computers

MDPI

Article

Enhancing Personalised Learning with Graph-Based Ensemble Prediction and Skill Cluster Mapping for Student Knowledge Completeness

Zhanibek Kozhirbayev [1,2,3,*] and Assel Omarbekova [1]

[1] Faculty of Information Technologies, L.N. Gumilyov Eurasian National University, Astana 010008, Kazakhstan; omarbekova_as@enu.kz
[2] National Laboratory Astana, Nazarbayev University, Astana 010000, Kazakhstan
[3] Institute of New Materials and Energy Technologies, Nazarbayev University, Astana 010000, Kazakhstan
* Correspondence: zhanibek.kozhirbayev@nu.edu.kz; Tel.: +7-(701)-344-24-70

Abstract

The increasing adoption of data-driven educational systems requires reliable methods to predict student readiness for future coursework and support personalised learning pathways. This study proposes a graph-enhanced ensemble framework that integrates curriculum structure and skill-gap awareness to estimate student course readiness. A global prerequisite directed acyclic graph (DAG) of university subjects was constructed to model curriculum dependencies, from which structural features including the PageRank, in-degree, out-degree, and prerequisite chain depth were derived. In parallel, a domain-informed skill cluster mapping grouped subjects into nine interpretable competency domains to enable skill-gap analysis. These curriculum-aware features were combined with academic history, behavioural engagement, and demographic indicators to produce 38 engineered features for each student–subject pair. Six models (CatBoost, XGBoost, LightGBM, FT-Transformer, MLP and TabPFN) were trained and combined using a weighted ensemble. Experiments on a real-world institutional dataset containing 20,581 students and 727,168 records achieved an AUC of 0.8908 for predicting course success. Ablation experiments demonstrate that graph-derived and skill-cluster features provide modest but statistically significant incremental value. The resulting model was integrated into a prototype personalised recommender that prioritizes curriculum-consistent learning pathways. The proposed framework provides an interpretable and curriculum-aware approach for personalised learning. While the model demonstrates strong overall performance, a moderate gender disparity in the false positive rate was observed. Results were obtained on a large longitudinal dataset from a single university, and external validation at other institutions is needed to confirm generalizability.

Keywords: curriculum graph modelling; educational data mining; ensemble learning; learning analytics; student performance prediction; personalised learning recommendation

Check for updates

Academic Editor: Stamatios Papadakis

Received: 4 May 2026
Revised: 24 May 2026
Accepted: 27 May 2026
Published: 28 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The rapid development of smart learning environments has intensified the need for systems capable of understanding and supporting individual student learning needs in a personalised and proactive manner. Higher education institutions are increasingly moving beyond traditional one-size-fits-all teaching models toward adaptive educational systems that account for differences in student preparedness, learning pace, and academic

background. A key challenge in this transformation is the ability to estimate student knowledge completeness, which in this study refers to a student's readiness to successfully engage with a target subject based on prior academic preparation and prerequisite mastery. Accurate estimation of such readiness enables early identification of at-risk students, timely pedagogical interventions, and the creation of coherent personalised learning pathways.

Despite notable progress in educational data mining, most existing studies continue to rely heavily on historical grades, basic demographics, or simple behavioural indicators. These approaches often overlook the underlying structure of the curriculum and the relationships between courses. Prerequisite dependencies naturally form a directed acyclic graph, yet few models explicitly incorporate this graph topology or higher-level skill domains when making predictions. As a result, current systems frequently lack interpretability and fail to deliver truly personalised recommendations that align with both student readiness and curriculum coherence.

To address these limitations, this study proposes a graph-enhanced ensemble framework for personalised learning recommendation. The approach integrates curriculum structure, domain knowledge, and modern tabular machine learning models. First, a global prerequisite directed acyclic graph (DAG) is constructed using NetworkX [1]. From this graph, structural indicators such as the PageRank, in-degree, out-degree, and prerequisite chain depth are derived to capture curriculum topology. Second, a domain-expert validated mapping groups subjects into nine interpretable skill clusters, enabling competency-level analysis of student preparedness. Third, a heterogeneous ensemble of six models (CatBoost, XGBoost, LightGBM, FT-Transformer [2], MLP, and TabPFN [3]) is trained on engineered features describing academic history, prerequisite readiness, behavioural engagement, demographic context, and curriculum structure.

The framework is evaluated on a large real-world dataset from L. N. Gumilyov Eurasian National University, comprising 20,581 students and 727,168 student–subject records. In addition to predictive performance, the system produces calibrated estimates of student knowledge completeness and supports a prototype personalised recommender that generates curriculum-aware learning pathways using a hierarchical scoring mechanism (same department > same faculty > same skill cluster).

The main contributions of this work are as follows:

- Curriculum-aware graph feature engineering: Construction of a university-wide prerequisite DAG and extraction of structural indicators (PageRank, in-degree, out-degree, and chain depth) to model course dependency structure.
- Domain-informed skill cluster mapping: Introduction of a mapping of academic subjects into nine interpretable competency domains, enabling skill-level analysis of student preparedness.
- Heterogeneous ensemble modelling: Development of an ensemble combining six complementary tabular learning architectures to improve robustness when modelling heterogeneous educational data.
- Large-scale institutional evaluation: Empirical validation on 20,581 students and 727,168 records, demonstrating improved predictive performance and the value of curriculum-aware features.
- Deployment-oriented recommendation framework: Integration of predicted knowledge completeness into a personalised recommender that respects prerequisite constraints and institutional curriculum structure.

The rest of the paper is organised as follows. Section 2 reviews related work in educational data mining and personalised learning systems. Section 3 describes the dataset. Section 4 presents the proposed methodology, including data preprocessing steps in detail, feature engineering, and the base and ensemble framework. Section 5 outlines the experi-

mental setup, which details the data splitting strategy, model training hyperparameters, and evaluation metrics. Section 6 reports the results, including the performance metrics, ablation analysis, interpretability, and fairness evaluation. Finally, Section 7 discusses the findings, limitations, and practical implications, while Section 8 concludes the paper and suggests directions for future research.

2. Related Work

Educational data mining (EDM) has evolved significantly as a key enabler for predicting student academic performance and supporting personalised learning in higher education. Traditional approaches primarily rely on historical grades, demographics, and basic behavioural indicators, employing supervised machine learning techniques such as decision trees, Random Forests, and neural networks [4–7]. While these features provide useful signals, they often treat courses as independent entities and overlook the structured relationships embedded within academic curricula. In practice, university curricula are inherently organised through prerequisite dependencies that form directed acyclic graphs representing the progressive accumulation of knowledge. Ignoring these dependencies limits the ability of predictive models to determine whether a student possesses the foundational knowledge required for future subjects, thereby reducing both predictive accuracy and interpretability.

