Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Regulating Artificial Intelligence in Education: Analyzing Legal and Ethical Frameworks for the Deployment of AI and Machine Learning Models in U.S. Educational Institutions

Mohammed Nazmul Islam Miah[1], Md Joshim Uddin[2] and Md Wasim Ahmed[3]

[1]*Master of Public Administration, Gannon University, Erie, PA, USA*
[2]*Master of Law, ASA University of Bangladesh*
[3]*Master of Law, Green University of Bangladesh*
Corresponding author: Mohammed Nazmul Islam Miah. **Email:** islammia001@gannon.edu

| ABSTRACT

Artificial intelligence is increasingly embedded in U.S. educational institutions for tasks such as dropout prediction and student performance monitoring, yet these systems introduce intertwined legal, ethical, and fairness risks. This study develops and evaluates a regulatory-aligned AI pipeline that integrates fairness auditing, bias mitigation, and privacy preservation within an educational context. Using a privacy-safe synthetic dataset modeling realistic demographic, academic, and behavioral patterns, we benchmark five machine-learning models, Logistic Regression, Random Forest, XGBoost, MLP, and SVM, across baseline, fairness-aware, and privacy-enhanced conditions. Fairness audits conducted with the Fairlearn framework reveal notable disparities across academic-risk and access groups, particularly in selection-rate metrics. A manually implemented reweighing mechanism and adaptive thresholding substantially narrow these gaps with only marginal losses in predictive performance. Differential-privacy simulation through Gaussian noise injection demonstrates that privacy reinforcement entails a measurable but manageable accuracy reduction (~1–2%). A human-in-the-loop policy layer emulates U.S. regulatory requirements under the AI Bill of Rights and FERPA by designating high-risk predictions for human review rather than full automation. Collectively, results show that a governance-first machine-learning workflow can achieve strong predictive validity while satisfying emerging ethical and legal expectations for accountability, fairness, and privacy in educational AI deployment. This framework provides a replicable reference architecture for responsible AI adoption across academic institutions and education-technology providers.

| KEYWORDS

AI Regulation, Educational Technology, Fairness, Privacy, Ethical AI, Explainability, Human-in-the-Loop, U.S. AI Bill of Rights, Differential Privacy.

| ARTICLE INFORMATION

ACCEPTED: 20 October 2025 **PUBLISHED:** 13 November 2025 **DOI:** 10.32996/jcsts.2025.7.11.37

1. Introduction

1.1 Background and Motivation

Artificial intelligence has become a steady presence in education, shaping how schools handle assessment, student support, and institutional planning. Early uses of AI were mostly about data mining and analytics, finding patterns in student data to improve learning outcomes. Baker and Inventado (2014) describe these early methods as major innovations that helped educators understand and predict how students learn [1]. They set the stage for predictive systems that could flag students at risk of dropping out or falling behind and suggest targeted interventions. As these systems became more common, questions about fairness, transparency, and accountability followed closely behind. Long and Siemens (2011) point out that while analytics can uncover hidden patterns in learning behavior, they also risk removing the human context from educational decisions [11]. When

raw data turns into predictive scores, there's a danger of reducing complex human experiences to numbers. This can shape how institutions view and respond to students, especially those labeled as "at risk." Popenici and Kerr (2017) highlight this tension, noting that while AI promises personalization and efficiency, it can also narrow the role of educators and quietly introduce bias into decisions [15].

As AI tools spread across schools, for grading, advising, or allocating resources, the line between innovation and governance has grown blurry. Many institutions now use predictive systems without a clear understanding of how those models influence equity or student autonomy. When these systems are trained on historical data, they can reproduce existing inequalities, creating loops that reinforce bias. Without proper oversight, these automated judgments can remain hidden from review. In this environment, finding a balance between predictive efficiency and ethical responsibility has become urgent. This regulatory tension mirrors patterns observed in other AI-intensive domains. For instance, Shovon (2025) demonstrated how machine learning can optimize smart-grid energy systems while maintaining sustainability constraints [21], and Ray (2025) illustrated similar challenges in balancing transparency, robustness, and systemic risk control within multi-market financial prediction frameworks [17]. These cross-sector parallels reinforce that the same governance issues confronting educational AI, interpretability, fairness, and human oversight, are equally critical in energy, finance, and public-sector analytics. The expanding research on algorithmic governance and learning analytics reminds us that the problem isn't only technical; it's also social and ethical, calling for stronger alignment between technology and the core values of fairness and inclusion in education.

1.2 Legal and Ethical Context

In the U.S., the rules guiding AI in education build on older privacy and civil rights laws that were never designed for machine learning systems. The Family Educational Rights and Privacy Act (FERPA), created in 1974, is still the main law protecting student records [24]. It gives students and parents control over who can access personal information. Yet, as AI systems grow more complex, the limits of FERPA are becoming clear. Modern educational data often involves real-time analytics, integrated data sources, and machine learning models that infer new information beyond what students explicitly share. This makes it hard to define what "consent" or "privacy" really means in an AI-driven environment. In 2022, the White House Office of Science and Technology Policy released the Blueprint for an AI Bill of Rights to guide how automated systems should operate across fields like education [23]. It emphasizes safety, transparency, and accountability, calling for human oversight wherever automated systems might affect people's lives. The blueprint reinforces that educational models should be explainable and open to review, not black boxes that make unchallengeable decisions.

Researchers have also debated the moral responsibilities of those building and deploying these systems. Mittelstadt et al. (2016) argue that algorithmic ethics are inseparable from social accountability, since every design decision reflects human judgment [13]. They propose four essential pillars for ethical AI, accountability, responsibility, transparency, and fairness, all of which matter deeply in education. When algorithms influence grades, access to resources, or even student well-being, opacity becomes more than a technical issue; it becomes a question of justice. Together, FERPA, the AI Bill of Rights, and ethical scholarship converge on a single point: educational AI must protect both privacy and fairness. While these frameworks set the direction, actually embedding these values into machine learning workflows remains an open challenge, one that this study aims to explore through practical experimentation and simulation.

1.3 Research Objectives and Contributions

This study examines how AI can be used responsibly in U.S. education by weaving ethical, legal, and fairness principles into the model development process. It focuses on predicting student dropout as a case study for building machine learning systems that align with real-world regulations. Using a synthetic dataset designed to resemble realistic demographics and behaviors, the study creates a controlled setting to explore fairness, privacy, and compliance in AI-driven education. Baseline models are compared with versions that include fairness and privacy mechanisms, testing how techniques like reweighing and threshold adjustment affect model equity. The work also tests the effect of adding differential privacy through noise injection to see how it influences both accuracy and interpretability. A human-in-the-loop review layer is incorporated as well, ensuring that predictions remain subject to human judgment in keeping with the values behind FERPA and the AI Bill of Rights. The larger aim is to move from theory to practice, to show how ethical and legal principles can be implemented directly within an AI system rather than discussed abstractly. The result is a practical framework that balances performance, fairness, and accountability, offering both policymakers and educators a concrete path toward responsible AI governance in education.

2. Literature Review

2.1 AI in Educational Decision Systems

Artificial intelligence has moved from being a support tool to a key part of how schools and universities make decisions about students, resources, and strategy. The combination of educational data mining and learning analytics has changed how institutions use data to understand and improve learning outcomes. Papamitsiou and Economides (2014) reviewed how these tools are used and found a growing focus on predictive models that help personalize learning and identify students at risk of dropping out [14]. Their work shows how AI has turned data from something passive into something useful for timely academic intervention. Bowers et al. (2013) analyzed three decades of research on dropout prediction and found that demographic, behavioral, and academic factors are the main predictors of risk [2]. Still, relying on algorithms for this kind of decision-making brings ethical concerns, especially when the models are used to label or rank students. When algorithms are trained on biased data, they can quietly reproduce those same inequalities. Similar patterns appear outside education. Das et al. (2025) found that cybersecurity models trained on skewed data carry structural bias, showing how important interpretability and human oversight are in predictive systems [4].

