Harnessing Explainable AI (XAI) to Inform Educational Decision-Making: A Machine Learning Study on Graduation Rates

Clement Appeadu
Unaffiliated/Independent Researcher
Swansea, United Kingdom
appeaduclement@gmail.com

Improving student outcomes and institutional performance is a central theme in Educational Data Mining (EDM), where machine learning (ML) models are increasingly used for data-informed policymaking. However, the lack of transparency and interpretability in many ML models limits their utility for non-technical stakeholders. In this study, I used institutional-level data from the U.S. Integrated Postsecondary Education Data System (IPEDS) between 2012 and 2017 to predict graduation rates—a key metric of institutional effectiveness. I applied six supervised ML algorithms, with Support Vector Regression (SVR) achieving the highest test R^2 of 81.12%, followed closely by ensemble methods such as XGBoost, Random Forest, and LGBM. To enhance interpretability, I incorporated Explainable AI (XAI) tools including SHAP, Permutation Importance (PI), Partial Dependence Plots (PDPs), and Individual Conditional Expectation (ICE) plots. These analyses identified prior retention rates, equity gaps in graduation by race and gender, tuition levels, and test scores as key predictors. Force plots and waterfall charts were used to provide local explanations of individual predictions. The study demonstrates how XAI can support transparent and actionable educational policy. Code and data are publicly available at: https://github.com/cappeadu/edm_grad_xai.

Keywords: educational data mining, machine learning, explainable artificial intelligence (XAI), policy decision-making, transparent AI, institutional-level data, graduation rate prediction

1. INTRODUCTION

With growing concerns around student success and institutional accountability, educational leaders increasingly rely on data-driven decision-making (Gullo, 2013). The widespread availability of student and institutional data—encompassing demographics, test scores, tuition, and performance metrics—has created new opportunities to derive actionable insights. An emerging field contributing to this shift is Educational Data Mining (EDM), an interdisciplinary domain integrating machine learning, statistics, (Mahajan and Saini, 2020) and education theory to extract knowledge from educational data.

At the core of EDM is the application of machine learning (ML) to predict student outcomes and assess institutional effectiveness. Previous research in EDM has explored a wide range

of supervised, unsupervised, and reinforcement learning techniques, applied across student- and institutional-level datasets, and these studies have demonstrated the importance of machine learning in predicting academic outcomes such as grades, graduation rates, retention, and dropout likelihood (Dol and Jawandhiya, 2024). Incorporation of workflows such as feature engineering, model comparison, and parameter tuning has further helped optimize predictive performance (Tatineni and Mulukuntla, 2017) in diverse settings. Ultimately, these advances have had a significant impact on the education sector (Bangare et al., 2022).

However, despite the value of predictive power, a major challenge remains where many of these machine learning models function as "black boxes", making it difficult for stakeholders to understand how predictions are made. This lack of interpretability can limit trust and restrict the usefulness of the model to inform policy (Guidotti et al., 2018). To address this limitation, researchers are increasingly turning to eXplainable AI (XAI) which are a set of techniques designed to make ML models more transparent and interpretable (Adadi and Berrada, 2018). XAI methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) offer insight into model decision-making. For example, these models can help institutions understand which variables drive graduation rates or what factors most influence individual student outcomes. Such insights can guide interventions, improve accountability, and ultimately support more effective policy formulation.

Although interest in XAI is growing, its integration, particularly in modelling graduation rates using institutional-level data, remains underexplored. While improving predictive accuracy (Gupta et al., 2024) has been a key focus, relatively few studies incorporate XAI tools to make results understandable and actionable for non-technical stakeholders, policymakers and education leaders. In addition, few studies have explored temporal patterns in graduation data over multiple years to uncover evolving institutional performance trends. Much of the existing research in EDM has focused on student-level datasets, often overlooking the rich and policy-relevant information embedded in institutional-level data across years.

To address these gaps, this study employs supervised machine learning and XAI techniques including SHAP, Partial Dependence Plots (PDP), Individual Conditional Expectation (ICE), and Permutation Importance on U.S. institutional data from the U.S. Integrated Postsecondary Education Data System (IPEDS) for the years 2012 to 2017. The goal is not only to predict the institutional graduation rates of students in their first year but also to uncover the underlying patterns and identify the key drivers of institutional performance. This work contributes to the growing emphasis on interpretable machine learning in education (Wang and Luo, 2024) and supports the development of transparent, data-driven policies aimed at improving student success.

This study is organized as follows. Section 2 introduces the Knowledge Discovery in Databases (KDD) framework, which guides the knowledge extraction process in this research, and reviews related literature. Section 3 outlines the methodology used, while Section 4 presents the results and discussion of the analysis. Finally, Sections 5 concludes the study as well as addressing limitations and future work.

2. Literature Review

2.1. Data Mining and Knowledge Discovery in Databases (KDD).

Data mining refers to the use of machine learning and statistical techniques to identify patterns and extract meaningful insights from large dataset (Han et al., 2012). Its use spans multiple sectors, including education, where it forms the foundation of Education Data Mining (EDM)—an emerging field focused on analysing data from educational settings (Koedinger et al., 2015) to improve learning outcomes and institutional performance. EDM typically involves extracting insights from repositories that contain data on student demographics, academic performance, financial records, admission processes, and more. The central aim of data mining is to support data-driven decision-making by policymakers, administrators, and researchers through the discovery of knowledge from databases. This introduces the field of Knowledge Discovery in Databases (KDD).

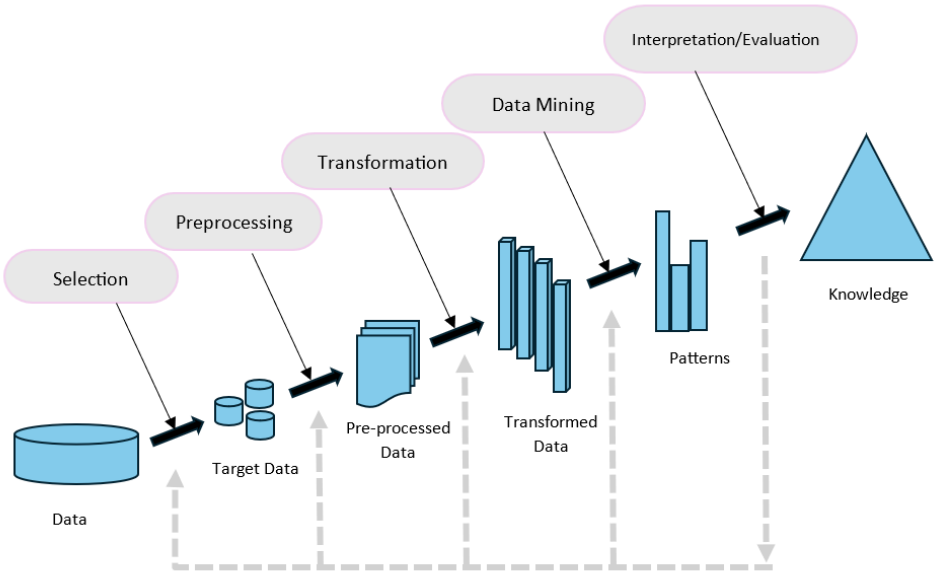

Figure 1: Overview of the Knowledge Discovery in Databases (KDD) process used to guide the methodology of this study, from data selection to interpretation.

Knowledge Discovery in Databases (KDD) is an iterative process designed to extract meaningful knowledge from raw data, with data mining serving as its core component. This study adopts the KDD methodology (shown in Figure 1) to guide the entire research process. As outlined by Fayyad et al. (1996), the five main steps are:

- **Data Selection:** Identifying the target data repository and understanding the available data in alignment with the goals and objectives of the stakeholders.
- **Data Preprocessing:** Applying techniques to improve data quality, including addressing missing values and reducing noise.

- **Data Transformation:** Transforming and engineering features using statistical methods to represent data in a suitable format for mining. It also involves selecting relevant variables.
- **Data Mining:** Choosing or combining algorithms (e.g. supervised, unsupervised, or reinforcement learning) that align with the project objectives to extract patterns or make predictions.
- **Interpretation and Evaluation:** Evaluating the resulting models and interpretation of their outputs. The iterative nature of KDD allows researchers to revisit earlier steps to fine-tune models or adopt alternative approaches. Insights are communicated through visualization and documentation, and findings are implemented or compared with existing research to resolve conflict and support decision making.

The intersection of KDD and data mining is critical as it ensures that the process yields interpretable and actionable insights (Shu and Ye, 2023).

