iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82712

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

BBAU Academic Admission Intelligence: An AI-Powered Decision Support System for Course Sustainability and Admission Optimization

Abhishek Rauniyar[1], G. Nitheesh Kumar[2], Ashish Verma[3], Abhishek Kumar[4], Astha Singh[5], Aman Kumar Sharma[6], Ashutosh Kumar[7]
Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Lucknow, India

Abstract: Public universities in India face unprecedented pressure to balance multiple competing constraints: rising demographic demand for higher education, constitutional equity mandates, fixed infrastructure capacity, and the policy demand for evidence-based governance. This paper presents BBAU Academic Admission Intelligence (BBAUI), an integrated AI-powered decision support system designed to address these challenges in real-time. The system combines machine learning for forecasting (Random Forest, XGBoost, ARIMA, Prophet), classical operations research for optimization (Linear Programming), and explainability frameworks (SHAP) to deliver strategic insights at the course level. Trained and validated on ten years of historical data from 106 courses across 23 departments, the system achieves forecasting accuracy exceeding 0.78 R-squared on held-out test data. The implementation spans 36 interactive analytical pages organized into four capability layers: descriptive analytics (dashboards and KPIs), predictive analytics (enrolment and outcome forecasting), prescriptive analytics (optimization recommendations), and equity auditing (reservation compliance and diversity monitoring). Results demonstrate that the data BBAU already collects is sufficient to predict next-year enrolment with acceptable accuracy, identify at-risk courses with explainable root causes, and recommend evidence-based interventions. The system runs on standard office hardware without requiring a database server, making it immediately deployable in resource-constrained academic settings. This research contributes to the emerging field of higher-education analytics in the Indian central-university context, where comparable decision-support systems remain rare.
Keywords: higher-education analytics, decision support system, machine learning, Random Forest, XGBoost, ARIMA, Prophet, SHAP, linear programming, sustainability index, equity audit, forecasting, course analytics, NEP-2020, Indian higher education

I. INTRODUCTION

A. Background

Higher education in India operates at the intersection of two structural pressures. On one hand, demographic expansion and rising aspirations have driven application volumes to unprecedented levels. On the other, Indian central universities operate under a constitutional mandate to provide access to socially and economically disadvantaged groups through reservation policies, with targets of 15% SC (Scheduled Caste), 7.5% ST (Scheduled Tribe), 27% OBC (Other Backward Classes), and 10% EWS (Economically Weaker Sections). These dual mandates collide with a third constraint: fixed infrastructure. Faculty strength, laboratory capacity and hostel beds cannot be expanded overnight. The result is that course-level admission decisions increasingly involve trade-offs: expanding a high-demand programme may strain laboratory resources; maintaining low intake in a declining programme wastes faculty load; closing a course impacts rural or minority representation.

These decisions are traditionally made on intuition and historical precedent. A department head receives pressure from the academic council to justify a course's continuation; a dean must decide whether to hire additional faculty; the registrar must forecast total applications to plan examination logistics. Each decision is made in isolation, without access to institution-wide patterns or predictive models. This paper describes BBAU Academic Admission Intelligence (BBAUI), a comprehensive AI-powered decision support system that brings data, models and stakeholder communication into a unified dashboard.

B. Motivation

Three observations motivated this work. First, vacant seats represent real institutional loss: wasted faculty effort, idle laboratory capacity and forgone fee revenue that could have funded infrastructure or scholarships.

Understanding and preventing vacancies directly improves institutional health. Second, equity auditing is tedious and error-prone when done manually. Reservation compliance requires tracking nine distinct demographic dimensions; identifying which courses are equity-weak requires reviewing hundreds of rows manually. Automating this task frees administrators to focus on remediation rather than data collection. Third, the National Education Policy 2020 (NEP-2020) explicitly mandates evidence-based governance and data-supported decision making. This system provides concrete tools for compliance with that mandate.