Recent advances have attempted to address this limitation through graph-based representations and educational knowledge graphs (EKGs), which encode relationships among courses, concepts, and learning outcomes [8,9]. Such approaches enable richer modelling of curriculum dependencies and have demonstrated improvements in both prediction and recommendation tasks. However, many of these methods rely on automatically constructed graphs or graph neural networks that require large-scale relational data, introduce additional architectural complexity, and may reduce the interpretability in real-world institutional deployments. Moreover, several studies evaluate their models on relatively small or synthetic datasets, limiting evidence of their scalability and applicability in operational university environments [10–12].

Complementary to graph-based modelling, skill-gap analysis has emerged as an important paradigm for understanding discrepancies between acquired and required competencies. By mapping courses or academic units to interpretable skill domains, these approaches enable higher-level reasoning about student readiness and support actionable educational recommendations [13,14]. Nevertheless, many prior studies employ unsupervised or purely data-driven clustering methods that may lack semantic interpretability or alignment with institutional academic structures. Domain-informed mappings grounded in faculty expertise remain comparatively underexplored despite their potential to enhance both interpretability and practical relevance [15,16].

In parallel, ensemble learning and modern machine learning models have become dominant approaches in student performance prediction due to their ability to handle heterogeneous data, capture complex feature interactions, and maintain robustness in real-world datasets [17,18]. Gradient boosting algorithms such as CatBoost, XGBoost, and LightGBM consistently demonstrate strong empirical performance, while recent deep tabular architectures (e.g., FT-Transformer) further improve representation learning capabilities [19–22]. However, these models are typically applied to flat tabular representations and rarely incorporate explicit curriculum topology or semantically meaningful skill abstractions within a unified predictive framework.

Personalised recommender systems for higher education have similarly evolved from simple grade-based recommendations toward more sophisticated approaches incorporating semantic relationships and learning trajectories. Knowledge graph-based recommender

systems exploit prerequisite structures and semantic similarity to generate improved course recommendations, while hybrid approaches integrate embeddings, sequence modelling, and collaborative filtering to construct coherent learning pathways [23–27]. More recent work incorporates reinforcement learning or ensemble attention mechanisms to generate adaptive multi-step recommendations and mitigate information overload [28,29]. Empirical evidence further suggests that graph-enhanced ensemble approaches can significantly improve predictive accuracy, recall for at-risk students, and probability calibration compared with traditional baselines [30,31].

Taken together, these studies highlight an important gap in the literature. While prior research has explored curriculum graphs, skill-based modelling, and machine learning approaches for student performance prediction, these components are typically investigated in isolation. Few studies integrate curriculum topology, interpretable skill abstractions, and heterogeneous tabular learning within a unified and scalable framework for personalised learning analytics. To address this gap, the present study develops a curriculum-aware predictive framework that combines prerequisite graph information, domain-informed skill representations, and ensemble machine learning models to estimate student knowledge completeness and support personalised learning recommendations.

3. Dataset

The dataset used in this study was collected from the institutional information systems of ENU, covering the academic years 2011–2025. It represents a substantial but not exhaustive subset of the university's student population. The data were provided by the university's Department of Digital Development. It constitutes a large-scale, real-world, longitudinal collection of student academic trajectories, behavioural indicators, demographic profiles, and curriculum structures.

No systematic inclusion or exclusion criteria were applied to students or courses beyond basic data integrity checks necessary for feature engineering. Unlike many publicly available educational data mining benchmarks that are often synthetic, MOOC-based, or limited in scope, this dataset reflects the full complexity of an operational university environment, including multi-level degree programs, multilingual instruction, and diverse student backgrounds. Full details of the raw data sources are provided in Appendix A.

The richness of the data, integrating grades, attendance, extracurricular activity, and explicit prerequisite relationships, enables the development of predictive models that go beyond simple grade forecasting to assess knowledge completeness while respecting curriculum dependencies and individual student contexts. This directly supports the project goal of creating digital services for identifying individual student learning needs. The key dataset statistics are summarized in Tables 1 and 2. The student distributions by degree type, study language, year of study, faculty affiliation, and gender are shown in Appendix A.

Table 1. Dataset overview.

Metric	Value
Total Students	20,581
Total Samples (Student–Subject Pairs)	727,168
Avg Samples per Student	35.3

All data were pseudonymised prior to analysis. No personally identifiable information was used. The study complies with institutional ethics guidelines.

Table 2. Student demographics.

Category	Group	Count	%
Gender	Female	11,833	57.5%
	Male	8748	42.5%
Special Needs	Yes	2628	12.8%
	No	17,953	87.2%

4. Methodology

This section describes the methodology used to predict student knowledge completeness, defined as the probability that a student will successfully complete a future course. The methodological pipeline consists of three main stages: data preprocessing, feature engineering, and predictive modelling.

4.1. Problem Formulation

We formulate the task as a supervised binary classification problem. Given a student s and a target subject t that the student has not yet taken, the goal is to predict the probability that the student will successfully complete the subject. Formally, for each student–subject pair$(s,\ t)$, we construct a feature vector $x_{s,t}$ and estimate:

$$\hat{p}_{s,t} = P\left(y_{s,t} = 1 \middle| x_{s,t}\right) \quad (1)$$

where $y_{s,t}$ is a binary outcome variable defined as

$$y_{s,t} = \begin{cases} 1, \ if \ \text{Mark_Total}_{s,t} \geq 75 \\ 0, \ otherwise \end{cases}$$

The feature vector $x_{s,t}$ represents the information available prior to enrollment in subject t, including the student's academic history, prerequisite readiness, behavioural engagement indicators, demographic context, and curriculum structure. The 75-point threshold is grounded in the institutional grading policy of ENU, where it represents satisfactory academic performance and indicates sufficient mastery for progression to subsequent coursework. While the minimal passing grade is 50, scores in the 50–74 range are considered marginal passes that do not reliably demonstrate the level of knowledge completeness required for downstream subjects in the curriculum. Therefore, we adopted the higher threshold of 75 to better align with the study's focus on predicting genuine readiness rather than minimal passing.

4.2. Data Preprocessing

Data preprocessing was performed using a custom pipeline. Key steps included lookup table creation, behavioural aggregation, construction of a global prerequisite DAG, skill cluster mapping, student-level merging and lookup table creation.

4.2.1. Auxiliary Data Integration

The supporting datasets were first cleaned and standardized to ensure consistent identifiers across institutional records. Lookup tables were created for subjects, faculties, departments, and degree types. These tables provide normalized metadata that enables efficient merging of heterogeneous university datasets. Key changes in the processed lookup tables are presented in Appendix B.

In addition, behavioural indicators were aggregated at the student level. Records related to attendance, internships, and scientific publications were aligned with individual

student grade histories. These variables capture aspects of academic engagement and extracurricular activity that may influence learning outcomes.