2.2. Educational Data Mining.

Several machine learning techniques and algorithms have been applied in Educational Data Mining (EDM), including supervised, unsupervised, and reinforcement learning approaches (Kalita et al., 2025). The following provides a concise overview of these three primary categories.

Supervised Learning. This involves predicting a labelled outcome (commonly referred to as the *dependent variable* or "y") using one or more input variables (*independent variables* or "X"). It typically takes two forms:

- **Regression:** Predict continuous numeric outcomes. For example, graduation rates can be predicted using institutional- or student-level data.
- **Classification:** Predict categorical outcomes, such as whether a student will drop out or persist.

Supervised learning enables both outcome prediction and the discovery of patterns in labelled data. Common algorithms include Generalized Linear Models (GLM), Support Vector Machines (SVM), tree-based models such as Decision Trees, Random Forests, and Gradient Boosting machines.

Unsupervised Learning. Unsupervised learning involves analysing data without a labelled outcome. It relies solely on input variables to uncover patterns, groupings, and the underlying structure in the data. Common techniques include:

- **Clustering:** Groups similar data points based on identified patterns. In educational context, students may be clustered based on academic performance and demographics. Methods include K-means (partitioning), agglomerative clustering (hierarchical), and DBSCAN (density-based).
- **Dimensionality reduction:** Reducing the number of features in a dataset while preserving important features. Techniques such as Principal Component Analysis (PCA) and t-SNE help simplify models, improve performance, and eliminate noise or redundancy.

- **Anomaly detection:** Identifies data points that deviate significantly from the norm. In education, this can be applied to detect irregular academic performance or outliers in engagement.

REINFORCEMENT LEARNING. Reinforcement learning is a machine learning paradigm in which an autonomous agent learns by interacting with its environment, receiving feedback in the form of rewards or penalties. The agent continuously updates its strategy to optimize performance towards a defined goal (Singh et al., 2022). In education, reinforcement learning is used in applications such as intelligent tutoring systems, adaptive learning platforms, and personalized curriculum design.

2.3. APPLICATION OF MACHINE LEARNING METHODS AND WORKFLOWS IN EDM.

This section presents previous research that has explored and implemented different machine learning methods and workflows in EDM. These workflows (such as handling missing data, feature engineering, model evaluation, and assessing feature importance) are essential steps for improving predictive performance and drawing meaningful insights.

In a study aimed at predicting student performance, Chen et al. (2023) applied both unsupervised and supervised ML methods to university student data across all majors for academic years 2017 and 2018. The authors first applied an unsupervised approach using a matrix-based bipartite network with Louvain clustering, identifying six distinct clusters representing trends in student performance. These clusters were subsequently used as class labels for a supervised classification task. When evaluated using accuracy, precision, recall, and F1 score, the model outperformed several established algorithms including K-nearest Neighbours (KNN), Support Vector Machines (SVM), Decision Trees, Multi-layer Perceptron (MLP), and Convolutional Neural Networks (CNN). Similarly, Feng et al. (2022) applied K-means clustering to three different datasets, identifying four clusters for datasets A and C and five for dataset B. These clusters were used as target labels for classification. The authors compared two evaluation strategies: random hold-out method and shuffle 5-fold cross-validation. Results indicated that the Dataset A achieved higher accuracy with the hold-out method (94.59%) compared to the shuffle method (91.22%). Conversely, Datasets B (95.92%) and C (93.9%) showed slightly better performance under the shuffle method compared to the hold-out method (94.29% and 93.29% for B and C respectively).

In a study that used solely supervised learning, Yağcı (2022) incorporated students' midterm grades as a key input variable to predict final exam scores in a Turkish university setting. The tested algorithms included Random Forest, SVM, KNN, Naïve Bayes, and Logistic Regression. Using a 10-fold cross-validation, Random Forest had the best accuracy (74.6%). Feature importance analysis confirmed that the midterm score was a significant predictor, suggesting its importance in providing early feedback to students for exam preparation.

Musso et al. (2020) explored predictive modelling for three student outcomes — retention, GPA, and degree completion — using three types of perceptron artificial neural networks. The study found that student background played a major role in retention prediction, while learning strategies were the most significant for predicting GPA and degree completion. Interestingly, coping mechanisms were consistently identified as influential across all three models. The models demonstrated perfect performance (100%) in terms of accuracy, recall, and precision, though such high metrics may warrant closer scrutiny for overfitting or data leakage. Ramaswami and

Bhaskaran (2010) investigated the most important features for predicting student performance. Their workflow involved feature selection using the Pearson chi-square test (threshold>100) followed by modelling with the CHAID algorithm (Chi-squared Automatic Interaction Detection) and a 10-fold cross-validation for evaluation. Important variables identified included medium of instruction, area of residence, and school location.

A key focus in prediction tasks is the comparison of different machine learning algorithms to determine the best-performing model. For instance, Nahar et al. (2021) evaluated six ML models (Decision Tree, Naïve Bayes, PART, Bagging, Boosting, and Random Forest) for classifying student categories and predicting final grades. The study involved data cleaning, integration, and feature engineering and selection. Decision Tree was best for classifying student categories (accuracy: 64.28%), while Naïve Bayes (accuracy:73.07%) was preferred for grade prediction due to its simplicity in interpretability according to the authors. They concluded that mid-term exam performance significantly influenced final outcomes. A study by Zhang et al. (2021) compared models for grade prediction using two datasets: one with a five-level grading system (Erasmus Grade Conversion) and another with three-level of classifications. The models included Naïve Bayes, KNN, Decision Tree, SVM, Bagging, and Random Forest. The researchers utilised GridSearch with 10-fold cross-validation to fine-tune hyperparameters. Random Forest emerged as the top model, with feature importance analysis highlighting pre-requisite courses, last semester performance, and current term courses as key predictors.

Finally, Pelima et al. (2024) conducted a systematic review of 70 studies between 2018 to 2023 focused on graduation rate prediction using academic performance and machine learning. The findings revealed that SVM was the most frequently used algorithm (appearing in 31 studies), followed by Random Forest (27 studies), Logistic regression (13 studies), and K-Nearest Neighbours (9 studies).

2.4. Dataset and Features Used in EDM.

Educational repositories offer rich datasets that can be leveraged for a variety of purposes within EDM. These databases typically include both student-level database including demographics, attendance, and examination scores, and institutional-level data covering tuition, financial aid, retention rates, and faculty-related metrics. However, it is worth noting that the choice of data depends heavily on the research objective and the level of analysis (individual or institutional).

Mehta et al. (2021) introduced a novel framework to predict graduation rates of first-time, full-time undergraduate students using an evolutionary feature selection method coupled with regression modelling, based on the Integrated Postsecondary Education Data System (IPEDS). IPEDS contains institutional-level data collected from U.S. institutions across 12 survey components. The initial dataset for the study comprised 152 features; after data cleaning and feature engineering, 143 features were retained, covering data from 903 universities. The study highlighted a notable gap in the application of regression analysis within EDM, advocating for its expanded use. The authors applied Decision Tree Regressor, Support Vector Regression (SVR), and Multiple Linear Regression models. The authors further emphasized the importance of hyperparameter tuning in boosting model performance.

In a related study, Goenner and Snaith (2003) used both institutional-level and student-level data to investigate factors influencing graduation rates at 258 Carnegie I doctoral universities over 4-, 5-, and 6-year periods. Student-level variables included SAT scores, the proportion of students graduating in the top 10% of their class, the percentage of out-of-state students, and

average student age. Institutional-level factors encompassed class size, faculty composition, education-related expenditure, urbanization, the percentage of students PhDs, student-to-faculty ratio, and institutional affiliation. Features that did not meet statistical significance were omitted. Their findings underscored that student and institutional-level data are essential for effectively modelling university graduation rates. Notably, full-time faculty percentage, tuition and fees, educational expenditure, and the student-to-faculty ratio emerged as strong predictors.

Francis and Babu (2019) conducted a study using a diverse collection of features to predict student performance in higher education. The models used included SVM, Naïve Bayes, Decision Tree, and Neural Networks. Features were categorized into academic, demographic, behavioural, and extra features. The latter included data from surveys on parental satisfaction with the institution, and student absence records. The authors constructed three composite models: the first combined behavioural and extra features, the second integrated academic, behavioural, and extra features, and the third incorporated all four categories, including demographics. The best-performing model, using Decision Trees, included academic, behavioural, and extra features and achieved an accuracy of 75.47%. This study highlights the value of enriching datasets with non-traditional variables, such as survey data and behavioural metrics, to improve model interpretability and performance.