C. Problem Statement

BBAU's administrative offices answer four critical questions informally each admission cycle: (1) What is the current state? (2) Why is this happening? (3) What will happen next year? (4) What actions should we take? At present, these are answered through ad-hoc spreadsheet analysis, departmental anecdotes and precedent. The problems with this approach are well-established: data quality issues go unnoticed, trends are missed, decisions lack an explicit evidence trail, and equity blind spots persist. The project addresses this gap by building an integrated system that answers all four questions rigorously, transparently and at scale.

D. Contributions

This work contributes to three dimensions. Academically, it integrates a wide spectrum of techniques— supervised learning, time-series forecasting, optimization, explainability and equity auditing—into a coherent pipeline, demonstrating how multiple ML techniques can be orchestrated toward a real-world problem. Practically, it produces an institutional artefact that BBAU can deploy immediately without additional engineering. Policy-wise, it aligns with NEP-2020 requirements for evidence-based governance, multidisciplinary programme planning, and equity-aware data use.

II. LITERATURE REVIEW

A. Enrolment Forecasting

Enrolment forecasting is the most extensively studied problem in higher-education analytics. Classical approaches use univariate ARIMA models on institution-wide time series; recent work combines time- series baselines with machine-learning residual models to capture demographic and policy shifts. This system employs Random Forest and XGBoost regression with lagged features (prior-year enrolment, applications, rural/urban composition) to account for both temporal autocorrelation and cross-sectional heterogeneity across courses.

B. Programme Sustainability

The notion of programme sustainability is borrowed from portfolio-management literature (Boston Consulting Group matrix). The system computes a Sustainability Index that combines fill rate, placement rate, enrolment growth and retention, weighted to reflect institutional priorities. This composite score provides a single number that synthesizes multiple dimensions of programme health.

C. Equity and Inclusion Audits

Equity in Indian higher education is rooted in constitutional mandates around reservation. Beyond reservation category, gender representation and rural/urban access are increasingly recognized as critical dimensions. The system audits performance across nine equity dimensions: SC, ST, OBC, EWS, GEN (General), gender, rural, urban, and PwD (disability). This multi-dimensional approach surfaces equity gaps that single-metric audits miss.

D. Decision Support Systems

A decision support system, in the classical formulation by Sprague and Carlson, comprises three layers: a database, a model base, and a user interface joining them. This system follows that pattern: Pandas DataFrames serve as the database layer, scikit-learn and statsmodels provide the model base, and Streamlit plus Plotly form the interface. This architecture is increasingly common in academic analytics.

E. Explainability in Policy Contexts

Black-box models are increasingly unacceptable in public-sector decisions. The system incorporates SHAP (SHapley Additive exPlanations) values to attribute model predictions to constituent features. This addresses auditability and accountability requirements that arise when AI informs policy decisions affecting student access and resource allocation.

III. OBJECTIVES AND SCOPE

A. *Primary Objectives*

The project has seven concrete objectives: (1) forecast next-year enrolment, applications and demographic composition per course with explicit accuracy metrics; (2) compute a sustainability index and health score for every course; (3) audit equity across nine dimensions; (4) diagnose root causes for at-risk courses with specific interventions; (5) optimize seat allocation using linear programming; (6) provide scenario- simulation capabilities; (7) deliver all of the above through a single web application running on standard office hardware.

B. *Scope*

The system covers tabular admission, demographic, infrastructure and outcome data for ten academic years (2015–2024) across approximately 106 courses in 23 departments. It operates at course-level granularity (not individual student level) to respect privacy constraints. The implementation spans 36 analytical pages organized into four layers: policy-operational pages (Dashboard, Forecast, Equity, Sustainability), advanced ML-research pages (SHAP, Survival Analysis, Clustering), data-management pages (Data Dictionary, Quality Audit), and support pages (Methodology, User Manual, AI Assistant).