4.2.2. Prerequisite DAG Construction

To capture the structural dependencies of the curriculum, we modelled subject prerequisites as a formal global DAG, $G = (V, E)$. In this representation, nodes V represent individual subjects, and directed edges $E = \{(u, v)\}$ represent administrative requirements where subject u must be completed before subject v. The graph was constructed from ENU institutional records using the NetworkX library (v3.6.1). This topological approach explicitly encodes the cumulative nature of the learning process, where knowledge completeness in ancestral nodes is a latent requirement for success in descendant nodes. As illustrated in Figure 1, the DAG captures non-trivial dependencies across different academic domains. For instance, foundational courses like Sociology (SubjID 8779) and Ecology (SubjID 8221) serve as structural precursors to more specialized subjects such as Environmental Biotechnology (SubjID 614).

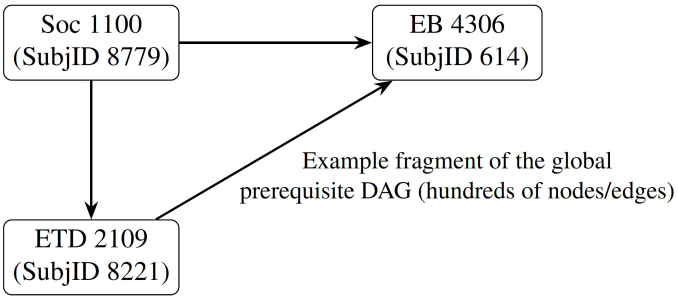


Figure 1. Example fragment of the global prerequisite DAG illustrating topological dependencies. Foundational subjects (**left**) feed into advanced subjects (**right**).

We quantified the structural significance of each subject using three distinct graph metrics:

1. PageRank ($\alpha = 0.85$): PageRank was used to identify central "anchor" subjects. A high PageRank indicates a subject that is not only a frequent prerequisite but is linked to other high-impact courses.
2. Out-degree (d_{out}): The out-degree identifies "bottleneck" foundational courses. Subjects with a high out-degree are critical path nodes required for a high volume of downstream specializations.
3. In-degree (d_{in}): The in-degree measures the "dependency load." High in-degree nodes represent advanced coursework that requires the integration of multiple prior competency streams.

Additionally, we extracted the prerequisite chain depth, defined as the longest path from a root node to the target subject, to serve as a proxy for the academic maturity required for specific enrolments.

4.2.3. Skill Cluster Mapping

To enable interpretable skill-gap analysis, subjects were grouped into nine domain-informed skill clusters based on faculty affiliation and academic taxonomies. This mapping was developed in collaboration with domain experts from the respective faculties.

We considered automatic clustering approaches, such as K-Means applied to subject feature representations (e.g., using embeddings of subject names and metadata). However, these data-driven methods produced clusters with limited semantic coherence and poor alignment with the university's official academic programs and pedagogical domains.

K-Means clusters often grouped unrelated disciplines together or created unbalanced partitions that lacked practical interpretability for educators and students.

Therefore, we intentionally adopted an expert-validated mapping that assigns each faculty to a semantically meaningful competency cluster. This expert-driven approach ensures direct alignment with the institutional curriculum organisation, improves interpretability for stakeholders, and supports more actionable skill-gap analysis in real educational settings. The complete faculty-to-cluster mapping, which includes the default "Other" category for non-specialized electives, is detailed in Appendix C.

4.2.4. Student-Level Merging

The final integrated student data is stored in students_dag.json as a nested structure containing the demographics, academic profile, detailed grade history with prerequisite chains, internships, and publications. Example of the nested structure for one student is presented in Table 3.

Table 3. Example of the nested structure for one student.

Component	Content
Demographics	sex: "male"; birth_year: 1994; citizenship: "Kazakhstan"; special_needs: false
Academic	faculty_id: 12; degree_id: 1; study_language: "Kazakh"; year_of_study: 2011; study_program: "Economics of Entrepreneurship and Business"
History	Subject 4: study_year = 2011, mark_total = 90.0, attendance_percent = 75.0, prereq_chain={direct_prereqs: [], chain_depth: 0}; Subject 719: study_year = 2011, mark_total = 95.0, prereq_chain = {direct_prereqs: [28112], chain_depth: 1}; . . .
Internships	[]
Publications	[]

4.3. Feature Engineering

To transform the nested longitudinal student records into a format suitable for supervised learning, we constructed a tabular dataset in which each row corresponds to a student–subject pair. From the academic trajectories, we engineered a total of 38 predictive features, organised into five conceptual categories: academic history, prerequisite readiness, graph-derived structural features, behavioural engagement, and demographic context (Table 4).

Table 4. Representative subset of the 38 engineered features.

Feature	Value	Description/How Engineered
prereq_readiness	1.00	Fraction of direct prerequisites passed with mark ≥ 50
prev_gpa	82.3	Mean of all previous marks earned by the student
subj_difficulty	78.4	Bayesian smoothed average mark for the target subject (global mean + subject-specific history)
difficulty_adjusted_gpa	1.05	prev_gpa normalized by subject difficulty
knowledge_coverage	0.75	Fraction of previous subjects passed with mark ≥ 50

Table 4. *Cont.*

Feature	Value	Description/How Engineered
knowledge_coverage_same_cluster	0.92	Coverage rate within the same skill cluster (faculty-based)
fail_rate	0.08	Proportion of previous subjects with mark < 50
gpa_std	6.5	Standard deviation of past grades (measures consistency)
consistency	0.133	Inverse of (1 + gpa_std); higher = more consistent performance
momentum	1.2	Recent GPA trend adjusted by fail rate
temporal_weighted_gpa	84.1	Time-decayed weighted average of past grades
subj_pagerank	0.0456	PageRank score of the target subject in the global prerequisite DAG
engagement_score	0.78	0.5 × current attendance + 0.25 × log(1 + internships) + 0.25 × log(1 + publications)
current_attendance	88.5	Attendance percentage for the current subject
age	22	Calculated as 2026—birth year
cross_faculty_transfer	0	Binary flag: 1 if target subject belongs to a different faculty
target_bin	1	Binary target: 1 if Mark_Total $\geq$ 75, else 0

4.3.1. Time-Decayed and Momentum Features

To capture the evolving nature of student performance, we implemented dynamic temporal features instead of static averages. The *temp_weighted_gpa* applies an exponential decay function to historical marks, giving higher weight to more recent performance and thereby better reflecting current knowledge readiness:

$$\text{temporal_weighted_gpa} = \frac{\sum_{i=1}^{n} mark_i \cdot \alpha^{(t_{current} - t_i)}}{\sum_{i=1}^{n} \alpha^{(t_{current} - t_i)}} \quad (2)$$

where α ($0 < \alpha < 1$) is the decay factor, $t_{current}$ is the prediction time, and t_i is the time of the i-th historical record.