2.5. Explainable AI (XAI) in Educational Data Mining.

The practical interpretation of machine learning models has become increasingly important, particularly in high-stakes domains like education where decisions can significantly impact policy and student outcomes. The ability to understand *why* and *how* ML models make predictions is essential for building trustworthy, transparent, and accountable systems. This has led to growing interest in Explainable Artificial Intelligence (XAI) within the educational data mining community.

XAI comprises techniques that provide insights into the complex inner workings of black-box models, making their predictions interpretable to stakeholders such as policymakers, administrators, and educators. Popular XAI tools include:

- **SHAP (SHapley Additive exPlanations)**: A unified approach based on cooperative game theory that attributes each feature's contribution to a model's prediction.
- **Partial Dependence Plots (PDP):** Visualizations that illustrate the marginal effect of a feature on the predicted outcome.
- **Permutation Feature Importance (PI):** Measures the change in model performance when a single feature's values are randomly shuffled, indicating how important that feature is to the model's prediction.
- **LIME (Local Interpretable Model-agnostic Explanations):** Generates local approximations of a model's decision boundary to explain individual predictions.

Jang et al. (2022) applied SHAP to uncover both global and local feature importance in predicting early at-risk students. Their work demonstrated how SHAP values could inform early intervention strategies by revealing the most influential variables for individual students. Similarly, Gunasekara and Saarela (2025) combined SHAP and LIME to explain student outcomes, whether a student would pass or fail, using data from the Open University learning Analytics

(OULA) datasets. Their approach highlighted the strengths of using multiple XAI tools to gain a more robust understanding of model behaviour. Nagy and Molontay (2024) used a suite of XAI techniques including PI, PDPs, LIME and SHAP to interpret ML models built to predict whether students would dropout or graduate. Their analysis provided both model-agnostic and model-specific explanations, reinforcing the usefulness of XAI in improving transparency in EDM applications.

These studies illustrate the increasing importance of explainability in educational data mining, especially when models are used for decision-making that affects institutional policy or student trajectories. XAI enables stakeholders to go beyond prediction, offering nuanced explanations that can be used for diagnosis, intervention design, and equitable policy development.

3. METHODOLOGY

The field of EDM involves the application of a variety of techniques, algorithms, and domain knowledge to draw insights from educational databases and support informed decision-making. I adopted the Knowledge Discovery in Databases (KDD) framework, using its sequential stages to structure and guide the methodology.

3.1. DATA

I obtained institutional-level datasets in CSV format from the National Center for Education Statistics (NCES) through the Integrated Postsecondary Education Data System (IPEDS). The data spanned these survey components: institutional characteristics, completions, student financial aid, graduation rates, fall enrollment, academic libraries, 12-month enrollment, human resources, outcome measures, finance, and admissions. The study focused on active universities (in a given year data was collected) awarding at least baccalaureate degrees, with particular emphasis on undergraduate students. The dataset spans from 2012 to 2017, forming a multi-year panel suitable for identifying temporal patterns.

3.2. DATA PREPROCESSING AND TRANSFORMATION

Preprocessing included identifying and handling erroneous, missing data, cleaning the dataset, extracting relevant variables, and preparing the data for machine learning algorithms. Due to the structure of the IPEDS surveys, it was necessary to merge data across multiple components. Each university has a unique identifier (unitid) and name, which facilitated linking records. Universities missing graduation rate data (the target variable) for a given year were excluded from that year's dataset. Additionally, only institutions present in at least four of the six years were retained to ensure sufficient longitudinal representation.

3.2.1. Feature Engineering and Missing Values

Feature engineering involved creating new variables, transforming existing ones, and selecting features that could improve model performance and enhance insights. Examples of engineered features in this study include average total price across student categories, average standardized test scores, interaction terms (such as average test score x average loan amount), and disparities in graduation rates between racial and gender groups from the previous year.

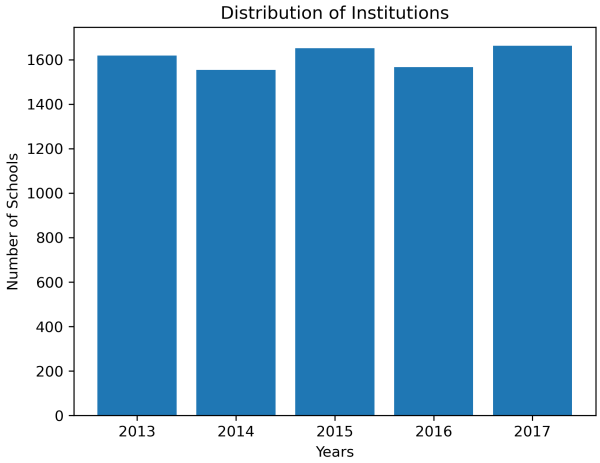

Figure 2: Bar plot with years (2013 to 2017) on the x-axis and number of universities on the y-axis. Each bar shows how many institutions were included for each year post-filtering. The plot shows that all years had a relatively stable number of institutions, around 1600, indicating a robust sample for modeling graduation rates.

To reduce dimensionality and multicollinearity, I used both Variance Inflation Factor (VIF) and correlation analysis. VIF identifies collinearity among independent variables. While multicollinearity does not inherently degrade predictive performance, it complicates interpretability and inflates the variance of coefficient estimates (Kyriazos and Poga, 2023). VIF being approximately 1 implies no correlation, VIF between 1 and 5 (inclusive) indicates moderate correlation. VIF greater than 5 suggests high correlation and features within this category are treated as candidates for removal. VIF threshold of 5 was used in this study with features exceeding this threshold being excluded unless retained due to domain significance or their relevance in interpretability (e.g., interaction features and features related to gender and race). Subsequently, 29 out of 93 features were selected for modelling and are detailed in the Appendix (see Table 2). Categorical variables were encoded using binary dummies (e.g., "Control of Institution" with public=1, private=0). Numerical features were standardized (mean=0, standard deviation=1) for models that require scaling (e.g., Linear regression and SVR) using StandardScaler to ensure compatibility across algorithms.

To support the engineering of features based on prior-year data (e.g., retention and graduation rates), data from 2012 was initially included in the merged dataset. However, because 2012 served as the baseline year, the engineered features capturing previous-year values resulted in missing data for that year. To ensure robust analysis and minimize bias from imputation, institutions with less than 90% non-missing data were excluded. As a result, institutions from 2012 were not included in the final analysis, which focused on the five subsequent years (2013–2017). The number of institutions that remained after this step is shown in Figure 2.

Missing values were imputed using the median of each feature, as several features exhibited wide variability and contained both small and large values. The median was chosen because it provides a more robust measure of central tendency in the presence of skewed distributions and outliers (Manikandan, 2011).

3.3. Data mining algorithms

The selection of machine learning models and workflow was based on prior literature and applicability to regression tasks. All modelling was conducted using Python 3.11.4 and the following algorithms were used:

1. **Linear Regression:** A generalized linear model using Ordinary Least squares (OLS) to estimate the relationship between features and the target variable. It is valued for its simplicity and interpretability.

2. **Decision Tree:** A non-parametric model that recursively partitions data based on feature splits. It learns hierarchical if-then rules and handles both linear and non-linear relationships.

3. **Random Forest:** An ensemble of decision trees built on bootstrap samples (bagging) with random feature selection, improving accuracy and generalization.

4. **eXtreme Gradient Boosting (XGBoost):** A highly efficient, scalable tree-boosting algorithm developed by Chen and Guestrin (2016). It is used in this study for its performance and feature importance insights.

5. **Support Vector Regression (SVR):** Based on Support Vector Machines, SVR uses kernel functions (e.g., radial basis, linear, polynomial) to model complex, high-dimensional relationships.

6. **LightGBM (LGBM):** A fast gradient-boosting algorithm based on decision trees, optimized for performance and scalability, especially with large feature sets.

3.4. Evaluation and Interpretation

3.4.1. Evaluation Metrics

The models were evaluated using three performance metrics:

- **R-Square (R^2):** Represents the proportion of variance in the target feature explained by the input features. Ranges from 0 to 1, with 1 indicating perfect prediction.

- **Root Mean Squared Error (RMSE):** Measures the standard deviation of prediction errors. It retains the same unit as the target variable.

- **Mean Absolute Error (MAE):** Captures the average absolute difference between predictions and actual values. Like RMSE, it is expressed in the same unit as the outcome.

R-Square was used to assess the goodness of fit, while RMSE and MAE evaluated prediction error magnitude. To improve model performance, hyperparameter tuning through Randomized Search with 5-fold cross-validation and 20-iterations was used. Each iteration sampled combinations of hyperparameters and scored them using R^2. This approach ensures robust performance (Muhajir et al., 2021) by evaluating models on multiple subsets of the training data.