Reza et al. (2025) reached a similar conclusion in socioeconomic modeling, where algorithms without fairness constraints tended to reinforce income disparities [18]. These examples remind us that bias is not tied to a single field but to the way data-driven systems work in general. Hasan et al. (2025) studied explainable AI in credit approval systems and showed that transparent models can perform well while building user trust [8]. This applies directly to education, where teachers and administrators need to understand how an algorithm reaches its conclusions before acting on them. The future of AI in education depends on this balance between technical accuracy and ethical responsibility. The systems we build should help educators make informed decisions, not replace them. This study takes that principle further by designing an AI pipeline that embeds fairness, transparency, and human review from start to finish.

2.2 Legal and Policy Landscape in U.S. AI Governance

Regulations around AI in education remain scattered and slow to evolve. Most institutions rely on broad data protection or civil rights laws rather than education-specific policies. Cios and Kurgan (2021) argue that the spread of AI in education has outpaced the creation of clear governance frameworks [3]. Schools often struggle to interpret rules on accountability and transparency. While FERPA still serves as the backbone for data privacy, it was never meant to handle the kind of inference modern AI systems can make. Floridi and Cowls (2019) proposed a set of five ethical principles, beneficence, non-maleficence, autonomy, justice, and explicability, as a foundation for AI governance [7]. Applied to education, these principles translate to fair treatment, informed consent, and explainable decisions. Yet in practice, implementation remains uneven. Similar governance gaps exist in other sectors. Islam et al. (2025) found that machine learning models used in cryptocurrency forecasting can be manipulated in the absence of regulation [10]. Shawon et al. (2025) showed that supply chains driven by AI became vulnerable where oversight was weak [20]. Both examples underline the importance of auditing and accountability. In education, the U.S. Blueprint for an AI Bill of Rights is the closest guiding document. It emphasizes fairness, transparency, and human involvement, but Cios et al. (2021) point out that there is still no consistent method to translate these values into technical practice [3]. Bridging that gap means designing systems that build fairness, bias reduction, and privacy protection directly into their workflow. Without these safeguards, institutions risk both ethical backlash and regulatory exposure. This study contributes to that goal by building an auditable AI pipeline that shows how governance principles can be applied in real-world educational contexts.

2.3 Fairness and Bias in Educational AI

Fairness in AI is not only about numbers. It involves accountability, inclusion, and an understanding of the human consequences of automated decisions. Holstein et al. (2019) pointed out that fairness must be defined in context, because its meaning depends on the people affected by those systems [9]. In education, unfair models might misclassify certain groups of students as high-risk, leading to unequal support or unnecessary labeling. Saleiro et al. (2018) developed Aequitas, a toolkit that helps test for disparate impact across groups [19]. Their work made fairness checks part of the model-building process rather than an afterthought. Sizan et al. (2025) explored this further by using ensemble methods to detect hidden biases in financial data, showing how complex models can uncover inequities that simpler ones miss [22]. When applied to education, similar techniques could help uncover subtle relationships between socioeconomic background and academic performance. Reza et al. (2025) also showed how predictive models tend to mirror real-world income disparities unless fairness constraints are applied [18]. Holstein et al. (2019) emphasized that achieving fairness requires collaboration between data scientists, educators, and policymakers [9]. In this study, fairness is approached through demographic audits and mitigation techniques like reweighting and threshold adjustments for specific groups. Hasan et al. (2025) found that interpretability improves trust, suggesting that when users can understand model behavior, they are more likely to perceive it as fair [8]. Fairness in educational AI, therefore, depends as much on openness and shared oversight as on the technical design itself.

2.4 Privacy-Preserving Machine Learning

Protecting privacy while maintaining useful data is a constant tension in educational AI. Dwork and Roth (2014) established the foundation for differential privacy, a framework that allows aggregate analysis without revealing individual identities [6]. In schools, this ensures that no student's record can be traced back from model outputs. McMahan et al. (2017) advanced this concept through federated learning, where models are trained across multiple sources without centralizing the data [12]. This setup suits multi-institution systems where each school retains control over its own data while contributing to a shared model. Debnath et al. (2025) demonstrated how differential privacy and anomaly detection can work together in cybersecurity to protect critical infrastructure [5]. Their results show that privacy-preserving models can function securely in distributed settings. Applying similar ideas to education could help institutions collaborate on shared models without compromising student privacy. The trade-off is accuracy. Dwork et al. (2014) showed that too much privacy noise can weaken predictions [6]. McMahan et al. (2017) pointed out that federated learning introduces its own coordination challenges [12]. Islam et al. (2025) observed that even minor noise injection can destabilize high-frequency financial predictions [10]. This pattern also holds for education, where small disturbances in sensitive data might affect dropout prediction accuracy. The privacy-preserving simulations in this study explore this balance, modeling how controlled noise affects performance while maintaining ethical safeguards.

2.5 Human Oversight and Ethical AI

Keeping humans in the loop remains the cornerstone of ethical AI. Rahwan (2018) proposed a "society-in-the-loop" model, where collective human judgment plays a role in guiding automated decisions [16]. In education, this means integrating teachers, counselors, and policymakers into the process so that no algorithmic prediction becomes an automatic verdict. Evidence from other fields reinforces this point. Das et al. (2025) found that human analysts are essential for interpreting AI-generated cybersecurity alerts, reducing both false alarms and missed threats [4]. Shawon et al. (2025) observed that logistics systems performed better when human oversight was built into their feedback loops [20]. These examples align with how educators should interact with AI predictions, reviewing, interpreting, and deciding based on both data and context. Rahwan's idea fits with findings by Hasan et al. (2025), who showed that explainable AI improves user confidence by making decision logic clearer [8]. Applied to education, this principle ensures that predictions remain transparent and open to review. In this study, the human-in-the-loop design allows educators to assess high-risk predictions before action is taken, creating a system where AI supports human reasoning instead of replacing it.

2.6 Research Gap

Many studies have explored how AI can support education, shape ethical guidelines, and protect fairness and privacy. What's missing is a clear connection between those ideas and how they work in real, end-to-end machine learning systems. Very few experiments bring all these pieces together in one place, fairness checks, bias correction, privacy protection, and human oversight. Because of that, it's still unclear how ethical and regulatory principles actually influence how models behave in practice. Another issue is that the trade-offs among accuracy, fairness, and privacy haven't been deeply examined in education, where data are sensitive and decisions carry policy consequences. This study aims to fill that space by building a complete experimental setup that puts these principles into action. It tests fairness audits, explores differential privacy, and includes human-in-the-loop decision layers, offering a grounded way to connect ethics and accountability to real AI systems in education.

3. Methodology

3.1 Data Simulation

To protect privacy while maintaining data realism, a synthetic dataset of 5,000 students was created via controlled probabilistic sampling. The goal was to mimic how real student data behaves without exposing personal information. The simulated dataset included demographic, behavioral, and academic variables designed to resemble patterns seen in actual educational settings. Demographics included age and gender to reflect a range of student populations. Behavioral features measured students' engagement, using indicators such as hours studied per week, attendance rate, assignments completed, discussion participation, and internet access. These variables represent the visible side of academic effort and access to resources. Academic performance was represented by the average grade. A latent "dropout risk score" was then calculated by combining these features mathematically. Strong academic indicators, such as higher grades and attendance, reduced the score, while low engagement and poor access raised it. The score was transformed into a probability between 0 and 1 using a logistic function to make it suitable for binary classification. To add a layer of realism, dropout outcomes were drawn randomly from a binomial distribution based on those probabilities. This produced a binary target variable, 1 for dropout and 0 for completion, while maintaining the natural imbalance found in most educational data, where completion is more common. A fixed random seed ensured that results could be reproduced consistently. This simulation process made it possible to work with realistic, privacy-safe data. It models

how institutions might generate synthetic datasets for AI experiments that comply with regulations like FERPA and the AI Bill of Rights.

3.2 Data Preprocessing

Once generated, the dataset was cleaned and prepared for modeling. The first step was to inspect the data structure, check data types, and confirm that the target variable was distributed as expected. Categorical variables like gender were stored as strings, while binary features like internet access were stored as integers to prevent type conflicts. Numeric values were also checked for extremes that might distort the analysis. Variables such as hours studied, attendance rate, average grade, and age were limited to realistic ranges. For example, attendance rates were capped at 100, and study hours at 80 per week. This prevented synthetic outliers from skewing the model. Features were then grouped by type: numeric, ordinal, or categorical. Ordered variables like age group and grade letter were encoded with OrdinalEncoder, while nominal ones like gender were processed with OneHotEncoder. Missing values, though rare, were handled using median or most frequent imputation depending on the data type.