IV. PROPOSED SYSTEM ARCHITECTURE

A. *Logical Architecture*

The system comprises four logical layers: (1) Data Layer: Master Excel file (1,060 rows × 93 columns) and applicant-district sidecar (133,681 synthetic rows); (2) Preprocessing Layer: column normalization, missing-value imputation, feature engineering (fill rate, dropout rate, growth indices, etc.); (3) Modelling Layer: supervised regression (Random Forest, XGBoost, Linear Regression), time-series forecasting (ARIMA, Prophet), clustering (KMeans), optimization (Linear Programming), survival analysis (Kaplan-Meier); (4) User Interface Layer: Streamlit application with Plotly charts, Folium maps, and PDF export.

B. *Data Flow*

On application startup, the Excel master file is loaded with openpyxl via Pandas. Columns are normalized, numeric types are coerced, missing values are handled, and derived features are computed. This preprocessing is cached using Streamlit's @st.cache_data decorator; subsequent page renders reuse the preprocessed DataFrame. When a user selects a page, the corresponding model (e.g., Random Forest for forecasting, ARIMA for time-series) is trained or retrieved from cache. Results are rendered as interactive plots and KPI cards. Expensive operations (model training, LP optimization, SHAP computation) are cached with the filter state as the cache key, ensuring that users see instant results even when adjusting filters.

C. *Key Features*

Global filter propagation: all 36 pages respond to department and course selections made in the sidebar. Composite scoring: Sustainability Index, Health Score and Equity Index combine multiple metrics with configurable weights. Root-cause cascades: a deterministic rule system diagnoses the causes of vacancy risk and maps them to specific interventions. What-If simulator: users adjust intake, fee, faculty or rural- outreach levers and observe predicted impact. Geographic analysis: a Folium map shows district-level demand patterns and conversion rates. AI Assistant: a natural-language chatbot answers ad-hoc queries about the data with fallback to rule-based responses. PDF export: one-click generation of a strategic briefing document.

V. METHODOLOGY AND TECHNICAL APPROACHES

A. *Machine Learning Models*

1) Random Forest Regression: An ensemble of decision trees trained on bootstrap samples, with random feature subsets at each split. Hyperparameters: n_estimators=150, random_state=42. Used for per-course enrolment, application and demographic forecasting.
2) XGBoost Regression: A gradient-boosted ensemble with L1/L2 regularization. Hyperparameters: max_depth=5, learning_rate=0.1. Selected for its handling of non-linear interactions and robustness to outliers.
3) ARIMA(1,1,1): Classical time-series model fitted independently per course. Requires minimum four yearly observations. Produces intuitive confidence intervals for forecasts.
4) Facebook Prophet: Additive model with trend, seasonality and holiday effects. Robust to missing data and outliers. Used for institution-level forecasting.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

5) KMeans Clustering: Unsupervised clustering with k=4 on standardized KPI matrix. Used to identify archetypal course profiles (Cash Cows, Stars, Question Marks, Dogs).
6) Kaplan-Meier Survival Analysis: Non-parametric estimation of retention curves stratified by demographic cohort. Addresses dropout as a time-to-event problem rather than binary outcome.

B. Optimization and Allocation

Linear Programming formulates seat allocation as: maximize total enrolment subject to faculty load constraints (UGC norm: ≤25 students per faculty), laboratory utilization constraints, and floor/ceiling bounds on per-course intake. The solver uses scipy.optimize.linprog with the interior-point method.

C. Explainability

SHAP (SHapley Additive exPlanations) values are computed for tree-based models using TreeExplainer. Feature attributions satisfy efficiency, symmetry and additivity axioms from cooperative game theory. Results are presented as mean-absolute-SHAP importance bars and SHAP beeswarm plots showing the direction and magnitude of feature effects.

D. Training and Evaluation Protocol

Data is split by year: rows with YEAR ≤ 2023 form the training set; YEAR = 2024 forms the held-out test set. Models are trained on the feature matrix and evaluated with mean absolute error (MAE), root mean squared error (RMSE) and R-squared.

Walk-forward backtesting evaluates a model trained on years <2019 on year 2019, then trained on years <2020 on year 2020, etc., providing a more conservative accuracy estimate. All predictions use learned feature relationships; no novel features are engineered for the test set.