3.4.2. Interpretation

To explain model behaviour and enhance transparency, several Explainable AI (XAI) techniques were adopted:

SHAP (SHAPLEY ADDITIVE EXPLANATION). SHAP was used for both global and local interpretations. Based on Shapley values from cooperative game theory, SHAP assigns importance scores to each feature's contribution to a prediction (Rozemberczki et al., 2022). SHAP summary plot and dependence plots were used for global explanations while SHAP force and waterfall plots were adopted for the explanation of individual predictions.

PERMUTATION FEATURE IMPORTANCE (PI). To complement SHAP, PI was used to measure the drop in model performance when a feature's values are randomly shuffled, revealing how critical the feature is to the model's output.

PARTIAL DEPENDENCE PLOTS (PDP) AND INDIVIDUAL CONDITIONAL EXPECTATION (ICE). These plots are used to visualize and analyse the interactions between an outcome variable and a feature of interest, with both plots assuming that the input feature of interest is independent from other variables. PDPs show the average marginal effect of a feature on the outcome variable, and this can be interpretated as the expected target response as a function of the feature of interest. However, ICE plots display how the prediction changes for individual samples.

4. RESULTS AND DISCUSSION

This section presents the outcomes of the analysis, including the evaluation metrics of six predictive models. The models were trained using data from 2013 to 2016, with 2017 data held out for testing.

Table 1: Performance comparison of six machine learning models on the test set using R^2, RMSE, and MAE metrics. SVR achieved the highest predictive accuracy, followed closely by XGBoost and Random Forest. Models are ranked based on R^2 and error magnitudes.

Model	**R^2**	**RMSE**	**MAE**	**Rank**
Linear Regression	0.7104	10.9572	7.6822	5
SVR	0.8112	8.8470	5.7218	1
Decision Tree	0.6906	11.3250	7.9173	6
Random Forest	0.7997	9.1123	6.0002	3
XGBoost	0.8024	9.0500	5.8177	2
LGBM	0.7991	9.1262	5.7313	4

4.1. METRICS

Table 1 displays the result of the test set after applying Randomized Search (hyperparameter tuning) on the training data. The best estimator for each model was selected based on cross-validation scores and subsequently evaluated on the test data.

Models are ranked from 1 (highest) based on the combination of R^2, RMSE, and MAE. SVR emerged as the top-performing model with an R^2 of 81.12%, RMSE of 8.847, and MAE of 5.7218, followed closely by XGBoost with an R^2 of 80.24%, RMSE of 9.05, and MAE of 5.8177. Random Forest and LightGBM also performed competitively. The slight performance

differences between the ensemble models—XGBoost, Random Forest, and LightGBM—were relatively small, highlighting the robustness of ensemble-based approaches in capturing complex patterns in institutional-level educational data.

The test results for all models were broadly consistent with those observed during cross-validation, suggesting strong generalization capabilities. However, it was observed that Random Forest and XGBoost showed slight variations in repeated runs, even with fixed *random_state* parameters and controlled cross-validation (e.g., KFold with shuffling and seed setting). This residual variability is likely due to internal stochastic elements which may not be fully governed by external random state controls. Nonetheless, the variation in performance metrics was marginal and did not materially affect model rankings or reliability.

Given its balance of performance and model structure suited for interpretability, I selected XGBoost for further analysis using Explainable AI (XAI) tools.

4.2. Global Interpretation

4.2.1. Feature importance

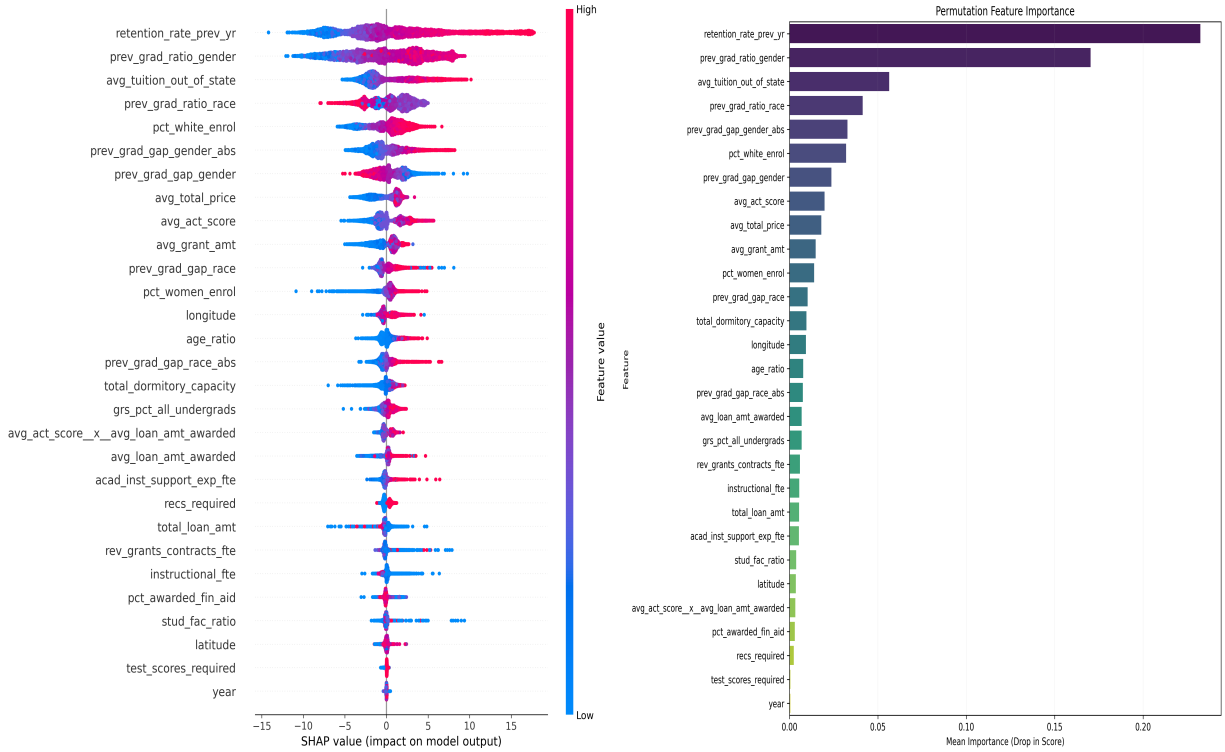

Figure 3: SHAP summary plot (left) and permutation feature importance plot (right) showing the top predictors of graduation rates. Both methods highlight previous retention rate, disparities in prior graduation rates across race and gender, tuition, and test scores as the most influential features.

To understand the global importance of features contributing to the prediction of graduation rates, this study used SHAP summary plots, complemented by permutation feature importance (PI). These tools provide insights into which institutional features most significantly impact graduation outcomes, offering valuable guidance to educators and policymakers. Figure 3 (left)

presents the SHAP summary plot, where features are ranked by their mean absolute SHAP values. The most predictive features include previous year's retention rates, racial and gender disparities in graduation rates from the prior year, average tuition, test scores, and average total price for students.

In SHAP summary plots, red points represent high feature values, while blue points represent low feature values. SHAP values on the right side of the axis push predictions toward higher graduation rates, while values on the left pull predictions lower. For example, high values of *retention_rate_prev_yr*, *avg_tuition_out_of_state*, *and average_act_score* increase the predicted graduation rate. However, for features like *prev_grad_ratio_race* and *prev_grad_gap_gender*, high values correspond to a negative impact—suggesting that large disparities in past graduation rates by race or gender are associated with lower graduation outcomes. This likely reflects the effect of equitable racial and gender graduation outcomes contributing positively to overall rates. Conversely, low values (i.e., smaller disparities or more balanced outcomes) are linked to higher graduation rates.

To confirm these insights, Permutation Importance (PI) was used (Figure 3, right). The top twelve features from PI closely match those from SHAP in both identity and relative importance. The consistency between SHAP and PI affirms the reliability of the identified features. However, it is worth noting that PI assumes feature independence, which may not always hold true in practice.

4.2.2. SHAP Dependence Plot, PDPs, and ICE.

To further understand how specific features affect predictions, SHAP dependence plots were used alongside Partial Dependence Plots (PDPs) and Individual Conditional Expectation (ICE) plots. While PDP and ICE plots provide useful visual summaries, they rely on the assumption of feature independence from others, which may not always hold in real-world educational datasets. Therefore, their interpretations are most meaningful when combined with techniques like SHAP that account for interactions.