Additionally, the *momentum* feature quantifies the short-term academic trajectory by computing the GPA trend over the two most recent semesters, adjusted by the student's historical failure rate:

$$momentum = \frac{GPA_{t-1} - GPA_{t-2}}{1 + fail_rate} \quad (3)$$

These dynamic features allow the model to better differentiate between students showing academic recovery and those whose performance is declining.

4.3.2. Curriculum-Aware Features

Using the global prerequisite DAG and the domain-informed skill cluster mapping, we derived high-level structural features. *prereq_readiness* measures the fraction of direct prerequisites completed with a passing mark ($\geq$50). Features like *knowledge_coverage* and *knowledge_coverage_same_cluster* quantify the student's overall and cluster-specific foundational preparation, respectively. The target subject's *subj_pagerank* and in-degree from the prerequisite DAG are included to enable the model to account for the subject's hierarchical importance and dependency load within the curriculum.

Additionally, we computed a subject difficulty score (*subj_difficulty*) using Bayesian smoothing of historical average grades for each subject. This feature, together with *dif-*

ficulty_adjusted_gpa, helps the model account for varying course rigor when assessing student readiness.

4.3.3. Temporal Integrity and Leakage Mitigation

To ensure the model can be deployed prospectively (i.e., predictions are made before course enrolment), we implemented a strict temporal separation between the historical data and the prediction point. A common risk in longitudinal educational data is look-ahead bias, where information observed during or after a course influences the prediction of that same course's outcome. This is addressed through a historical institutional imputation strategy. For each student–subject pair $(s,\ t)$ at the prediction time point (t_1 in Figure 2):

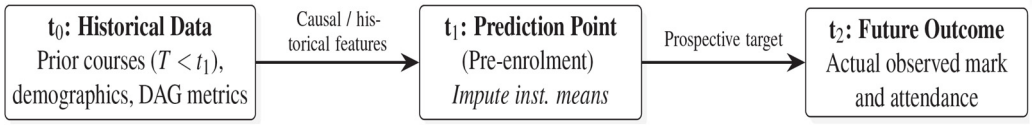


Figure 2. Timeline illustrating temporal data separation. At deployment (t_1), only information available before course commencement is used. Unobservable features are replaced with historical institutional averages to maintain causal integrity.

1. All historical features (GPA statistics, prerequisite information, DAG metrics, and prior marks) are computed exclusively from records completed before t_1.
2. Unobservable target-specific features (*current_attendance* and *engagement_score*) are replaced with neutral defaults corresponding to the historical institutional means (80.0 and 0.75, respectively). The impact of this mitigation strategy was quantified by training a leakage-safe variant of the model (see Section Leakage-Safe Variant (Prospective Deployment)), which excludes these two features entirely.

To quantify the impact of this mitigation strategy, we trained a leakage-safe variant of the model that completely excluded the two unobservable features.

4.4. Base Models

To establish strong baselines and capitalize on the complementary strengths of different model families, we trained eight models in total. This includes two classical baselines and six advanced models spanning gradient boosting, deep tabular architectures, and prototype-based methods.

Classical baselines (Logistic Regression and Random Forest) were included to provide a reference against traditional machine learning approaches commonly used in educational data mining. They were used only for benchmarking and were not included in the final ensemble.

The six advanced models were deliberately selected to form a heterogeneous ensemble:

- Gradient boosting models: CatBoost, XGBoost, and LightGBM were selected. These were chosen for their excellent handling of heterogeneous tabular data, built-in support for categorical features, robustness to outliers, and strong empirical performance on educational datasets.
- Deep tabular models: FT-Transformer and MLP were selected. These models were selected because they can automatically learn complex non-linear interactions and higher-order feature combinations that traditional tree-based models may miss.
- Prototype-based model: TabPFN was selected. This was included as a strong, fast baseline that leverages in-context learning and has demonstrated competitive performance on small-to-medium tabular tasks with minimal hyperparameter tuning.

Each model was trained on the complete set of 38 engineered features. CatBoost was trained directly on the original feature set, taking advantage of its native categorical

support. For the remaining models (XGBoost, LightGBM, Logistic Regression, Random Forest, and TabPFN), categorical variables were encoded using Ordinal Encoding. For FT-Transformer and MLP, categorical features were handled internally through embedding layers within the PyTorch Tabular framework (v1.1.0).

4.5. Heterogeneous Ensemble Framework

The core predictive engine is a heterogeneous ensemble of six diverse architectures. The final prediction is obtained through a weighted ensemble:

$$\hat{p}_{ensemble} = \sum_{i=1}^{6} \mathrm{w_i} \cdot \mathrm{p_i(x)} \tag{4}$$

where $\mathrm{p_i(x)}$ represents the probability output of the i-th model. The weights $\mathrm{w_i}$ were determined by constrained convex optimization on the validation set that maximizes the ensemble AUC under simplex constraints ($\sum \mathrm{w_i} = 1,\ w_i \geq 0$). Specifically, we used the SLSQP solver to solve the following optimization problem:

$$\underset{w}{arg\max}\ AUC\left(\sum_{i=1}^{6} \mathrm{w_i} \cdot \mathrm{p_i(x)}, \mathrm{y_{val}}\right)\ s.t.\ \sum_{i=1}^{6} \mathrm{w_i} = 1,\ w_i \geq 0 \tag{5}$$

This weighting scheme ensures that models with higher intrinsic capabilities for handling high-cardinality tabular data and class imbalance contribute more significantly to the final decision. The specific weight assignments resulting from this optimization are detailed in Section 6.

This ensemble design was motivated by several important considerations:

- Diversity: Gradient boosting models excel at capturing complex feature interactions and handling missing values, while deep tabular models are capable of discovering higher-order and more abstract representations. TabPFN further contributes a fundamentally different inductive bias based on prior-data fitting.
- Robustness: Combining multiple architectures reduces the risk of overfitting to any single modelling assumption and improves generalization on heterogeneous educational data.
- Calibration and reliability: The weighted combination, with weights optimised specifically for the AUC on the validation set, tends to produce better-calibrated probabilities than any individual model.
- Practical performance: Ablation experiments and the final results demonstrate that the ensemble consistently outperforms the best single model, particularly in terms of the AUC and Brier score.

The resulting ensemble serves as the core predictive engine for both knowledge completeness prediction and the downstream personalised recommender system.

5. Experimental Setup

This section presents the experimental setup used for the validation of the proposed predictive framework. First, the student-level data splitting strategy designed to prevent temporal leakage is described. Next, the model training configuration and hyperparameter settings are outlined. The personalised recommender system, which leverages ensemble predictions to generate curriculum-aware learning pathways, is then introduced. Finally, the evaluation metrics, statistical testing procedures, and ablation study used to quantify the impact of the proposed graph-based and skill-cluster features are defined.