VI. DATASET DESCRIPTION AND PREPROCESSING

A. Master Dataset

The master dataset contains 1,060 rows (one per course-year combination) and approximately 93 columns organized thematically: course metadata (name, department, level, duration, fees), admission data (applications, enrolled, dropouts), reservation data (SC, ST, OBC, EWS, GEN counts and fill rates), demographic data (gender, rural/urban, PwD, minority), infrastructure (faculty, labs, classrooms, hostels), outcomes (placement rate, average package, pass rate, higher-studies count), and state-representation columns (students from 17 major Indian states).

B. Preprocessing Pipeline

Preprocessing applies the following transformations in order: (1) load with openpyxl, (2) normalize column names, (3) coerce numerics with pd.to_numeric(errors='coerce'), (4) compute fill rate = enrolled / intake (clipped to [0,1]), (5) compute dropout rate, (6) derive lag features (prior-year rural, urban, applications), (7) compute Sustainability Index as weighted average of fill, placement, growth, retention, (8) classify vacancy risk into High (fill < 0.60), Medium (fill 0.60–0.85), Low (fill ≥ 0.85), (9) fill remaining NaNs with column- wise medians.

C. Feature Engineering

Engineered features are designed to capture two phenomena: serial correlation (lag features capture the fact that this year's applications strongly predict next year's applications) and cross-sectional heterogeneity (rural_ratio, placement_rate capture differences between courses).

The feature set used for supervised learning comprises: LAG_APPS, LAG_RURAL, LAG_URBAN, LAG_ENROLLED, FILL_RATE, DROPOUT_RATE, PLACEMENT_RATE, RURAL_RATIO, FACULTY_LOAD_RATIO, FEMALE_PCT, and SUSTAINABILITY_INDEX.

D. Data Quality Audit

The system includes a live Data Quality page that reports: (1) percentage missing values per column, (2) duplicate-row count, (3) univariate outliers (values >3σ from mean), (4) numeric type violations, (5) logical inconsistencies (enrolled > intake). All columns are documented in an interactive Data Dictionary.

VII. SYSTEM IMPLEMENTATION

A. Architecture Overview

The system is implemented as a single Python file (bbau_app.py) of approximately ten thousand lines. This monolithic architecture was chosen for simplicity of deployment: a single file plus a requirements.txt and an Excel data file can be run with 'streamlit run bbau_app.py' on any machine with Python 3.10+. The codebase uses Streamlit's page-navigation pattern and session-state management for persistent state across reruns.

B. Page Modules

The 36 pages are organized into eight functional groups: (1) Insights pages (Dashboard, Course Health Monitoring, Sustainability, Equity, Regional Demand); (2) Forecasting pages (Forecast 2026, Model Comparison, ARIMA, Prophet); (3) Optimization pages (Infrastructure & LP, What-If Simulator, Scenario Simulator); (4) Equity & Inclusion pages (Inclusion & Equity, Diversity, Reserved Category Analysis); (5) ML Research pages (Advanced Predictions, Survival Analysis, Clustering, Model Explainability/SHAP); (6) Data pages (Data Dictionary, Data Quality, Correlation Analysis); (7) Support pages (Methodology, User Manual, AI Assistant); (8) Export pages (PDF Strategic Report Generator).

C. Dashboard Interface

The landing page presents a comprehensive overview with global filters (department, course, year) in the left sidebar, four KPI cards (total enrolment, applications, fill rate, placement rate), two time-series charts (enrolment trend 2015–2024, department fill rates), and a vacancy-risk distribution bar. This design surfaces the most important institutional metrics at a glance; deeper analysis is one click away on specialized pages.

D. Caching and Performance

Every expensive operation is cached. Excel loading, preprocessing and model training are wrapped with @st.cache_data decorators; the cache key includes the filter state so adjusted filters trigger model re- training. On a typical office laptop (Intel i7, 16GB RAM), first-load time is 4–6 seconds; subsequent page renders are sub-second. This performance is crucial for adoption in administrative workflows.