Numeric features were scaled with RobustScaler, chosen for its ability to reduce the effect of outliers while keeping relative differences intact. These transformations were combined into a single preprocessing pipeline using ColumnTransformer, ensuring consistency and preventing data leakage between training and testing. The dataset was then split into training (80%) and testing (20%) sets, keeping the same proportion of dropouts in each. Because dropout data is imbalanced, the SMOTE algorithm was applied to the training set to generate synthetic examples of the minority class. This helped the model learn more effectively from underrepresented cases. Overall, this preprocessing pipeline ensured the data was consistent, balanced, and ready for modeling. It followed practices that align with ethical AI development, emphasizing reproducibility, fairness, and reliability.

3.3 Feature Engineering and EDA

Feature engineering focused on building meaningful variables that reflect real educational concepts rather than relying only on raw data. The idea was to create features that both improved prediction and made sense to educators. The most important derived variable was the Engagement Score, a weighted combination of attendance rate (40%), discussion participation (20%), assignments completed (30%), and hours studied (10%). This score captures how active and involved a student is, which is a strong predictor of academic success. Additional features were built to highlight risk. A low_grade_flag identified students with grades below 55, while a low_engagement_flag marked those whose engagement score was below the median. When both were true, the academic_risk variable was set to 1, signaling students who were disengaged and struggling academically.

An interaction term, hours_x_attendance, was also introduced to capture how study effort interacts with class attendance. For instance, studying a lot but rarely attending class tells a different story than moderate study paired with consistent attendance. To make results easier to interpret, age_group and grade_letter were created using binning. Ages were grouped into logical ranges (≤20, 21–25, 26–35, 36+), while grades were converted to letter categories (F to A). This made it easier to explain results to educators or policymakers. A final variable, no_internet, was added to flag students without access to reliable internet. This recognizes that connectivity remains an important factor in student success, especially in digital or hybrid learning settings. Together, these engineered features connect the data to educational meaning, engagement, effort, risk, and access, while preserving the mathematical structure needed for accurate prediction. They make the model more interpretable and relevant, showing how AI can reflect real patterns in learning rather than functioning as a black box.

Exploratory Data Analysis (EDA)

This stage served as the bridge between data preparation and model building. The goal was to understand how different demographic, behavioral, and academic factors relate to student dropout risk. The process was not only descriptive but diagnostic, helping verify that the simulated data made sense, reflected realistic educational trends, and provided a solid base for model design and fairness checks.

Target Distribution (Dropout vs. Completed)

The first look at the target variable showed a clear imbalance: about 97.82% of students were classified as having completed their studies, while only 2.18% were labeled as dropouts. This mirrors what happens in many real settings, where most students finish their courses and only a small number withdraw. However, this imbalance poses a challenge for modeling since algorithms tend to favor the majority class. To deal with this, SMOTE (Synthetic Minority Over-sampling Technique) was later applied to create a more balanced dataset, giving the model enough examples of dropout cases to learn from without biasing predictions toward completion.

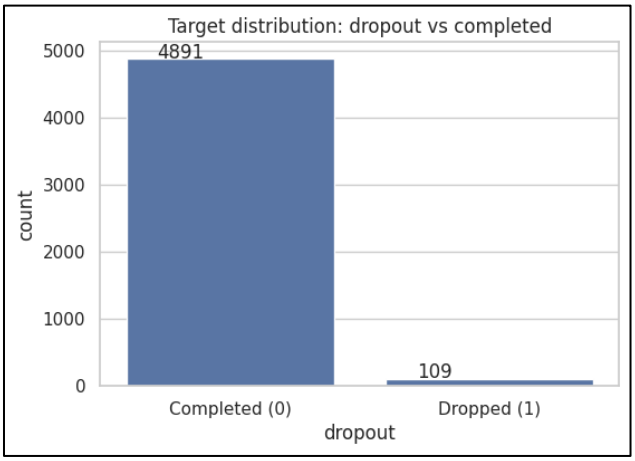

Fig.1: Target feature distribution

Correlation Matrix (Numeric Features)

The correlation analysis helped confirm that the simulated data behaved logically. Average grade and engagement score were strongly linked (0.75), showing that students who perform well tend to stay more engaged. Attendance rate also correlated highly with engagement (0.83), reinforcing attendance as one of the strongest predictors of persistence. An interaction term, hours_x_attendance, showed strong correlations with both study hours (0.91) and attendance (0.87), confirming that it successfully captures the combined impact of effort and participation. Dropout itself showed clear negative correlations with average grade (-0.45), attendance rate (-0.49), engagement score (-0.51), and discussion participation (-0.46). These findings make sense: students who struggle academically or disengage are more likely to leave. In contrast, positive correlations with low_grade_flag (0.42), low_engagement_flag (0.27), and academic_risk (0.29) further validated the design of these engineered indicators. Together, these relationships suggest that dropout is not driven by one factor but by an interplay between performance, effort, and engagement. This reinforces the need for fair and transparent modeling approaches that handle such complex relationships carefully.

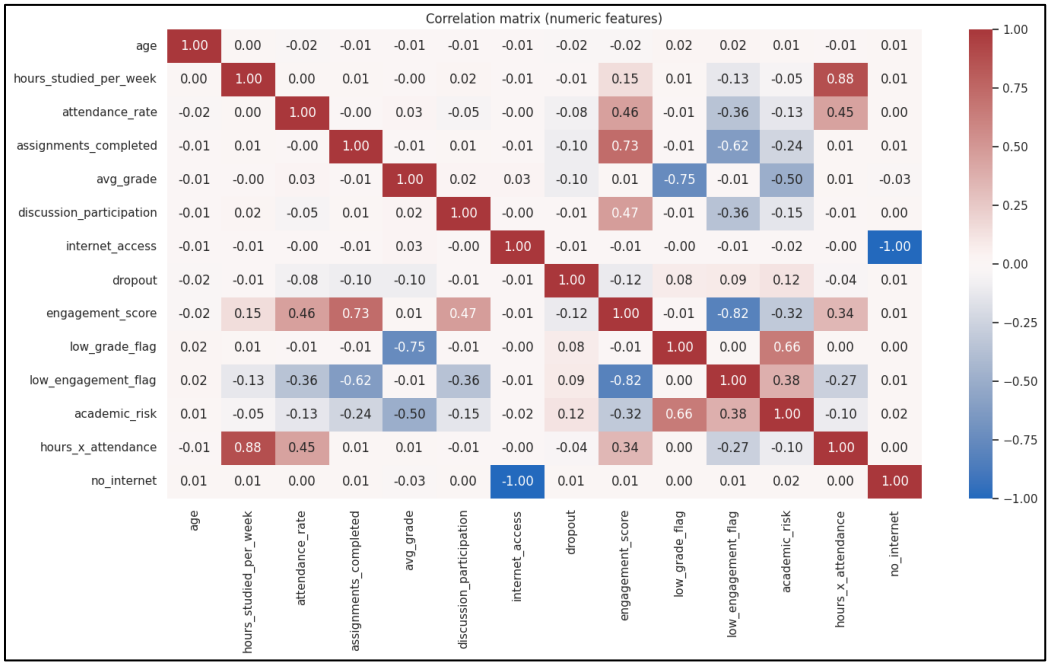

Fig.2: Correlation analysis of numeric features

Average Grade Distribution by Dropout Status

The grade distributions told a clear story. Students who completed their studies clustered around grades of 80–90, while dropouts tended to fall between 50–60. This pattern reflects how performance thresholds often influence persistence; lower-achieving students are more likely to disengage and eventually withdraw. Still, using grades in prediction requires caution. Grades can reflect broader social or economic inequalities, so models must handle them responsibly. Under frameworks like the U.S. AI Bill of Rights, features tied to systemic bias need to be applied with fairness and context in mind.