5.1. Data Splitting Strategy

To prevent data leakage from temporal dependencies within student histories while avoiding student-specific information leakage across splits, we applied a strict student-level split (70% train, 15% validation, and 15% test) using *train_test_split* with a fixed random seed (42). This ensures that all records belonging to a test student appear only in the test set. Within each student's records, features for a given student–subject pair are computed strictly from data available before the enrolment date of the target subject (see Section 4.3.3 for details on temporal integrity and leakage mitigation). Thus, while the split is performed at the student level, temporal causality is strictly preserved at the feature engineering stage (Table 5).

Table 5. Dataset split statistics.

Split	Students	Students	Pos. (%)	Pos. Count	Neg. Count
Train	14,406	508,807	87.95	447,480	61,327
Dev	3087	109,168	87.80	95,824	13,344
Test	3088	109,193	88.00	96,097	13,096
Total	20,581	727,168	87.95	639,401	87,767

The final dataset is large-scale (more than 700,000 records), high-dimensional (38 features), moderately imbalanced (≈88%), heterogeneous, and strongly temporal. The high positive rate reflects institutional grading distributions where the majority of students achieve passing or satisfactory marks. This imbalance is typical for real-world higher education datasets and motivates the use of metrics such as PR-AUC, negative-class recall, and MCC. These properties closely mirror real-world educational systems and make the dataset particularly suitable for developing robust, curriculum-aware predictive models.

5.2. Special Cases

Freshman students (no prior course history) and subjects with no prerequisites receive special handling to ensure fair and educationally meaningful predictions. In feature engineering, freshmen are flagged with *is_freshman* = 1 and receive a neutral prior GPA of 75.0. For subjects with no prerequisites, *prereq_readiness* = 1.0 and *avg_prereq_importance* = 0.0. In the personalised recommender, such cases receive an explicit *freshman_bonus* (value = 2.0) in the hierarchical pathway score and are prioritised when the target belongs to the student's dominant skill cluster or department.

5.3. Personalised Recommender System

The validated ensemble serves as the predictive engine for a personalised recommender system designed to identify optimal academic trajectories for students. Rather than recommending subjects based solely on difficulty, the system employs a hierarchical ranking mechanism that prioritizes curriculum coherence and institutional alignment. Potential next-step subjects are ranked according to a multi-level priority logic derived from the institutional structure:

- Institutional Proximity: Candidate subjects are first filtered by academic proximity, with strict priority given to those within the same Cafedra (department), followed by the same faculty, and finally the same domain-specific skill cluster.
- Knowledge Readiness: Within each proximity tier, subjects are ranked by their predicted probability of success ($\hat{p}_{s,t}$), ensuring that recommendations align with the student's current knowledge completeness.

- Curriculum Significance: For subjects with similar readiness scores, the global PageRank metric is used as a tie-breaker to prioritize foundational subjects that serve as critical nodes in the prerequisite DAG.

The ranking logic and evaluation protocol are described in detail in Section 6.7.

Integration and Student Interaction

The recommender is designed as a module that will be integrated into the existing Learning Management System (LMS) of ENU. When a student selects a subject for the upcoming semester during the course registration process in the LMS, the system automatically triggers in real time. It analyses the student's academic history and computes knowledge completeness for the selected subject using the ensemble model.

The interface then displays:

- The predicted success probability for the chosen subject along with a risk level (LOW/MEDIUM/HIGH).
- A ranked list of recommended alternative or preparatory subjects (top five), prioritized by the hierarchical pathway score (same Cafedra → same faculty → same skill cluster).
- Brief justifications for each recommendation (e.g., "High readiness based on strong prerequisite coverage in the same skill cluster").

Students can view prerequisite gaps (if any), accept the suggested subjects, or manually adjust their selection. Academic advisors have access to the same recommendations for review and approval. This seamless integration allows students to receive immediate, curriculum-aware guidance directly within the familiar LMS environment during registration.

5.4. Model Training and Hyperparameters

All models were trained on a DGX-1 server with GPU acceleration where applicable. Due to the substantial size of the dataset (over 700,000 records), we avoided exhaustive grid search. Instead, we performed limited hyperparameter optimization through manual tuning and sensitivity analysis focused on key parameters (the learning rate, tree depth, number of estimators, and regularization terms) using the validation set. Early stopping was applied based on the validation AUC (patience = 100 for GBDT models) or validation loss (for deep models).

The final configurations used in all experiments are as follows:

- Logistic Regression: Default parameters with *max_iter* = 1000 and *solver* = '*lbfgs*'.
- Random Forest: 500 trees, *class_weight* = '*balanced*', *max_depth* = *None*, *min_samples_split* = 2.
- CatBoost: 5000 iterations, *learning_rate* = 0.05, *depth* = 8, *l2_leaf_reg* = 3, *task_type* = "*GPU*", *eval_metric* = "*AUC*", random seed 42, early stopping (patience = 100).
- XGBoost: 1500 estimators, *learning_rate* = 0.08, *max_depth* = 7, *subsample* = 0.85, *colsample_bytree* = 0.85, *tree_method* = "*hist*", GPU support, *scale_pos_weight* adjusted for class imbalance.
- LightGBM: 1500 estimators, *learning_rate* = 0.08, *max_depth* = 8, *num_leaves* = 128, *subsample* = 0.85, *colsample_bytree* = 0.85, GPU acceleration, *scale_pos_weight* for imbalance.
- FT-Transformer: Batch size 1024, learning rate 0.0001 (with AdamW optimizer), 25 maximum epochs, early stopping on validation loss, dropout rate 0.1.
- MLP: Hidden layers *[512, 256, 128]*, ReLU activation, batch size 1024, learning rate 0.001, dropout rate 0.2, maximum 50 epochs, early stopping on validation loss.
- TabPFN: Used default parameters and trained on the most recent 40,000 training samples (due to its quadratic memory scaling).

To assess the robustness of these hyperparameter choices, we conducted a limited sensitivity analysis on the validation set. For each model, we varied the key hyperparameters around the final values: the learning rate by ±50%, maximum tree depth by ±2 levels, and number of estimators by ±30%. Across all models, the validation AUC varied by no more than ±0.003 to ±0.007. This small variation demonstrates that the reported performance is stable and not overly sensitive to precise hyperparameter tuning, thereby supporting the reliability of our modelling pipeline.

Freshman students and subjects with no prerequisites received special handling, as described in Section 5.2. All final models were retrained on the combined train + validation set using the selected configurations before evaluation on the held-out test set.

5.5. Evaluation Metrics and Statistical Analysis

Performance was assessed using a comprehensive set of metrics: Area Under the ROC Curve (AUC) as the primary metric, along with the accuracy, Balanced Accuracy, F1-score, Matthews Correlation Coefficient (MCC), LogLoss, and Brier score. Bootstrap resampling (n = 1000) was used to compute 95% confidence intervals for the AUC.

Model differences were evaluated using permutation tests with 10,000 permutations. The significance threshold was set at $\alpha = 0.05$. When multiple comparisons were performed (e.g., pairwise model comparisons and the ablation study), we applied the Bonferroni correction to control the family-wise error rate.

Model interpretability was analysed using SHAP (SHapley Additive exPlanations) values, providing insights into feature importance and interaction effects. Calibration quality was assessed via calibration curves and the Brier score. Additionally, advanced fairness analysis was conducted by examining the predictive parity (positive predictive rate), true positive rate (TPR), and false positive rate (FPR) across sensitive attributes (sex and special_needs).

5.6. Ablation Study Design

To rigorously evaluate the contribution of the proposed graph and skill-gap features, we conducted a systematic ablation study using CatBoost as the base learner. Four configurations were compared:

- Baseline 1: Raw GPA only;
- Baseline 2: GPA + attendance features;
- Baseline 3: All non-graph features (excluding *subj_pagerank*, *knowledge_coverage_same_cluster*, *cross_faculty_transfer*, *avg_prereq_importance*, and *temporal_weighted_gpa*);
- Full model: Complete feature set including graph and skill-gap components.

6. Results

This section presents the empirical evaluation of the proposed curriculum-aware predictive framework. The analysis is organised around five aspects: (1) the predictive performance of individual models, (2) gains from the ensemble model, (3) the contribution of curriculum-aware features, (4) the probability calibration and interpretability, and (5) the fairness analysis together with offline evaluation of the personalised recommender system.

All experiments were conducted on a held-out test set comprising 3,088 students and 109,193 student–subject pairs. The strict student-level split ensures that no student appears across multiple partitions. This evaluation protocol eliminates information leakage and reflects a realistic deployment scenario where predictions are made for unseen students.

6.1. Predictive Performance of Individual Models

We first established performance benchmarks by evaluating eight models trained on the full engineered feature set. This includes two classical baselines (Logistic Regression and Random Forest) and six advanced models: three gradient-boosted decision tree (GBDT) variants (CatBoost, XGBoost, and LightGBM), two deep tabular architectures (FT-Transformer and MLP), and a prior-data Transformer (TabPFN). The primary classification results are summarized in Table 6. Performance is evaluated across discrimination (AUC and MCC), robustness to class imbalance (Balanced Accuracy, F1-score), and probability calibration (LogLoss, Brier score). *p*-values are derived from DeLong's test for AUC comparisons using LightGBM as the reference model. The extended classification metrics, including the class-specific recall, specificity, and negative predictive value (NPV), are provided in Appendix D.

Table 6. Predictive performance of individual base models on the test set.

Model	AUC [95% CI]	Balanced Acc	MCC	F1	LogL	Brier	*p*-Val
LightGBM	0.8893 [0.887–0.892]	0.806	0.447	0.877	0.390	0.131	-
CatBoost	0.8845 [0.882–0.887]	0.640	0.391	0.943	0.247	0.075	0.009
XGBoost	0.8827 [0.880–0.886]	0.793	0.449	0.894	0.359	0.118	<0.001
FT-Transformer	0.8826 [0.880–0.885]	0.649	0.395	0.942	0.249	0.076	<0.001
MLP	0.8798 [0.877–0.883]	0.641	0.390	0.942	0.251	0.077	<0.001
Random Forest	0.8752 [0.872–0.878]	0.660	0.401	0.941	0.264	0.077	<0.001
Logistic Regression	0.8624 [0.859–0.865]	0.637	0.367	0.940	0.266	0.080	<0.001
TabPFN	0.8634 [0.860–0.867]	0.586	0.297	0.940	0.266	0.081	<0.001

The experimental results indicate that GBDT methods consistently outperform both classical baselines and other model families in this domain. LightGBM achieved the highest discriminative power (AUC = 0.8893) and more robust handling of the minority class (Balanced Accuracy = 0.806). However, CatBoost demonstrated the best probability calibration among all models, yielding the lowest Brier score (0.075) and LogLoss (0.247). The two classical baselines (Logistic Regression and Random Forest) were substantially outperformed by the modern models, confirming the advantage of advanced tabular architectures on this large, heterogeneous educational dataset.

Although LightGBM achieved the highest discriminative performance (AUC = 0.8893), we selected CatBoost as the reference model for SHAP interpretability analysis and the ablation study. This decision was motivated by several important considerations. First, CatBoost demonstrated the best probability calibration among all models, with the lowest Brier score (0.075) and LogLoss (0.247). Reliable probability estimates are essential in educational applications where predicted success probabilities are used for risk assessment and personalised recommendations. Second, CatBoost's native handling of categorical features and ordered boosting mechanism produces more stable and consistent SHAP value attributions, particularly when features exhibit strong interdependencies (e.g., various GPA-related features and curriculum graph metrics). This stability enhances the trustworthiness of feature importance explanations for academic stakeholders. In contrast, while LightGBM excels in ranking performance, its SHAP attributions showed higher variance across runs. Therefore, CatBoost provides a more suitable foundation for interpretable insights that can meaningfully inform pedagogical interventions.

6.2. Performance of the Weighted Ensemble

To exploit the complementary strengths of the individual models, we construct a weighted ensemble optimized on the validation AUC under simplex constraints (see

Section 4.5 for the optimization procedure). The final ensemble weights are: CatBoost ($w = 0.25$), TabPFN ($w = 0.20$), XGBoost ($w = 0.15$), LightGBM ($w = 0.15$), FT-Transformer ($w = 0.15$), and MLP ($w = 0.10$). This configuration prioritizes models with strong calibration and robustness while maintaining diversity across model families.

The weighted ensemble achieved an AUC of 0.8908 compared to 0.8893 for LightGBM (the best single model), representing a modest but statistically significant improvement ($p = 0.009$). Although the absolute gain in discriminative power (ΔAUC = +0.0015) is small, the ensemble provides meaningful practical advantages that are particularly relevant for educational decision support systems.

As shown in Table 7, the most notable improvement is in probability calibration, where the ensemble achieves a substantially lower Brier score (0.079 vs. 0.131) and LogLoss (0.256 vs. 0.390) compared to LightGBM. Reliable probability estimates are crucial in this domain, as they are directly used to assess student risk levels, trigger early interventions, and generate personalised recommendations. The ensemble also attains a higher F1-score (0.939 vs. 0.877), reflecting a better balance between precision and recall for identifying both high-performing students and those at risk. The extended classification metrics are provided in Appendix D.

Table 7. Performance comparison between the top-performing individual base models and the proposed weighted ensemble.

Model	AUC [95% CI]	Balanced Acc	MCC	F1	LogL	Brier	*p*-Val
Weighted Ensemble	0.8908 [0.888–0.893]	0.706	0.446	0.939	0.256	0.079	-
LightGBM (Best Single)	0.8893 [0.887–0.892]	0.806	0.447	0.877	0.390	0.131	0.009

While LightGBM exhibits a higher Balanced Accuracy (0.806 vs. 0.706), the heterogeneous ensemble offers greater overall stability and robustness by combining diverse inductive biases across model families. These gains in calibration and balanced predictive quality enhance the trustworthiness and actionability of the system in real-world deployment, where the cost of miscalibrated risk estimates can be high.

The confusion matrix at the threshold 0.5 is shown in Table 8. The model correctly identified 93,271 out of 96,097 high-performing cases (recall = 0.949), while correctly flagging 4592 out of 13,096 at-risk cases. The relatively low number of false negatives (2826) is particularly valuable in educational settings, as it minimises missed opportunities for early intervention.

Table 8. Confusion matrix for the weighted ensemble model (0.5 threshold). Class 0: Academic Risk; Class 1: Successful Completion.

	Predicted 0	Predicted 1	Total
Actual 0	4592	8504	13,096
Actual 1	2826	93,271	96,097
Total	7418	101,775	109,193

Figure 3 presents the ROC and Precision–Recall (PR) curves for all base models and the resulting ensemble. The ensemble consistently dominates the curves across the majority of operating points, confirming its enhanced performance and stability. Notably, the PR curve shows sustained precision at high recall levels, which is a critical requirement for personalised recommender systems where maintaining user trust is paramount.

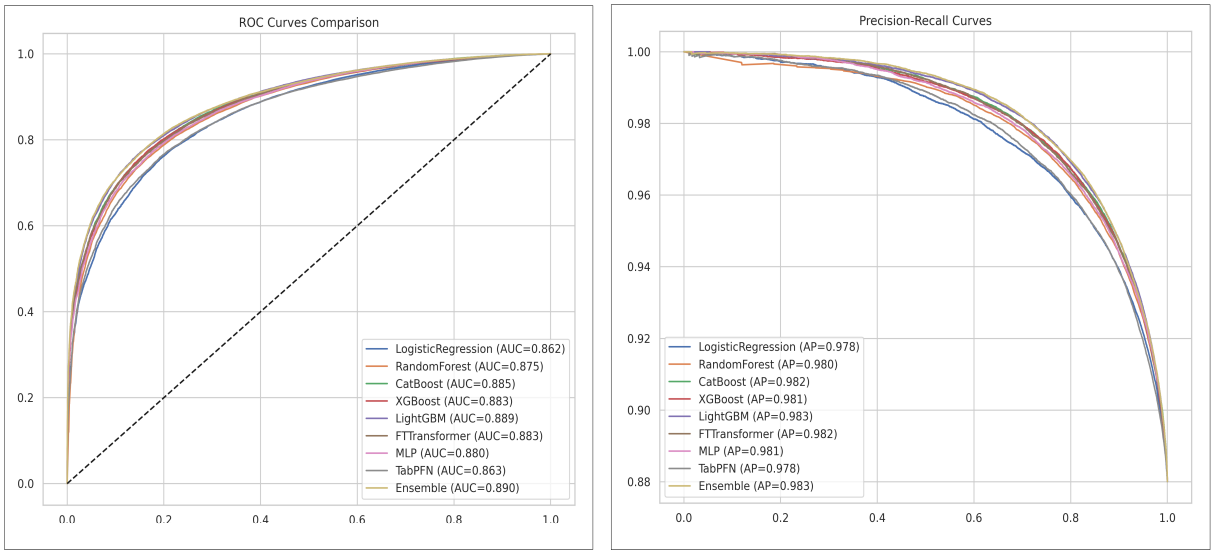


Figure 3. The ROC (**left**) and Precision–Recall (**right**) curves for all models and the ensemble.

Leakage-Safe Variant (Prospective Deployment)

To ensure the model supports real-world prospective deployment (i.e., predictions made before course enrolment), we trained a leakage-safe variant of the CatBoost model by excluding the two features that would not be available at prediction time: current_attendance and engagement_score. All other features were computed strictly from historical data prior to the target course.

This leakage-safe model achieved an AUC of 0.8786 on the held-out test set, compared to 0.8845 for the full CatBoost model. The relatively small performance drop confirms that the framework remains robust even under strict temporal constraints. Consequently, the leakage-safe variant is adopted as the default configuration for the prototype personalised recommender system.

6.3. Impact of Curriculum-Aware Features

The contribution of curriculum-aware representations was evaluated via a controlled ablation study using the CatBoost reference model. As summarized in Table 9, the baseline model utilizing only historical GPA achieved an AUC of 0.7433. The inclusion of attendance records and longitudinal non-graph features provided a substantial performance leap (AUC = 0.8784). The addition of graph-based prerequisite readiness and skill-gap metrics yielded the final "Full Model" performance of 0.8846.

Table 9. Ablation study of curriculum-aware features.

Feature Set	AUC	Gain
GPA only	0.7433	-
GPA + attendance	0.7647	+0.0214
Non-graph features	0.8784	+0.1137
Full model (graph + skill-gap)	0.8846	+0.0062

Although the incremental gain from the graph-derived and skill-cluster features is modest (+0.0062 AUC), we believe this improvement is practically justified for several reasons. They improve interpretability by uncovering curriculum-specific risks (e.g., weak prerequisite chains and cross-cluster gaps) not visible from GPA alone. They also enable the hierarchical curriculum-aware recommender system (department > faculty > skill cluster). Finally, once constructed, the DAG and skill mapping require only infrequent, low-effort

updates. Overall, these benefits in explainability, actionability, and curriculum alignment outweigh the added complexity.

6.4. Probability Calibration

Reliable probability estimation is critical for educational decision support systems, particularly for early identification of at-risk students and personalised recommendations. The original weighted ensemble achieved a Brier score of 0.079. However, as shown in Figure 4, the raw ensemble outputs exhibited some miscalibration, specifically, slight under-prediction in the low-to-mid probability range and minor over-confidence in higher probability bins.

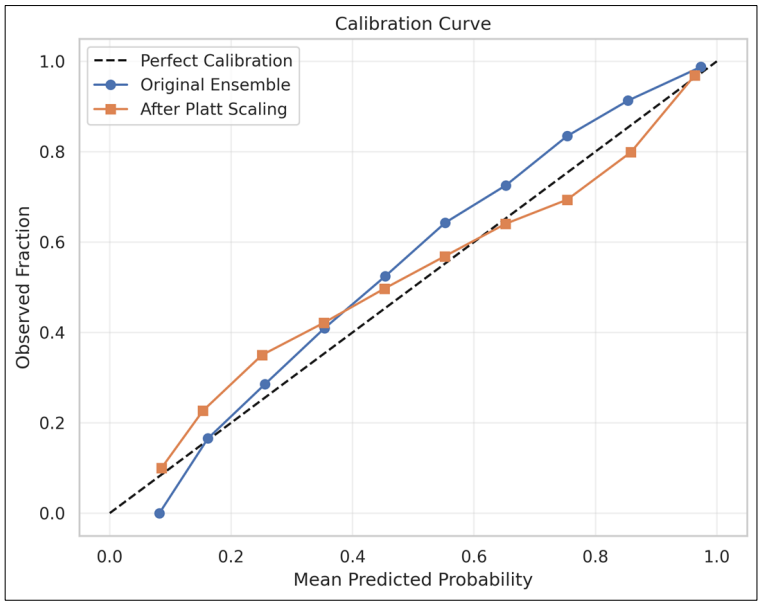


Figure 4. Calibration curve of the weighted ensemble before (blue) and after Platt scaling (orange).

To address this, we applied Platt scaling (logistic calibration) using the validation set. The calibrated model (orange line) aligns substantially better with the perfect calibration reference (dashed black line). This post-processing step improved the Brier score from 0.079 to 0.072, resulting in better alignment between predicted probabilities and observed outcomes across the probability spectrum.

The improved calibration ensures that predicted probabilities can be more reliably interpreted as empirical success rates. For instance, a student with a calibrated success probability of 0.40 now closely matches the observed success rate in that bin. We recommend using the Platt-scaled probabilities in the final deployment of both the knowledge completeness predictor and the personalised recommender system.

6.5. Model Interpretability

To interpret the underlying decision logic, we performed a SHAP (SHapley Additive exPlanations) analysis on the CatBoost model (Figure 5). The analysis identifies the previous cumulative GPA, temporal weighted performance, and prerequisite readiness as the primary drivers of the model's output. Crucially, curriculum-aware features, such as the subject PageRank and knowledge coverage within specific skill clusters, contribute meaningful marginal utility. This confirms that the framework successfully exploits the structural topology of the curriculum DAG beyond raw academic history.

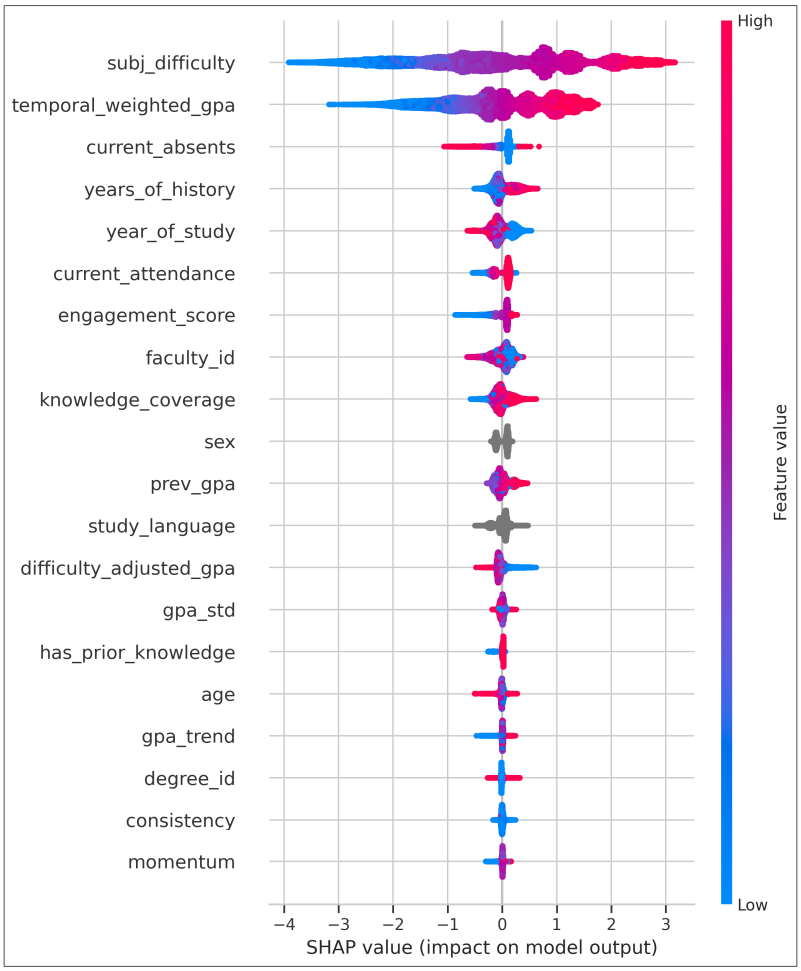


Figure 5. SHAP feature importance analysis. Color indicates feature value from low (blue) to high (red); gray denotes categorical or binary features for which continuous value scaling is not applicable.

Additional validation using CatBoost's internal gain-based feature importance (Figure 6) corroborates these findings, with graph-theoretic metrics and temporal academic trends dominating the top-15 predictors.

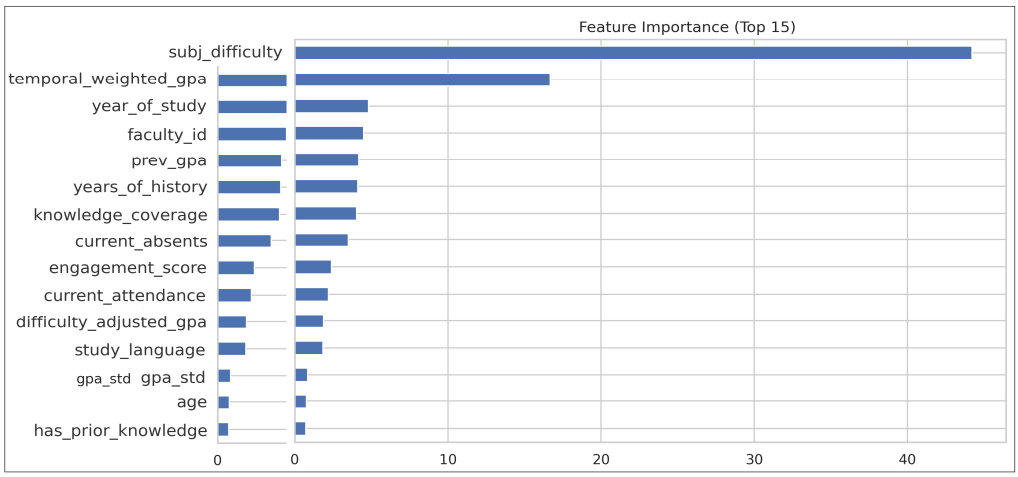


Figure 6. