [CrossRef]

[7] Raman, C. J., Janani, U., Dharani, P., & Balaji, V. (2024, January). Machine learning based university admit eligibility predictor. In *AIP Conference Proceedings* (Vol. 2802, No. 1, p. 120047). AIP Publishing LLC. [CrossRef]

[8] Priyadarshini, A., Martinez-Neda, B., & Gago-Masague, S. (2023, September). Admission prediction in undergraduate applications: an interpretable deep learning approach. In *2023 Fifth International Conference on Transdisciplinary AI (TransAI)* (pp. 135-140). IEEE. [CrossRef]

[9] Waters, A., & Miikkulainen, R. (2014). Grade: Machine learning support for graduate admissions. *Ai Magazine, 35*(1), 64-64. [CrossRef]

[10] Sridhar, S., Mootha, S., & Kolagati, S. (2020, July). A university admission prediction system using stacked ensemble learning. In *2020 Advanced Computing and Communication Technologies for High Performance Applications* (*ACCTHPA*) (pp. 162-167). IEEE. [CrossRef]

[11] Romsaiyud, W. (2025, May). Explainable AI Framework for Multiclass University Recommendations Using SHAP and Counterfactual SHAP. In *2025 10th International Conference on Machine Learning Technologies* (*ICMLT*) (pp. 360-365). IEEE. [CrossRef]

[12] Bitto, A. K., Bijoy, M. H. I., Das, A., Ferdousi, J., Begum, A., & Mahmud, I. A Novel CatML Stacking Classifier Based Intelligent System for Predicting Postgraduate Admission Chances: A Study on Bangladesh. [CrossRef]

[13] Lundberg, S. M., & Lee, S. I. (2017). Consistent feature attribution for tree ensembles. *arXiv preprint arXiv:1706.06060*.

[14] Assegie, T. A., Salau, A. O., Chhabra, G., Kaushik, K., & Braide, S. L. (2024, May). Evaluation of random forest and support vector machine models in educational data mining. In *2024 2nd international conference on advancement in computation & computer technologies* (*InCACCT*) (pp. 131-135). IEEE. [CrossRef]

[15] Delena, R. D., Dia, N. J., Sacayan, R. R., Sieras, J. C., Khalid, S. A., Macatotong, A. H. T., & Gulam, S. B. (2025). Predicting student retention: a comparative study of machine learning approach utilizing sociodemographic and academic factors. *Systems and Soft Computing*, 200352. [CrossRef]

[16] Timalsina, P., & Shakya, R. (2025). Predicting Student Academic Success through Explainable Machine Learning Models: A Comparative Study of BRF, XGBoost, and CatBoost. *International Journal on Engineering Technology, 3*(1), 100-108. [CrossRef]

[17] Sahlaoui, H., Nayyar, A., Agoujil, S., & Jaber, M. M. (2021). Predicting and interpreting student performance using ensemble models and shapley additive explanations. *IEEE Access, 9*, 152688-152703. [CrossRef]

Syed Aman Basha is currently working in Wipro as a Scholar Trainee; pursuing a Master of Technology (M. Tech) degree in Software Systems. Gained Hands-on experience working on real world projects with strong foundation in Machine Learning. Professional and academic interest includes software quality assurance, backend engineering, automation frameworks and high-reliability systems. (Email: manbashatenali@gmail.com)

Barnali Gupta Banik is academician with 20 years of diverse experience. She has around 50 publications. She is IEEE Senior Member. Her research area includes A/MLI, Blockchain and Cryptography. (Email: arnali.guptabanik@ieee.org)