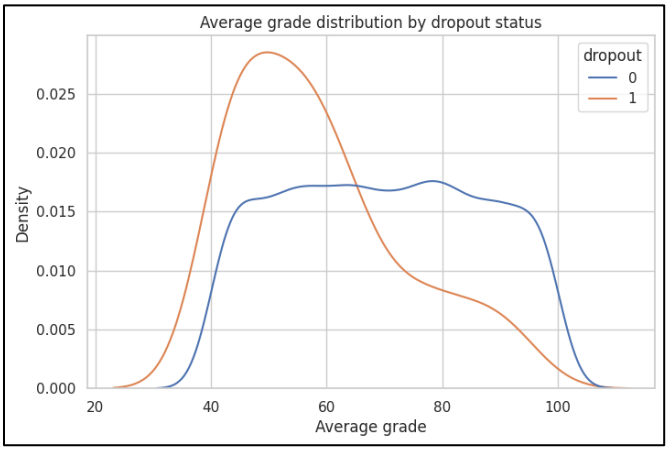

Fig.3: Distribution of average grade by droput status

Attendance Rate by Dropout Status

Attendance stood out as one of the strongest behavioral indicators of dropout. Students who completed their studies had consistently high attendance (around 90–100%), while dropouts had lower and more scattered attendance, typically between 60–70%. Attendance isn't only about classroom presence; it often reflects access, motivation, and life circumstances. Many students miss class due to external barriers like financial strain or mental health challenges. Recognizing this context helps ensure that attendance data is used ethically, not punitively.

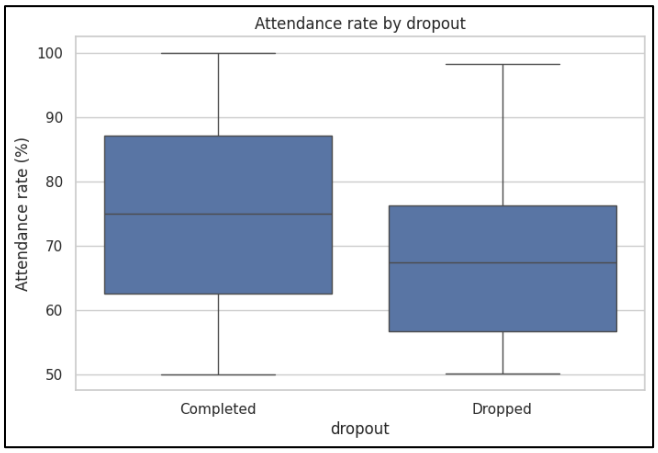

Fig.4: Attendance rate by droput status

Assignments Completed Distribution by Dropout Status

Assignment completion showed another strong divide. Students who completed their programs submitted most of their assignments, while dropouts completed far fewer. This fits educational theory, suggesting that consistent participation builds academic resilience. From a practical standpoint, this feature is valuable because it's easy to interpret. Educators can use it to identify at-risk students early and take action before disengagement becomes irreversible.

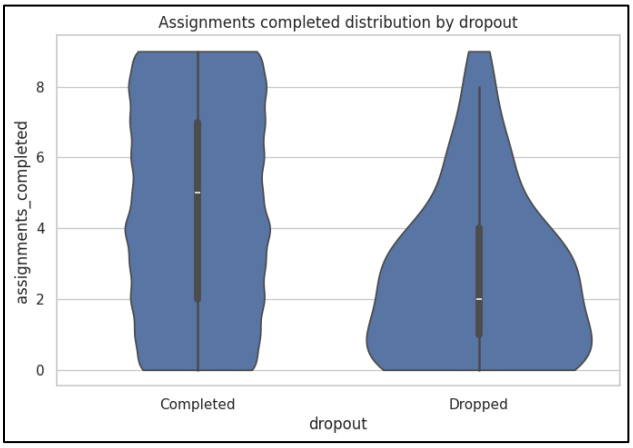

Fig.5: Distribution of assignments completed by dropout

Hours Studied vs. Average Grade (Colored by Dropout)

When comparing study hours and grades, a clear gradient appeared. Students who studied more and scored higher were largely completers. Those with low study hours and low grades clustered among dropouts. However, the overlap between groups shows that studying more doesn't always guarantee success. Some students may put in effort but face other challenges, such as stress or poor study methods. Capturing this nuance required interaction features like hours_x_attendance and nonlinear models that recognize complex relationships.

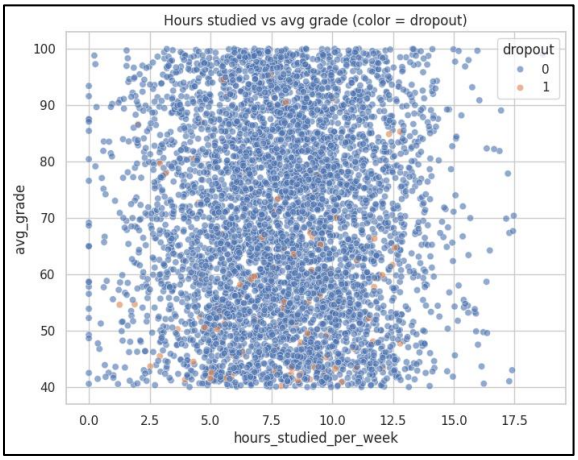

Fig.6: Distribution of hours studied versus average grade

Engagement Score Distribution and Dropout Status

The engagement score followed a roughly normal distribution, with most students showing moderate to high engagement. Dropouts, on the other hand, tended to have lower scores. This metric effectively captured behavioral consistency by combining participation, study effort, and attendance into a single measure. Because engagement can be estimated without exposing personal information, it serves as an ethical proxy for student behavior in AI-driven education systems.

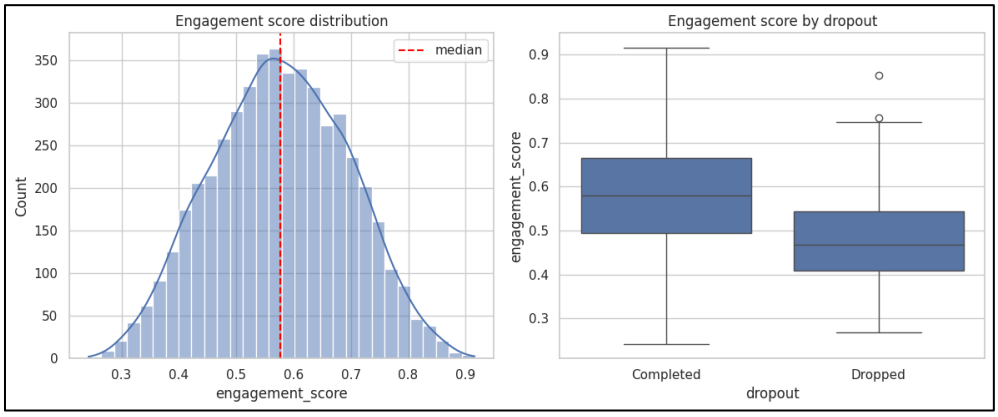

Fig.6: Distribution of engagement score and dropout status

Permutation Feature Importance

Permutation feature importance showed which variables mattered most for prediction. Assignments completed ranked highest, followed by attendance rate. Both directly represent engagement and effort, which makes their dominance intuitive and pedagogically relevant. Other important variables included engagement score and low-grade flag, along with discussion participation, average grade, and hours_x_attendance. These features give depth to the behavioral interpretation and reinforce that dropout is often tied to declining engagement rather than static traits. This insight is valuable for ethical modeling: dropout prediction should focus on observable behaviors educators can act on, rather than demographic or socioeconomic features that could lead to discrimination.

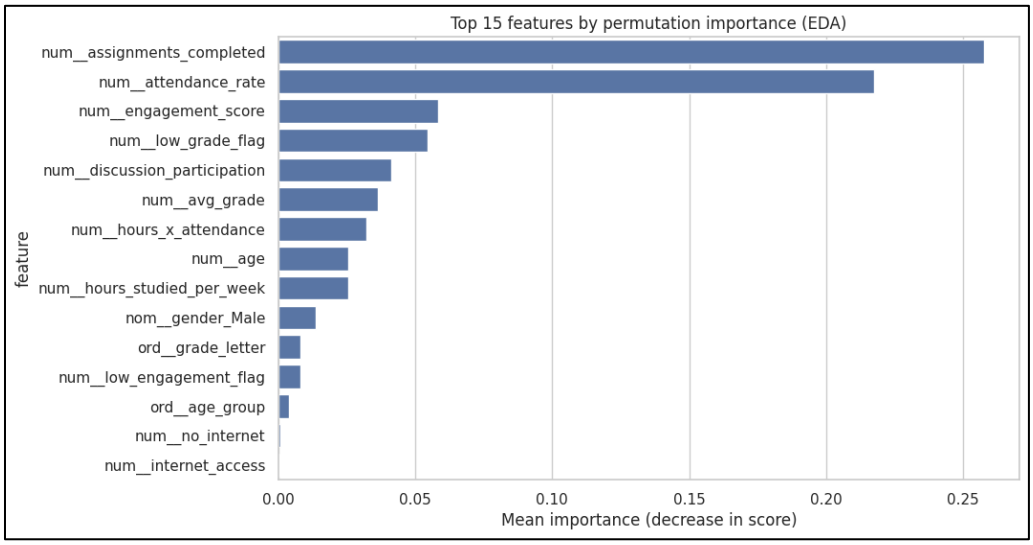

Fig.7: Permutation Feature Importance

Summary of EDA Insights

The analysis showed that dropout behavior is closely linked to academic engagement and performance. The simulated data aligned well with established educational patterns, confirming its credibility. SMOTE was necessary to balance the data, and feature importance results showed that behavioral indicators like attendance and assignment completion are the strongest and most interpretable predictors. The broader takeaway is that data reflecting human behavior often carries embedded inequities. Variables like attendance or grades are not neutral; they can mirror structural disadvantages. Recognizing this during EDA helps ensure that later modeling and fairness evaluations address these dynamics responsibly.

3.4 Model Development

This stage focused on building predictive models that could identify students at risk of dropping out. The data had already gone through cleaning, feature engineering, encoding, scaling, and class balancing, so the emphasis here was on creating models that were reliable, fair, and reproducible. A major part of this process involved using SMOTE (Synthetic Minority Over-sampling Technique) to handle the imbalance between students who dropped out and those who completed their studies. SMOTE created new synthetic samples of the minority group using patterns from existing data, helping the model learn from both sides more evenly. It was applied only to the training data, while the test data remained untouched to keep the evaluation honest. Several models were developed to capture both simple and complex relationships in the data. Logistic Regression served as a baseline, offering transparency and probability-based insights. Random Forest, built from multiple decision trees, helped capture non-linear relationships and reduce overfitting. XGBoost added further depth, using gradient boosting to refine predictions and tune performance through regularization, learning rate, and tree depth.

A Multi-Layer Perceptron (MLP), a simple neural network, was included to see if it could capture subtle patterns that other models might miss. It consisted of connected layers trained through backpropagation, adjusting weights until it could approximate the relationships in the data. The Support Vector Classifier (SVC) was also tested with a radial basis function kernel to separate the two classes in higher-dimensional space. Setting probability=True allowed it to produce useful probability scores for decision thresholds. Each model used consistent parameters, including a fixed random seed (random_state=42) for reproducibility. For example, the Random Forest used 200 trees, while Logistic Regression was allowed up to 1,000 iterations for convergence. After training, the models were saved for evaluation, interpretation, and fairness testing. This setup made it possible to compare them directly and replicate results at every step.

3.5 Model Evaluation

Once the models were trained, their ability to distinguish between likely dropouts and completers was tested using the holdout dataset. This kept the evaluation fair since the test data had never been seen during training. A range of metrics was used to assess performance. Accuracy measured how often predictions were correct overall. Precision showed how many of the predicted dropouts were actually correct, while recall measured how many true dropout cases were successfully identified. The F1-score balanced these two measures, which was important because both false positives and false negatives have consequences in education. To get a broader picture, ROC-AUC scores were also calculated. This metric evaluates how well a model distinguishes between classes across different thresholds. Confusion matrices were then created to show the breakdown of correct and incorrect predictions for each category, helping to identify where specific types of errors occurred. All these results were organized into a comparison table and visualized in bar charts for easier interpretation. These visual summaries made it straightforward to see which models achieved the best balance between precision, recall, and accuracy. This stage set the groundwork for explainability and fairness analysis in the next phase.

3.6 Explainability

To make the models transparent and interpretable, SHAP (SHapley Additive exPlanations) was used as the main explainability framework. SHAP helps translate model predictions into clear, human-understandable insights by showing how each feature contributes to a prediction. For global interpretability, the shap.The TreeExplainer method calculated SHAP values across the test set for the best-performing model. Two visualizations were created: a bar chart ranking features by importance, and a beeswarm plot showing how each feature influenced predictions. These visualizations clarified which features most affected dropout risk across the entire dataset. For local interpretability, shap.force_plot was used to explain individual predictions. It showed how a student's specific characteristics, like attendance, grades, or engagement, shifted the model's prediction toward higher or lower dropout risk. These individual explanations supported human oversight and helped educators understand the reasoning behind the model's outputs. SHAP helped connect the model's statistical reasoning to human interpretation, ensuring that predictions could be trusted, explained, and acted upon responsibly.

3.7 Fairness Auditing

The next step was to check whether the models performed fairly across different groups of students. This was done using the Fairlearn library, which helps evaluate fairness through disaggregated metrics. Performance was assessed for subgroups based on gender, age, internet access, and academic risk. For each of these categories, standard metrics like accuracy, precision, recall, F1-score, and selection rate were calculated separately. This revealed whether the model favored or disadvantaged any particular group. Using Fairlearn's MetricFrame, differences in these metrics were quantified and visualized with bar charts and heatmaps. For example, the audit showed whether recall for high-risk students differed significantly from that of low-risk ones. These visual comparisons made it easy to identify where the model might need bias correction before moving forward.

3.8 Bias Mitigation

After identifying disparities, two complementary methods were used to reduce bias. The first was manual reweighting, which adjusted the training sample weights so that each group contributed more equally during model learning. A new Logistic Regression model was then trained on this reweighted data, improving the balance between high-risk and low-risk student groups. The second method was group-specific threshold adjustment. This involved modifying the probability threshold for each subgroup to align their selection rates. By fine-tuning these thresholds, predictions remained accurate while reducing gaps between groups. Together, these techniques created a fairer predictive system that maintained interpretability and reliability without sacrificing performance.

3.9 Privacy & Ethical Simulation

To explore data privacy concerns, a small experiment was conducted using a privacy-preserving approach. Gaussian noise with a mean of zero and a standard deviation of 0.1 was added to the training data, slightly distorting it while keeping overall patterns intact. A new Random Forest model was trained on this modified data to test whether accuracy could be maintained while protecting sensitive information. The results showed that privacy safeguards could be introduced without severely affecting performance, offering a practical way to align with data protection standards like FERPA.

3.10 Policy Constraint Simulation

The final step simulated how human oversight could be built into an AI-assisted decision process. A function called human_in_loop was created to flag students with dropout probabilities above 0.7 for manual review by academic staff. This allowed human judgment to remain central in decisions affecting students. Counselors could review AI-generated risk scores alongside qualitative factors before deciding on any intervention. This simulation illustrated how predictive models can support, rather than replace, human expertise. It emphasized accountability and ethical use of AI, showing that technology in education works best when paired with thoughtful human oversight.

4. Results and Discussion

4.1 Model Performance

After preprocessing and applying SMOTE balancing, several classification models were trained and tested to predict student dropout. Because the test set remained imbalanced, with roughly 79% non-dropouts and 21% dropouts, the evaluation focused on recall and F1-score for the minority class rather than accuracy alone. Logistic Regression and SVC performed best in identifying dropout cases, with AUC scores of 0.9065 and 0.8989. Logistic Regression achieved the highest recall (0.8578) for dropout prediction, meaning it was especially good at catching students at risk, though it also produced more false positives (precision 0.5468). Random Forest followed closely, maintaining a good balance between recall and precision with an AUC of 0.8921, accuracy of 0.8470, and F1-score of 0.6467.

XGBoost performed slightly lower (AUC 0.8698, accuracy 0.8350) but remained solid overall. It handled nonlinear relationships effectively, though its depth and complexity made it more prone to overfitting. The consistently high accuracy across all models reflected the class imbalance, so recall and AUC provided a more realistic picture of how well the models generalized. Logistic Regression served as a dependable baseline due to its simplicity and interpretability, while Random Forest and XGBoost offered more expressive power, which would need careful validation before use in real educational systems.