VIII. RESULTS AND DISCUSSION

A. Forecasting Performance

On the held-out 2024 test set, Random Forest achieves R-squared scores of 0.78–0.86 across primary targets (total enrolment, applications, rural students, urban students), exceeding the project's 0.75 SRS target. XGBoost performs marginally better (0.80–0.88); Linear Regression lags (0.60–0.72). Walk-forward backtesting (training on years $<y$, testing on year y for $y \in \{2019,...,2024\}$) reports mean absolute percentage error of 11–15% across folds, indicating robust performance on sequentially-held-out years. This conservative estimate is more realistic for operational use than single-split test-set accuracy.

B. Model Feature Importance

Feature-importance analysis reveals that lagged features (LAG_RURAL, LAG_URBAN, LAG_APPS) are the strongest predictors, accounting for 40–50% of importance. This confirms the serial-correlation hypothesis: historical behaviour is the strongest signal for next-year trends. FILL_RATE and PLACEMENT_RATE together account for 20–25%, reflecting that course health metrics carry signal. This importance ranking guides practitioner decisions about data collection priorities: if the system were deployed at other institutions, collecting accurate prior-year figures would be more critical than many other data fields.

C. Vacancy Risk Distribution

In the latest year, approximately 15 courses fall into the High vacancy-risk band (fill rate < 0.60), 40 into Medium (0.60–0.85), and the remainder into Low. The distribution varies by department; STEM departments tend toward Low risk, while newer or lower-fee programmes cluster in Medium and High. The root-cause analysis identifies that approximately 40% of High-risk courses suffer from demand drop (declining year-on-year applications), 30% from weak placement outcomes, and 20% from high dropout rates. This breakdown guides intervention design: demand-drop courses need geographic outreach; placement-weak courses need industry partnerships; dropout-prone courses need mentorship systems.

D. Sustainability Index Findings

The Sustainability Index distribution is centered around 0.55 with a long left tail. Approximately 12 courses fall below the 0.40 At-Risk threshold; these are candidates for academic-council review or restructuring. The index successfully identifies courses that multiple stakeholders would intuitively flag as vulnerable, validating the choice of constituent metrics (fill, placement, growth, retention). The fact that index values cluster in the 0.40–0.70 range, rather than showing bimodal or uniform distribution, suggests the index captures genuine programme-health variation rather than being a noisy combination of uncorrelated metrics.

E. Equity Audit Results

Reserved-category fill rates show a consistent pattern: General-category seats fill at >95%, OBC and EWS at 75–85%, ST at 50–70%, and SC at 60–75% on average. This pattern reflects both application patterns (more applicants in unreserved categories) and reservation-policy compliance constraints. The system identifies that three courses show severely imbalanced fill rates (e.g., ST fill <40% while GEN fill >90%), flagging them as equity-weak and recommending targeted outreach or reservationpolicy review. Gender representation averages 35% female across the institution; five departments fall below 25% female, warranting targeted interventions. Rural representation averages 28%; seven departments fall below 15%, indicating geographic reach gaps.

F. Linear Programming Recommendations

The LP optimizer categorizes courses as: Expand (enrolment gap >10%, indicating demand-constrained programmes where increasing intake is feasible), Maintain (gap within ±10%), or Reduce (gap >-10%, indicating supply-constrained or declining-demand programmes). Results categorize approximately 20 courses for expansion (high fill rate, excess application demand), 45 for maintenance, and 30+ for reduction. These recommendations respect UGC faculty-load norms (max 25 students per faculty) and infrastructure constraints. Comparison with historical intake decisions shows the LP algorithm would have approved 18 of 20 expansions that actually occurred, validating its utility as a decision-support tool rather than a final arbiter.

G. Statistical and Correlational Findings

Pearson correlation analysis confirms several intuitive relationships: placement rate correlates strongly with fill rate ($r \approx 0.58$, $p<0.01$), confirming that programmes with good placement outcomes attract applicants. Faculty load ratio correlates with dropout rate ($r \approx 0.42$, $p<0.05$), suggesting that overcrowded programmes have higher attrition. Rural ratio shows weak negative correlation with placement rate ($r \approx -0.15$), reflecting documented challenges rural students face in leveraging campus networks for jobs. Independent t-tests find significant differences in dropout rate between rural and urban cohorts ($p<0.01$), rural students dropping out 2–3% more frequently. This statistical evidence quantifies disparities that equity audits flag qualitatively.

H. Scenario Simulation Findings

Running the What-If simulator across representative courses shows that a uniform 10% rural-outreach increase (modelled as +10% on LAG_RURAL feature) produces an average enrolment lift of approximately 4.6%, while a 10% fee reduction produces an average lift of 3.2%. These elasticity estimates are derived from the Random Forest model's learned relationships and can guide policy design. The scenario simulator's named scenarios (Aggressive Growth, Cost Cut, Stability) allow administrators to explore the institutional implications of strategic choices before committing to policy.

IX. CONCLUSION AND FUTURE SCOPE

A. Conclusion

This paper presents BBAU Academic Admission Intelligence, an integrated AI-powered decision support system for course-level analytics in a central Indian university. The system combines forecasting, optimization, equity auditing and explainability into a unified platform that runs on standard office hardware. Quantitative results show that machine-learning models beat target accuracy thresholds, while qualitative feedback from departmental coordinators confirms that the system produces actionable insights. The principal insight is that the data BBAU already collects is sufficient—without acquiring external feeds or new sensors—to forecast enrolment with acceptable accuracy, identify at-risk courses with interpretable root causes, and recommend evidence-based interventions. The system is immediately deployable, maintainable and aligned with NEP-2020 governance mandates.

B. Key Contributions

1) A 36-page integrated dashboard for course-level decision support in a central university context

2) An ensemble forecasting pipeline achieving >0.78 RZ on held-out test data with walk-forward MAPE of 11–15%
3) A multi-dimensional equity audit covering nine demographic dimensions
4) A root-cause diagnostic cascade mapping vacancy risks to specific recommended interventions
5) A linear-programming seat-allocation optimizer respecting UGC and infrastructure constraints
6) SHAP-based feature-attribution explainability addressing auditability requirements
7) A fully reproducible system deployable on institutional infrastructure without a dedicated IT team

C. Future Scope

1) ERP Integration: Replace Excel data layer with live SQL connection to institutional ERP for real- time analytics
2) Student-Level Dropout Prediction: Privacy-preserving predictive models to identify at-risk individuals early
3) Multi-Tenant Federation: Generalize architecture for cross-university benchmarking
4) Reinforcement-Learning Optimization: Multi-period seat-allocation policy optimizing long-term sustainability
5) Vernacular Interface: Hindi and Awadhi native support for broader administrator accessibility

REFERENCES

[1] Brinkman, P. T., & McIntyre, C. (1997). Methods and techniques of enrolment forecasting. New Directions for Institutional Research, 93, 67–80.
[2] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
[4] Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37–45.
[5] Henderson, B. D. (1970). The product portfolio. Boston Consulting Group Perspectives.
[6] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774).
[7] Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481.
[8] Box, G. E., & Jenkins, G. M. (1970). Time Series Analysis: Forecasting and Control. Holden-Day.
[9] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
[10] Ministry of Education, Government of India. (2020). National Education Policy 2020.
[11] NITI Aayog. (2023). Vision India @ 2047: Towards a Developed India.
[12] Sprague, R. H., & Carlson, E. D. (1982). Building Effective Decision Support Systems. Prentice- Hall.
[13] University Grants Commission. (2018). UGC Regulations on Minimum Qualifications for Appointment of Teachers.
[14] Davidson-Pilon, C. (2019). Lifelines: Survival analysis in Python. Journal of Open Source Software, 4(40), 1317.
[15] Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 40(6), 601–618.