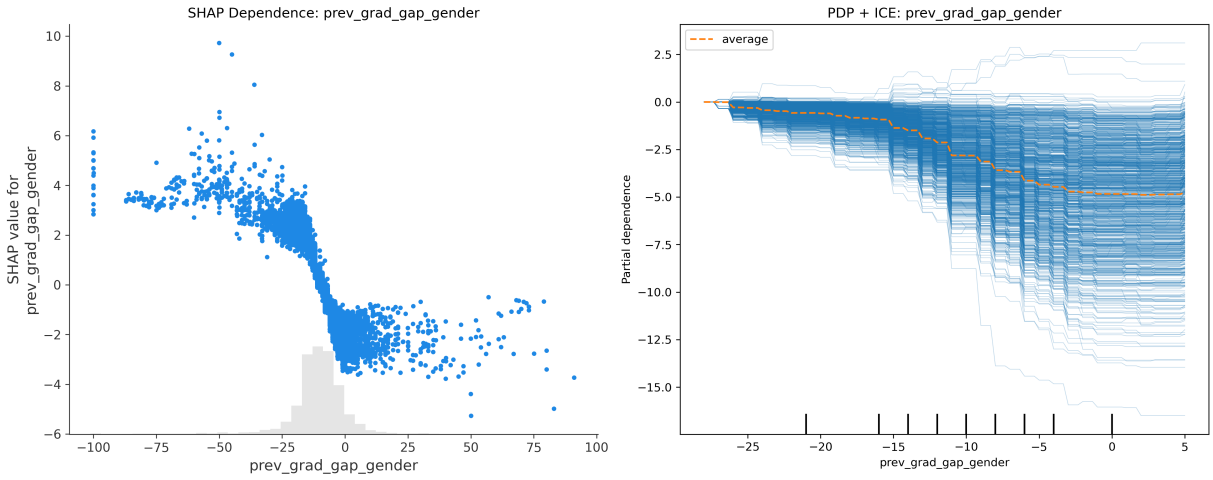

Figure 4: SHAP dependence plot (left) and PDP+ICE (right) for the gender-based graduation gap feature. These plots show that higher graduation rates are associated with negative gaps (i.e., women graduating at higher rates than men).

PREVIOUS GRADUATION GAP BETWEEN MEN AND WOMEN. This feature (*prev_grad_gap_gender*) is defined as the difference in graduation rates between men and women from the previous year. Negative values indicate higher graduation rates for women.

Figure 4 shows that as the gender gap increases (i.e., men outperform women), the predicted graduation rate declines. SHAP dependence plot (Figure 4, left) shows a downward trend in SHAP values as the feature values increases. The PDP+ICE plots confirm this: the average marginal effect (orange PDP line) declines with higher gender gap values. Individual institutions (blue ICE lines) follow a similar pattern, although with some variability in magnitude. Both plots suggest that higher institutional graduation rates are associated with negative values, i.e., where women graduate at higher rates than men. This aligns with existing literature that consistently shows women's graduation rates tend to be higher (Ma and Pender, 2023).

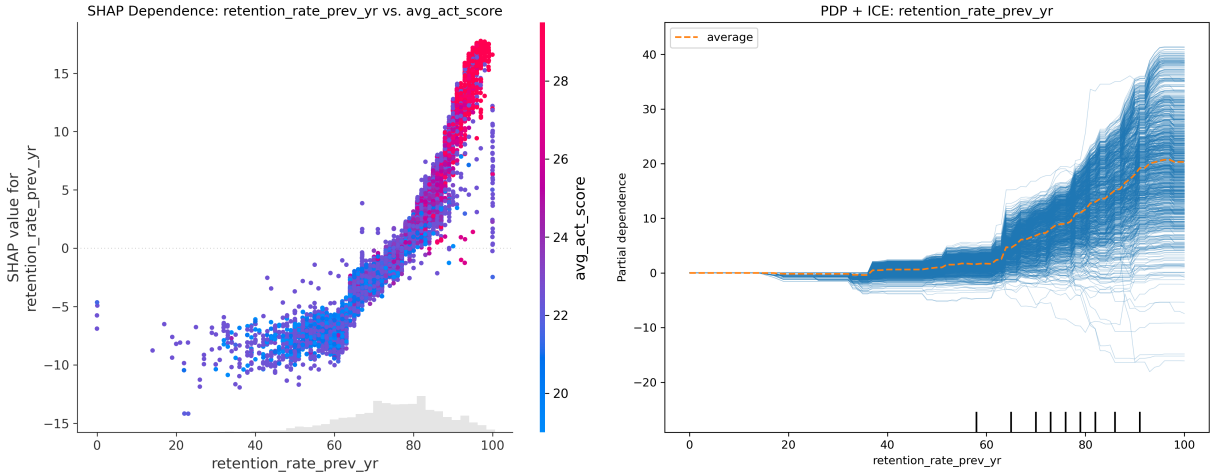

Figure 5: SHAP dependence plot (left) and PDP+ICE (right) for previous year's retention rate. Higher retention rates correspond with higher predicted graduation rates, especially when test scores are also high.

RETENTION RATES AND TEST SCORES. Figure 5 illustrates that higher retention rates from the previous year are associated with higher predicted graduation rates. The SHAP dependence plot reveals this positive relationship, especially when retention rates are paired with high *avg_act_score* (shown in red). The interaction between retention rate and test scores amplifies their predictive impact. The PDP on the right supports this, but the ICE lines show that in a few institutions, even high retention rates correspond to lower graduation predictions, indicating variability that may need local-level investigation.

Figure 6 shows that ACT scores above approximately 22 are consistently associated with higher graduation rate predictions. This confirms earlier insights from Figure 5, reinforcing test score importance in model outputs.

TUITION AND GRANT AMOUNT. Figure 7 demonstrates that higher average tuition is generally linked to higher graduation rates as SHAP values increase with tuition. Moreover, this effect is strengthened when combined with high grant support, as shown in the interaction colouring. This may indicate that institutions charging higher tuition but offering substantial grants are more likely to achieve better student outcomes.

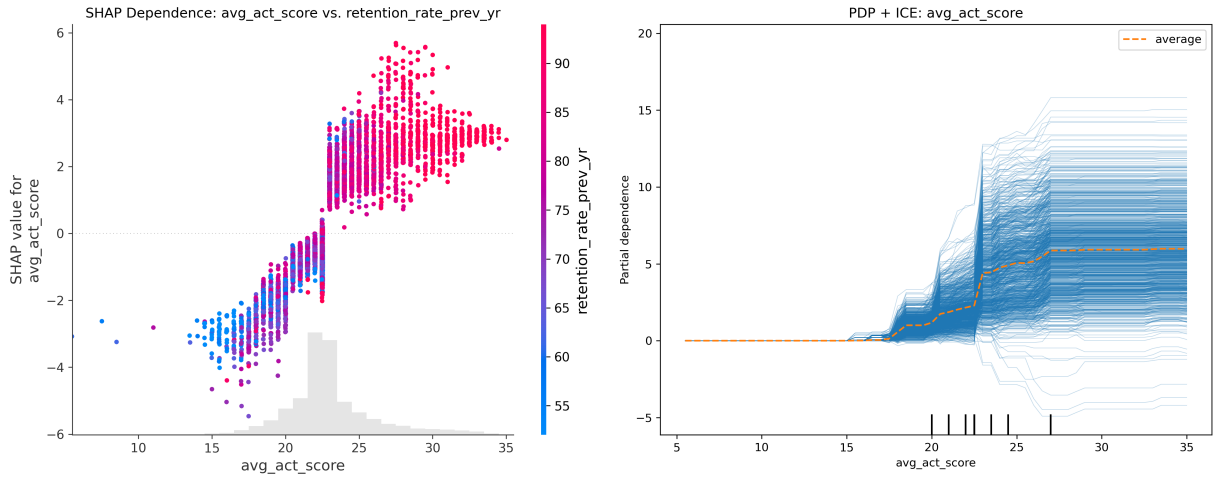

Figure 6: SHAP dependence plot (left) and PDP+ICE (right) for average ACT scores. Both plots indicate that test scores above approximately 22 significantly boost predicted graduation rates.

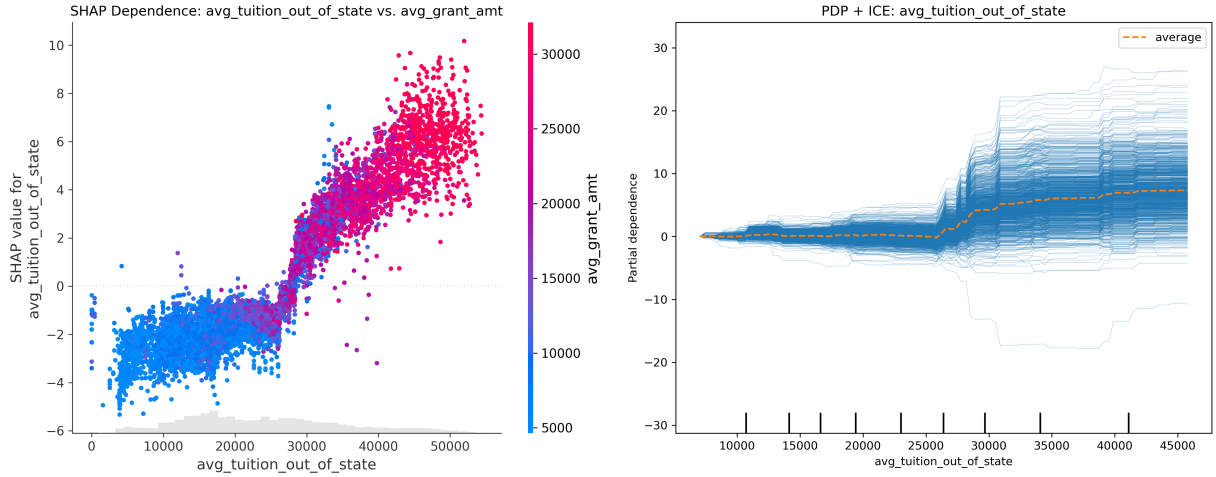

Figure 7: SHAP dependence plot (left) and PDP+ICE (right) for average tuition, with interaction from grant amounts. Higher tuition is associated with higher graduation rates, especially when paired with high grant support.

4.3. Local Explanation of Individual Predictions

To provide personalized insights at the institutional level, local interpretation methods such as SHAP force plots and waterfall plots were employed. These help interpret how individual feature values contribute to the model's prediction for a single institution. Figure 8 presents force plots for four institutions. These show how each feature contributes positively (red) or negatively (blue) to the final prediction. For example, at *school_idx* = 800, features such as *prev_grad_ratio_gender* and *avg_tuition_out_of_state* increased the predicted graduation rate, while *retention_rate* and *prev_grad_ratio_race* reduced the prediction.

Figure 9 and 10 use waterfall plots to visualize predictions for *school_idx* = 70 and 75. Here, feature values and contributions are displayed side by side, with less important features collapsed into a single row. For *school_idx* = 75, the most significant negative impact came from

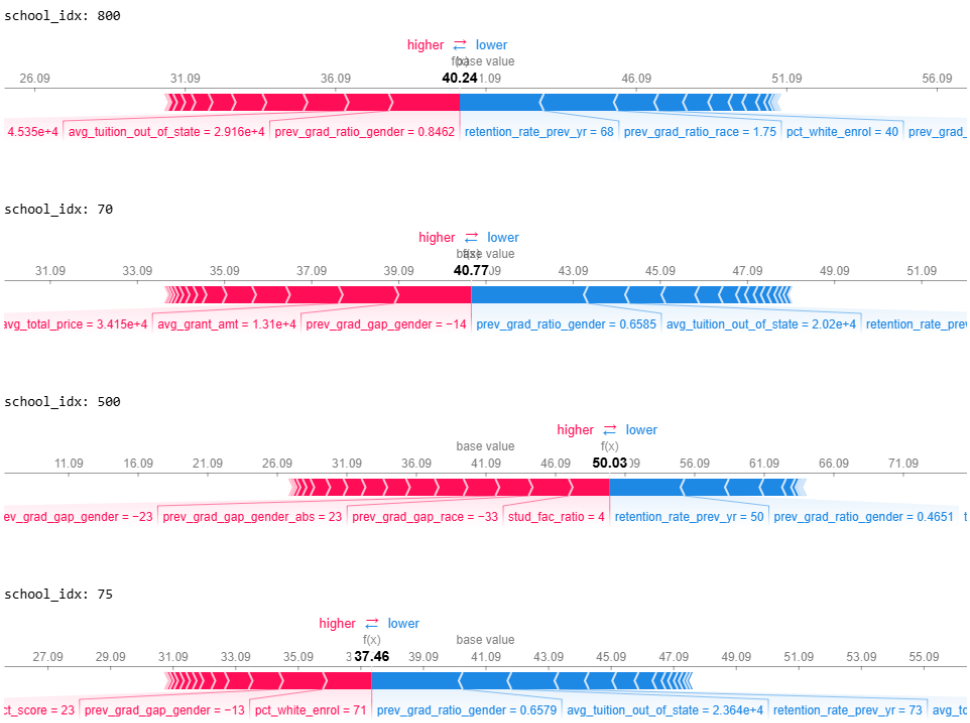

Figure 8: Force plots showing how different features contribute to graduation rate predictions for four institutions. Features in red push predictions up, while blue features push them down.

prev_grad_ratio_gender. Other negative contributors include: *avg_tuition_out_of_state*, moderate *retention_rate_prev_yr*, and *avg_total_price*. Positive contributions included percentage of white enrolment (*pct_white_enrol*), gender gap favouring women (*prev_grad_gap_gender*), *avg_act_score*, and *pct_women_enrol*.

DISCUSSION. This study has underscored the importance of regression analysis in EDM, advocated by (Mehta et al., 2021), and its integration with explainable AI (XAI) techniques to predict and interpret institutional graduation rates in U.S. universities. SVR, from the SVM family—which was identified as the most commonly used algorithm in the systematic review by Pelima et al. (2024)—emerged as the best-performing model, achieving an R^2 of 0.8112 on the test set. It was closely followed by XGBoost, Random Forest, and LGBM, with the results reaffirming the effectiveness of ensemble-based algorithms in handling complex educational data. Despite the performance of SVR, XGBoost was selected for XAI analysis due to its structured approach to feature importance, which supports better interpretability.

The use of SHAP, PDP, and ICE plots allowed both global and local interpretability. SHAP global plots confirmed the overall feature importance rankings, while force and waterfall plots demonstrated how specific institutional characteristics influenced individual predictions. A major finding of this study is the significant role of previous year's retention rate as the most important predictor of graduation rates. This aligns with prior research which constantly links retention rate to eventual completion (Crawford, 2015; Zhang and Koshmanova, 2020). This finding highlights the need for institutions to invest in early student support and engagement

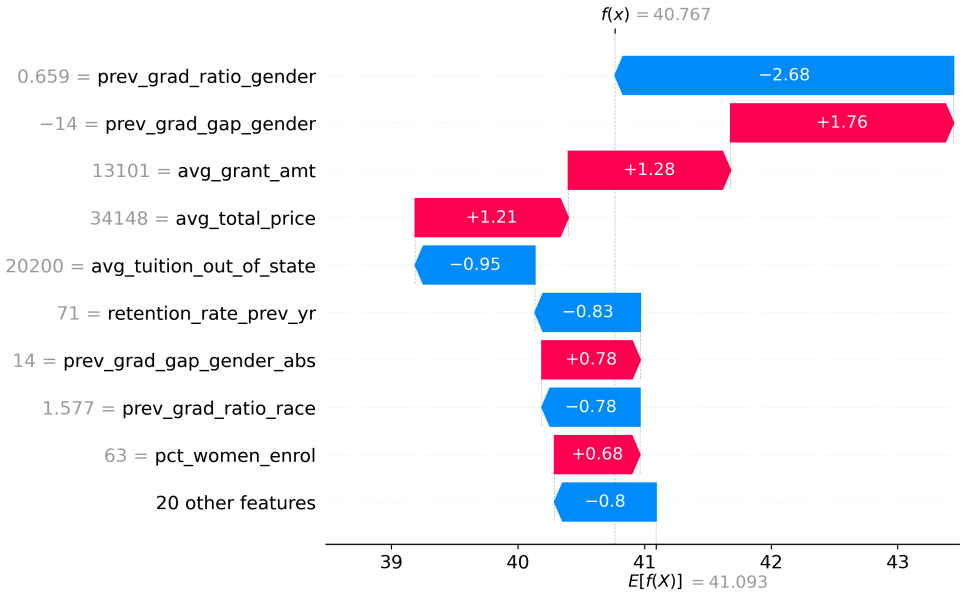

Figure 9: SHAP waterfall plot for institution 70 illustrating the cumulative effect of features on its graduation rate prediction. The plot distinguishes features that increase or decrease the final prediction.

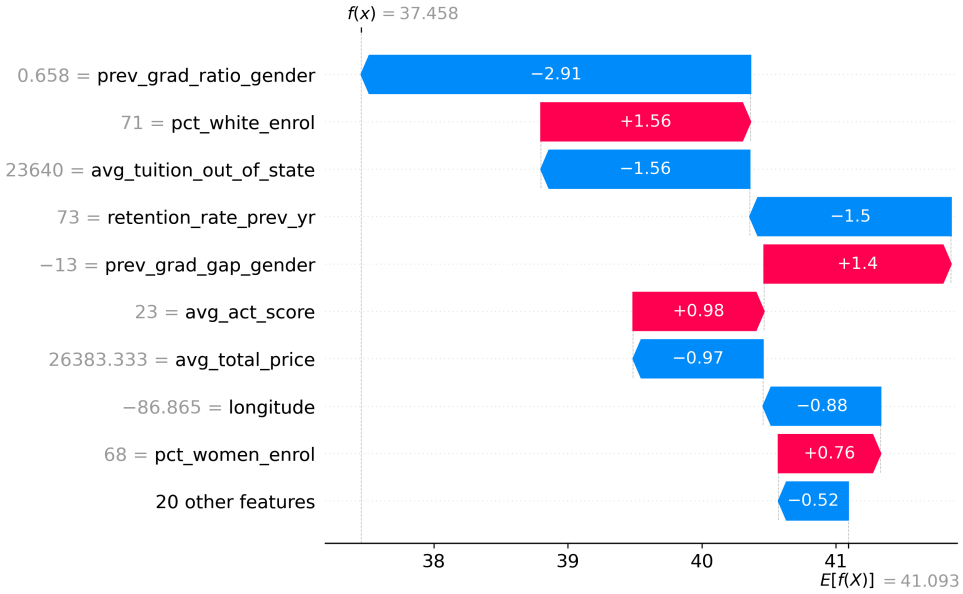

Figure 10: SHAP waterfall plot showing how feature values contribute to the graduation rate prediction for institution 75. The plot highlights both positive and negative influences, centered around the model's expected value.

strategies to boost retention rates (Hanover-Research, 2014), particularly in the first year.

The study also revealed that graduation rates disparities between gender and racial groups are also key predictive factors, with *prev_grad_gap_gender* and *prev_grad_ratio_race* providing insights into how equity outcomes within institutions affect overall graduation rates. SHAP

dependence plots showed that institutions where women outperform men in graduation tend to see higher overall graduation rates. Also, smaller gaps in graduation rates between white and non-white students were associated with better institutional outcomes. This emphasizes that equity in educational attainment is not only a moral imperative but also a strategic driver of institutional success (OECD, 2019).

Analysis of test scores highlighted the critical importance of improving students' preparedness for standardized assessments, as high average test scores were positively associated with high graduation rates. Policymakers should take proactive steps to enhance this readiness by providing targeted preparatory programs, resources, and support services for prospective students.

Furthermore, the alignment of tuition levels with eligible grant support emerged as a key factor influencing graduation outcomes. This suggests that institutions and policymakers should consider financial aid structures that effectively offset tuition costs, particularly for students from underrepresented or economically disadvantaged backgrounds, to help improve educational attainment.

Although force and waterfall plots show a single sample worth of data, it is most efficient to combine it with SHAP dependence plots to understand the impact of changing feature values, taking advantage of interaction features. By aligning these local explanations with global insights from Figures 3 - 7, institutions can make data-informed decisions. For instance, an institution with an ACT score of less than 23 and a retention rate below 73% may fall short of the optimal range seen in SHAP dependence plots (Figure 6). Similarly, an institution with suboptimal values for tuition and grants may also fall short. Even among institutions with similar retention rates or tuition levels, the interactions with other features such as test scores and demographic composition could shift predictions significantly.

These local explanations illustrate how machine learning and XAI tools can be used not just for institutional benchmarking, but also for crafting targeted interventions at the institutional level by properly aligning features to refine policy programs related to student support, tuition, or financial aid. This level of insight also offers policymakers and administrators a practical tool for strategic planning, enabling them to identify not just which features matter, but how they matter and for which institutions. However, such insights must always be contextualized within broader educational policies to avoid overgeneralization from isolated feature patterns.

The integration of machine learning and XAI has proven to be a useful approach for transforming educational data into actionable knowledge. While prior research has largely focused on student-level predictors (Ujkani et al., 2024; Ramaswami et al., 2022; Hoq et al., 2024) this study highlights the importance of institutional-level analysis in understanding systemic trends and informing high-level policy decisions.

5. Conclusion and Limitations

This study applied machine learning and explainable AI techniques to predict graduation rates using institutional-level data from IPEDS spanning 2012-2017. Among six models, SVR was identified as the best performing algorithm based on R^2 and error metrics but interpretation was based on XGBoost which performed similarly well. Through the utilization of global and local interpretability tools such as SHAP, Permutation Importance, PDP, and ICE plots, the study uncovered relevant insights:

- Retention rate is the strongest predictor of graduation rate, supporting the importance of early student engagement.
- Equity indicators, such as gender and graduation gaps have a crucial role in shaping institutional success.
- Financial factors including tuition and grant amounts affect graduation, especially when considered in tandem.
- Test scores are also relevant in predicting graduation rates.
- XAI tools offer stakeholders the opportunity to move beyond black-box predictions, allowing interpretable and transparent models suitable for real-world policy applications.

These findings can inform institutional strategies by identifying actionable drivers of student success, such as retention-focused interventions, equitable academic support, and financial aid optimization. Furthermore, the integration of XAI ensures that predictive insights are transparent and interpretable, enabling educational leaders and policymakers to make evidence-based decisions with confidence. By bridging predictive modeling with explainability, this work contributes to the responsible application of AI in education and supports scalable, ethical, and data-driven reforms.

LIMITATIONS AND FUTURE WORK. While this study offers valuable insights into the use of machine learning and explainable AI for predicting graduation rates, several limitations should be acknowledged. First, SHAP and other XAI tools used in the analysis explain model behavior rather than establishing causal relationships; as such, the findings should not be interpreted as evidence of cause and effect. Future research could incorporate causal inference frameworks to more robustly evaluate potential policy interventions. Additionally, the dataset is limited to U.S. institutions, which may restrict the generalizability of the results to other educational systems or international contexts. The study also did not include certain potentially influential factors—such as access to mental health services, campus engagement, or regional economic conditions—due to limitations in data availability. Finally, the data spans up to 2017 and therefore does not reflect more recent transformations in higher education, particularly those prompted by the COVID-19 pandemic. Addressing these limitations in future work by integrating student-level data, longitudinal causal models, and a broader range of contextual variables would provide a more comprehensive understanding of the factors influencing institutional graduation outcomes.

DECLARATION OF GENERATIVE AI SOFTWARE TOOLS IN THE WRITING PROCESS

During the preparation of this work, the author used https://chatgpt.com/ *in all sections to improve readability and ensure fluent texts. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.*

ACKNOWLEDGMENT

This work was inspired by the author's MSc dissertation, although the current analysis involves a distinct dataset and newly engineered features.

REFERENCES

ADADI, A. AND BERRADA, M. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access 6*, 52138–52160.

BANGARE, M. L., BANGARE, P. M., RAMIREZ-ASIS, E., JAMANCA-ANAYA, R., PHOEMCHALARD, C., AND BHAT, D. A. R. 2022. Role of machine learning in improving tourism and education sector. In *Materials Today: Proceedings*. Vol. 51. Elsevier Ltd, 2457–2461.

CHEN, T. AND GUESTRIN, C. 2016. Xgboost: A scalable tree boosting system. *CoRR abs/1603.02754*.

CHEN, Z., CEN, G., WEI, Y., AND LI, Z. 2023. Student performance prediction approach based on educational data mining. *IEEE Access 11*, 131260–131272.

CRAWFORD, G. A. 2015. The academic library and student retention and graduation: An exploratory study. *portal: Libraries and the Academy 15*, 41–57.

DOL, S. M. AND JAWANDHIYA, P. M. 2024. Systematic review and analysis of edm for predicting the academic performance of students. *Journal of The Institution of Engineers (India): Series B 105*, 1021–1071.

FAYYAD, U., PIATETSKY-SHAPIRO, G., AND SMYTH, P. 1996. From data mining to knowledge discovery in databases) (© aaai). *17*, 37–54.

FENG, G., FAN, M., AND CHEN, Y. 2022. Analysis and prediction of students' academic performance based on educational data mining. *IEEE Access 10*, 19558–19571.

FRANCIS, B. K. AND BABU, S. S. 2019. Predicting academic performance of students using a hybrid data mining approach. *Journal of Medical Systems 43*.

GOENNER, C. F. AND SNAITH, S. M. 2003. Predicting graduation rates: An analysis of student and institutional factors at doctoral universities. Tech. rep.

GUIDOTTI, R., MONREALE, A., RUGGIERI, S., TURINI, F., GIANNOTTI, F., AND PEDRESCHI, D. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys 51*.

GULLO, D. F. 2013. Improving instructional practices, policies, and student outcomes for early childhood language and literacy through data-driven decision making. *Early Childhood Education Journal 41*, 413–421.

GUNASEKARA, S. AND SAARELA, M. 2025. Explainable ai in education: Techniques and qualitative assessment. *Applied Sciences (Switzerland) 15*.

GUPTA, V., SINGHAL, P., AND KHATTRI, V. 2024. Enhancing predictive accuracy in education: A detailed analysis of student performance using machine learning models. Tech. rep.

HAN, J., KAMBER, M., AND PEI, J. 2012. *Data Mining: Concepts and Techniques, 3rd Edition (A volume in The Morgan Kaufmann Series in Data Management Systems)*, 3 ed. Morgan Kaufmann.

HANOVER-RESEARCH. 2014. Strategies for improving student retention. Tech. rep., Hanover Research.

HOQ, M., BRUSILOVSKY, P., AND AKRAM, B. 2024. Explaining explainability: Early performance prediction with student programming pattern profiling. *Journal of Educational Data Mining 16*, 115–148.

JANG, Y., CHOI, S., JUNG, H., AND KIM, H. 2022. Practical early prediction of students' performance using machine learning and explainable ai. *Education and Information Technologies 27*, 12855–12889.

KALITA, E., OYELERE, S. S., GAFTANDZHIEVA, S., RAJESH, K. N., JAGATHEESAPERUMAL, S. K., MOHAMED, A., ELBARAWY, Y. M., DESUKY, A. S., HUSSAIN, S., CIFCI, M. A., THEODOROU, P., HILČENKO, S., HAZARIKA, J., AND ALI, T. 2025. Educational data mining: a 10-year review.

KOEDINGER, K. R., D'MELLO, S., MCLAUGHLIN, E. A., PARDOS, Z. A., AND ROSÉ, C. P. 2015. Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science 6*, 333–353.

KYRIAZOS, T. AND POGA, M. 2023. Dealing with multicollinearity in factor analysis: The problem, detections, and solutions. *Open Journal of Statistics 13*, 404–424.

MA, J. AND PENDER, M. 2023. Education pays 2023 the benefits of higher education for individuals and society. Tech. rep.

MAHAJAN, G. AND SAINI, B. 2020. Educational data mining: A state-of-the-art survey on tools and techniques used in edm. Tech. rep.

MANIKANDAN, S. 2011. Measures of central tendency: Median and mode. *Journal of Pharmacology and Pharmacotherapeutics 2*, 214–215.

MEHTA, M. H., C, C. N., AND GOKHALE, A. 2021. Predicting institute graduation rate with genetic algorithm assisted regression for education data mining. *ICTACT Journal on Soft Computing 11*, 2266–2278.

MUHAJIR, D., AKBAR, M., BAGASKARA, A., AND VINARTI, R. 2021. Improving classification algorithm on education dataset using hyperparameter tuning. In *Procedia Computer Science*. Vol. 197. Elsevier B.V., 538–544.

MUSSO, M. F., HERNÁNDEZ, C. F. R., AND CASCALLAR, E. C. 2020. Predicting key educational outcomes in academic trajectories: a machine-learning approach. *Higher Education 80*, 875–894.

NAGY, M. AND MOLONTAY, R. 2024. Interpretable dropout prediction: Towards xai-based personalized intervention. *International Journal of Artificial Intelligence in Education 34*, 274–300.

NAHAR, K., SHOVA, B. I., RIA, T., RASHID, H. B., AND ISLAM, A. H. 2021. Mining educational data to predict students performance: A comparative study of data mining techniques. *Education and Information Technologies 26*, 6051–6067.

OECD. 2019. *PISA 2018 Results (Volume II): Where All Students Can Succeed*. OECD Publishing.

PELIMA, L. R., SUKMANA, Y., AND ROSMANSYAH, Y. 2024. Predicting university student graduation using academic performance and machine learning: A systematic literature review. *IEEE Access 12*, 23451–23465.

RAMASWAMI, G., SUSNJAK, T., AND MATHRANI, A. 2022. Supporting students' academic performance using explainable machine learning with automated prescriptive analytics. *Big Data and Cognitive Computing 6*.

RAMASWAMI, M. AND BHASKARAN, R. 2010. A chaid based performance prediction model in educational data mining. *IJCSI International Journal of Computer Science Issues 7*.

ROZEMBERCZKI, B., WATSON, L., BAYER, P., YANG, H. T., KISS, O., NILSSON, S., AND SARKAR, R. 2022. The shapley value in machine learning. In *IJCAI International Joint Conference on Artificial Intelligence* (Vienna). International Joint Conferences on Artificial Intelligence, 5572–5579.

SHU, X. AND YE, Y. 2023. Knowledge discovery: Methods from data mining and machine learning. *Social Science Research 110*.

SINGH, B., KUMAR, R., AND SINGH, V. P. 2022. Reinforcement learning in robotic applications: a comprehensive survey. *Artificial Intelligence Review 55*, 945–990.

TATINENI, S. AND MULUKUNTLA, S. 2017. Optimizing machine learning workflow efficiency: Comprehensive tooling and best practices. *3*, 2454–2016.

UJKANI, B., MINKOVSKA, D., AND HINOV, N. 2024. Course success prediction and early identification of at-risk students using explainable artificial intelligence. *Electronics (Switzerland) 13*.

WANG, S. AND LUO, B. 2024. Academic achievement prediction in higher education through interpretable modeling. *PLoS ONE 19*.

YAĞCI, M. 2022. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments 9*.

ZHANG, W. AND KOSHMANOVA, T. 2020. Relationship between factors and graduation rates for student success in the u.s. colleges. In *The European Conference on Education (ECE 2020)* (London). The European Conference on Education.

ZHANG, Y., YUN, Y., AN, R., CUI, J., DAI, H., AND SHANG, X. 2021. Educational data mining techniques for student performance prediction: Method review and comparison analysis. *Frontiers in Psychology 12*.

APPENDIX

Table 2: Descriptions of selected institutional-level features used in the predictive modeling of graduation rates. These features include academic, financial, demographic, geographic, and resource-related variables, as well as engineered features.

Feature	**Description**
retention_rate_prev_yr	Retention rate from the previous year
avg_tuition_out_of_state	Average tuition for out-of-state students
avg_act_score	Average 25th and 75th ACT composite scores
pct_white_enrol	Percent of undergraduate enrollment that are white
avg_grant_amt	Average amount of federal, state, local or institutional grant aid awarded
grs_pct_all_undergrads	Full-time, first-time, degree/certificate seeking undergraduates (GRS cohort) as percent of all undergraduates
acad_inst_support_exp_fte	Academic and institutional support, and student services expense per full-time-equivalent students
total_dormitory_capacity	The maximum number of students that the institution can provide residential facilities for
age_ratio	Ratio of students under 25 years to students 25 years and over
pct_women_enrol	Percent of undergraduate enrollment that are women

Feature	Description
longitude	The longitude of the institution
latitude	The latitude of the institution
stud_fac_ratio	Student-to-faculty ratio
avg_loan_amt_awarded	Average loan amount awarded to full-time first-time undergraduates
rev_grants_contracts_fte	Revenues from grants and contracts per full-time-equivalent students
instructional_fte	Full-time instructional staff
total_loan_amt	Total loan amount awarded to full-time first-time undergraduates
avg_act_score__x__avg_loan_amt_awarded	Interaction between average ACT scores and average loan amount
recs_required	Indicates if recommendations are required for admission
pct_awarded_fin_aid	Percent of full-time first-time undergraduates awarded any financial aid
test_scores_required	Indicates if standardized test scores (SAT, ACT, etc.) are required
year	Cohort year
avg_total_price	Average total price for all students
prev_grad_gap_race	Difference in previous year's graduation rate between whites and non-whites
prev_grad_gap_race_abs	Absolute value of the race graduation rate gap
prev_grad_ratio_race	Ratio of previous year's graduation rate of whites to non-whites
prev_grad_gap_gender	Difference in previous year's graduation rate between men and women
prev_grad_gap_gender_abs	Absolute value of the gender graduation rate gap
prev_grad_ratio_gender	Ratio of previous year's graduation rate of men to women
grad_rate	4-year graduation rate of full-time, first-time students seeking a bachelor's or equivalent degree.