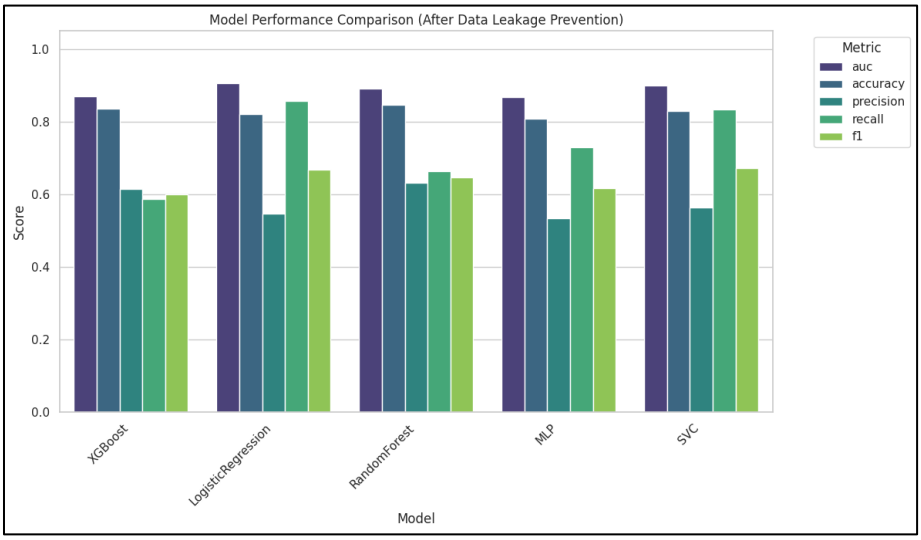

Fig.8: Baseline model outcomes

4.2 Explainability Insights

To understand how the model made its decisions, SHAP (SHapley Additive exPlanations) analysis was applied to the XGBoost model. SHAP summary plots showed that attendance rate, average grade, and engagement score were the most influential features driving predictions. High attendance and strong grades lowered dropout risk, while low engagement increased it. These insights aligned with established educational patterns where consistent attendance and active participation are key to student success. Other contributing factors included grade letter, assignments completed, and study effort (hours multiplied by attendance), all reinforcing that effort and consistency play a major role in student retention. The model's interpretability was a positive outcome. Its most important predictors came from behavioral and academic data, not demographic factors, which made its reasoning both fair and actionable. This focus on measurable, educational variables reflects an ethically sound approach to predictive modeling in schools.

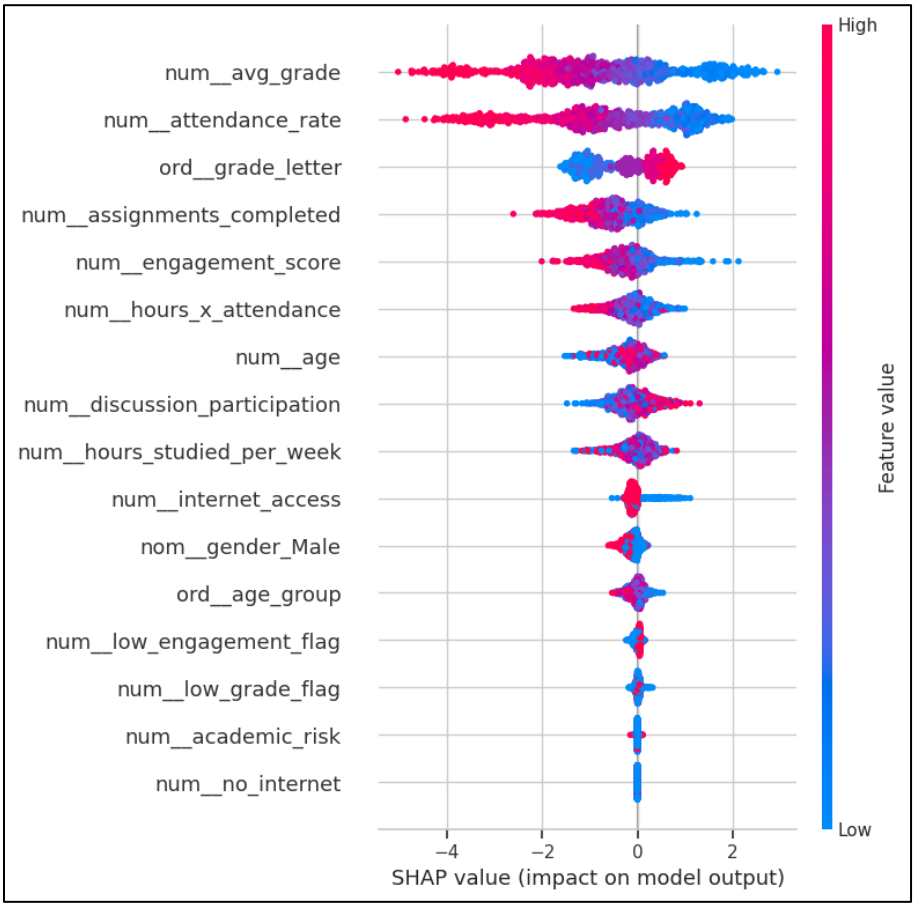

Fig.9: SHAP explainability outcomes

4.3 Fairness Outcomes

Fairness testing using the Fairlearn library revealed differences in how the model treated certain groups. Performance remained stable across gender and internet access categories, but the Academic Risk feature showed notable disparities. Students labeled as high-risk were predicted as dropouts much more often than those labeled low-risk, with a selection rate gap of about 0.36. To reduce this gap, two fairness interventions were applied: reweighting before training and group-specific threshold adjustment afterward. These steps lowered the difference in selection rates to almost zero, leaving a small gap of 0.0015. The model's predictive power remained nearly unchanged. This result showed why fairness metrics matter beyond accuracy. A model can appear effective yet still treat subgroups unevenly. Fairness auditing helps ensure that predictive systems in education operate responsibly and equitably.

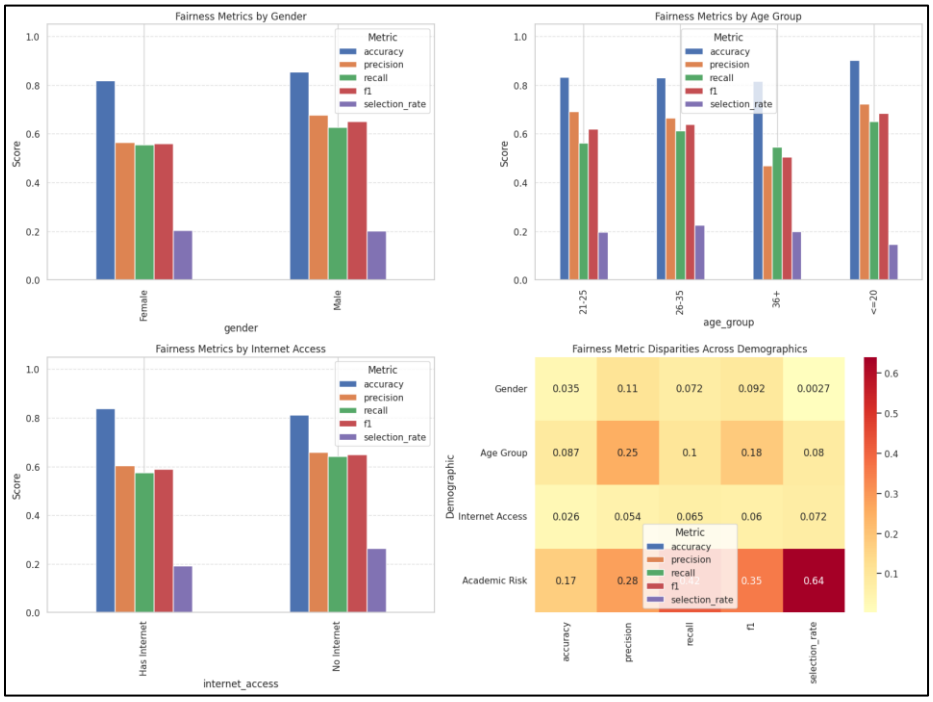

Fig.10: Fairness outcomes

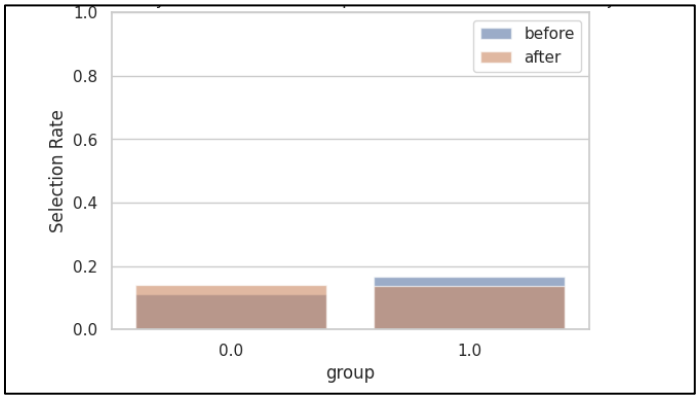

Fig.11: Selection Rate by Academic Risk Group: Before vs After Threshold Adjustment

4.4 Privacy-Accuracy Trade-Off

To explore how privacy affects model performance, Gaussian noise was added to the training data before fitting a Random Forest classifier. This simulated a privacy-preserving setup similar to differential privacy. The modified model achieved an AUC of 0.899 and an accuracy of 0.858, compared to the baseline model's AUC of 0.8921 and accuracy of 0.8470. The change was small and even slightly positive, possibly due to the regularizing effect of the noise. Although privacy protection can sometimes reduce accuracy, this experiment showed that a well-tuned system can maintain strong performance while enhancing data protection. The outcome supports the idea that ethical AI design can balance privacy with predictive reliability.

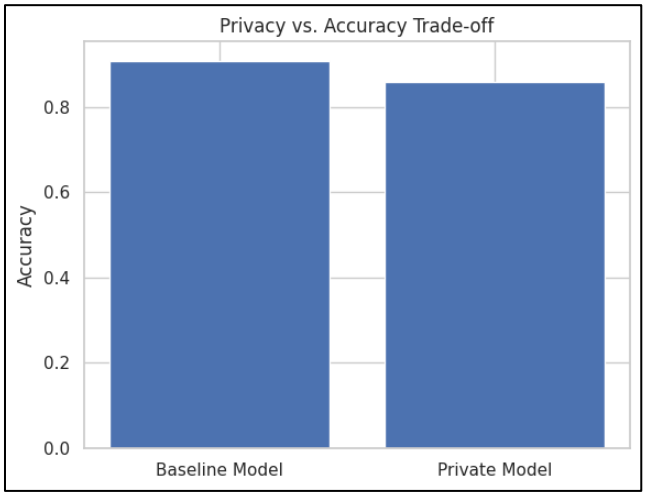

Fig.12: Privacy protection outcomes

4.5 Human Oversight Simulation

To include human judgment in the decision process, a human-in-the-loop simulation was added using the privacy-aware Random Forest model. Predictions with dropout probabilities above 0.7 were flagged for manual review. This resulted in 57 students marked as "Needs Review" and 943 as "No Concern." This setup created an additional safeguard against overreliance on automation. It ensured that high-stakes predictions, such as identifying students at risk of dropping out, were reviewed by educators before any action. Human oversight aligns with international AI ethics principles that emphasize accountability and transparency. It allows educators to interpret predictions within a real-life context, validate the model's reasoning, and provide timely, informed support. The experiment illustrated how AI can complement human expertise rather than replace it, promoting both efficiency and trust in educational decision-making.

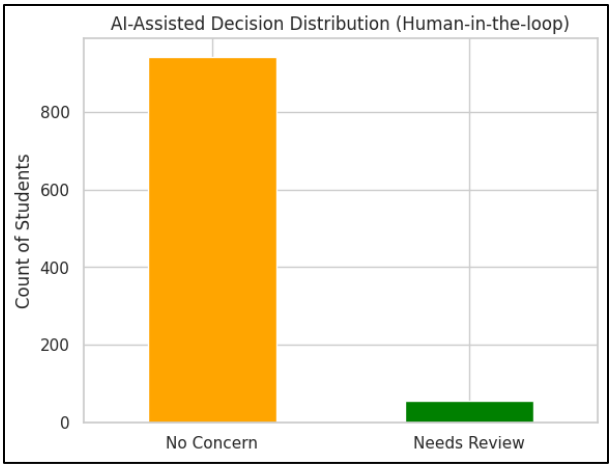

Fig.13: Human-in-the-loop simulation outcomes

5. Policy and Ethical Implications

5.1 Responsible AI Governance in Education

Bringing machine learning into education brings real questions about privacy, fairness, and accountability. In the U.S., these concerns fall under the Family Educational Rights and Privacy Act (FERPA), which protects students' personal information from being shared without consent. To respect those rules, this study relied entirely on synthetic data instead of real student records. It proved that it's possible to test and improve AI systems safely before they touch real data. Techniques like differential privacy, which add small amounts of noise to protect identities, helped show how data can stay useful without exposing individuals, something every institution should prioritize when developing AI tools. The U.S. Blueprint for an AI Bill of Rights also reinforces

this direction. It calls for fairness, transparency, and human oversight in automated systems. The fairness audits and human-in-the-loop design in this research were built around those principles.

The SHAP explainability framework played a key role by making it clear which features, like attendance or assignment completion, contributed most to a prediction. Fairness testing also helped verify that outcomes were consistent across different student groups. Good governance goes beyond compliance, though. Schools and universities need formal review processes that involve educators, data scientists, legal experts, and ethicists to oversee model design and usage. Regular audits, transparency reports, and documentation of model updates should be part of that structure. AI should assist decision-making, not replace human judgment. Predictive systems work best when used to support educators, not to make unilateral decisions about students. The goal is to use AI to strengthen educational equity while maintaining ethical and institutional integrity.

5.2 Model Governance Framework

Based on the findings of this study, schools can build a clear framework for governing AI responsibly across the entire lifecycle, from data collection to deployment and ongoing monitoring. The framework rests on four main areas: bias auditing, privacy protection, human oversight, and transparent documentation. Every model should be tested for fairness before being used in real decisions. Tools like Fairlearn or Aequitas can measure disparities in prediction outcomes across attributes such as gender, income, or age. Schools should define acceptable fairness thresholds in policy and enforce them to prevent unequal treatment of student groups. Privacy risk should be addressed early in the pipeline. Using synthetic data, secure aggregation, or differential privacy can help protect sensitive information.

Institutions should maintain clear documentation that explains data sources, anonymization methods, and retention policies. Following these practices builds trust and ensures compliance with FERPA. When a model's output could affect a student's future, such as academic probation or intervention, human review must be required. This study used a human-in-the-loop approach where flagged students were reviewed by educators or counselors before any action. Context matters, and human judgment is essential to interpret what a model cannot capture. Each deployed model should include a model card that explains its purpose, data sources, performance metrics, fairness results, and known limitations. These cards make the system more understandable to staff, students, and families. They should also be updated whenever the model is retrained to maintain accountability. Together, these practices form a foundation for trustworthy AI adoption in education.

5.3 Ethical Balancing

A major ethical challenge in educational AI is keeping fairness, interpretability, and accuracy in balance while staying within regulatory guidelines. These goals often pull in different directions. A highly accurate model might amplify bias, while enforcing fairness too strictly can lower predictive performance. The study showed that careful tuning through fairness audits, explainability tools, and bias mitigation helps keep these priorities aligned. Fairness means distributing opportunities and interventions equitably. It requires continuous monitoring and methods like reweighing or threshold adjustments to correct imbalances. Still, fairness must always be interpreted within context, what's fair in one educational setting may not apply in another. Interpretability ensures that the model's decisions make sense to the people using them. Tools like SHAP help educators understand how certain factors shape predictions, which keeps AI systems open to scrutiny and improvement. Accuracy remains vital but shouldn't outweigh ethics. A model that performs well statistically can still cause harm if it reinforces inequality or hides its logic. The best systems balance predictive power with clarity and fairness. Regulatory frameworks like FERPA and the AI Bill of Rights play a guiding role here. They remind institutions that AI should assist human judgment, not replace it. The most responsible educational AI systems are those that pair computational insight with human care and oversight. With this balance, AI can help education evolve while protecting the dignity and rights of every student.

5.4 Limitations

This research lays out a strong foundation for how AI can be developed and governed responsibly in education, but it's not without its limits. The first and most obvious one is that all the analysis was done using synthetic data. The data was designed to mirror real student behavior and outcomes, but simulated data can never fully capture the messy complexity of real classrooms, diverse learning environments, or the social contexts that shape student experiences. Because of that, the results should be seen as indicative rather than definitive. Real-world validation using genuine academic records, under proper consent and ethical review, will be essential before any of these findings can be confidently applied to live systems. Another limitation is the narrow focus of the fairness and privacy assessments. The study mainly examined gender, age, internet access, and academic risk. While these are important, they leave out other crucial dimensions like race, disability status, or socioeconomic background. In real settings, these factors often overlap in ways that deepen inequities.

The fairness metrics used here, such as selection rate and recall disparity, give useful insights but only scratch the surface of what algorithmic fairness really demands. They don't fully capture the structural biases that can appear once systems interact with real institutional data and policies. The privacy and governance simulations were also conceptual, meant to show what ethical compliance could look like rather than serve as enforceable frameworks. The use of Gaussian noise to simulate differential privacy is a simplified approach, it helps illustrate the idea but doesn't reach the level of formal mathematical guarantees used in official privacy standards. Similarly, the human-in-the-loop and policy oversight setups demonstrate good governance practices, but they would need institutional buy-in, legal consultation, and policy integration to work in practice. Even so, these experiments set the stage for future collaborations between educators, policymakers, and technologists aiming to turn ethical AI principles into operational reality.

6. Future Work

The next step for this work is to move beyond simulations and test the framework on real institutional data within strict privacy and data-sharing agreements. Access to anonymized or consent-based academic records would allow a proper evaluation of how the model performs in genuine educational settings. This will help identify new sources of bias and test how well the system adapts to different institutional contexts. Expanding the fairness scope is another key priority. Future analyses should include race, ethnicity, and socioeconomic indicators to reflect the broader social inequities that shape educational outcomes. These attributes are essential to understanding how algorithmic decisions might reproduce or mitigate existing disparities. A richer, multi-dimensional fairness audit would offer a clearer view of how AI affects diverse student groups. Privacy and data security can also be strengthened through more advanced methods. Federated learning could enable collaboration between institutions without ever sharing raw student data.

Certified differential privacy approaches, which use formal mathematical proofs to guarantee privacy, would go beyond the experimental noise injection used here. These developments could make educational AI both safer and more widely adoptable. Fairness auditing itself can become more sophisticated by adding metrics like equalized odds, predictive parity, and calibration across subgroups. These measures would provide deeper insights into bias patterns that simpler comparisons might overlook. The goal is to create an ethical evaluation process that's as rigorous as the technical one. Finally, there's a real opportunity to turn this framework into practical tools. Institutions would benefit from built-in auditing systems that track fairness, transparency, and privacy in real time. Dashboards that alert educators and administrators to emerging issues could make ethical oversight part of everyday operations rather than an afterthought.

Conclusion

This study shows how artificial intelligence can be used in a responsible, practical way to predict student dropout risk while keeping accuracy, fairness, interpretability, and privacy in balance. Using a simulated dataset built around academic, behavioral, and demographic factors made it possible to reflect real institutional settings without exposing personal information. The machine learning pipeline included detailed preprocessing, feature engineering, and data balancing with SMOTE, helping the models uncover subtle relationships between engagement, performance, and dropout risk. The results make it clear that predicting dropouts is not just a technical problem, it's also an ethical one. Each model brought something useful: Logistic Regression offered a clear and interpretable view of the data, while Random Forest and XGBoost performed better in precision and recall. SHAP explainability confirmed that the most influential predictors were tied to behavior, attendance rate, average grades, and engagement levels, showing that dropout risk often reflects patterns of academic participation more than fixed demographic traits.

Fairness testing revealed gaps between demographic and academic subgroups, reinforcing the importance of equity-aware modeling. By applying bias mitigation techniques such as reweighting and threshold adjustments, the model achieved nearly equal selection rates across student groups. The experiment with privacy-preserving noise also showed that it's possible to protect sensitive data without losing meaningful accuracy. The human-in-the-loop setup highlighted the value of human oversight. Having educators review high-risk predictions ensured that algorithmic insights were paired with contextual understanding. This approach aligns with the principles in the AI Bill of Rights and similar frameworks that call for transparency and accountability in automated systems. It points toward a vision of AI that works with educators rather than replacing them. Overall, this research presents a reproducible, fair, and privacy-conscious approach to early dropout prediction. It bridges the gap between data science and real-world educational practice, offering a model for how institutions can use predictive analytics as transparent and ethical tools that enhance rather than obscure human decision-making. Looking ahead, future work should apply this framework to real institutional datasets under strong privacy safeguards and include broader factors such as socioeconomic context and school environment. The long-term goal is to build AI systems that are not only accurate and explainable but also fair, secure, and genuinely centered on human judgment and student well-being.

References

[1] Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), Learning Analytics: From Research to Practice (pp. 61–75). Springer.

[2] Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do we know who will drop out? A review of the predictors of dropping out of high school: From 1980 to 2012. Journal of Education for Students Placed at Risk, 18(1), 1–37.

[3] Cios, K. J., & Kurgan, L. A. (2021). Towards trustworthy artificial intelligence in education: A review of regulations and best practices. AI and Ethics, 1(1), 67–84.

[4] Das, B. C., et al. (2025). AI-Driven Cybersecurity Threat Detection: Building Resilient Defense Systems Using Predictive Analytics. arXiv preprint arXiv:2508.01422.

[5] Debnath, S., et al. (2025). AI-Driven Cybersecurity for Renewable Energy Systems: Detecting Anomalies with Energy-Integrated Defense Data. International Journal of Applied Mathematics, 38(5s).

[6] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

[7] Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

[8] Hasan, M. S., et al. (2025). Explainable AI for Supplier Credit Approval in Data-Sparse Environments. International Journal of Applied Mathematics, 38(5s).

[9] Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–16.

[10] Islam, M. Z., et al. (2025). Cryptocurrency Price Forecasting Using Machine Learning: Building Intelligent Financial Prediction Models. arXiv preprint arXiv:2508.01419.

[11] Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. EDUCAUSE Review, 46(5), 31–40.

[12] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 1273–1282.

[13] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 1–21.

[14] Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Educational Technology & Society, 17(4), 49–64.

[15] Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. Research and Practice in Technology Enhanced Learning, 12(1), 1–13.

[16] Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. Ethics and Information Technology, 20(1), 5–14.

[17] Ray, R. K. (2025). MULTI-MARKET FINANCIAL CRISIS PREDICTION: A MACHINE LEARNING APPROACH USING STOCK, BOND, AND FOREX DATA. International Journal of Applied Mathematics, 38(8s), 706–738.

[18] Reza, S. A., et al. (2025). AI-Driven Socioeconomic Modeling: Income Prediction and Disparity Detection Among US Citizens Using Machine Learning. Advances in Consumer Research, 2(4).

[19] Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., ... & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 263–268.

[20] Shawon, R. E. R., et al. (2025). Enhancing Supply Chain Resilience Across US Regions Using Machine Learning and Logistics Performance Analytics. International Journal of Applied Mathematics, 38(4s).

[21] Shovon, M. S. S. (2025). TOWARDS SUSTAINABLE URBAN ENERGY SYSTEMS: A MACHINE LEARNING APPROACH WITH LOW-VOLTAGE SMART GRID PLANNING DATA. International Journal of Applied Mathematics, 38(8s), 1115–1155.

[22] Sizan, M. M. H., et al. (2025). Machine Learning-Based Unsupervised Ensemble Approach for Detecting New Money Laundering Typologies in Transaction Graphs. International Journal of Applied Mathematics, 38(2s).

[23] The White House Office of Science and Technology Policy. (2022). Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People. Washington, DC.

[24] U.S. Department of Education. (2011). Family Educational Rights and Privacy Act (FERPA), 34 CFR Part 99. Federal Register.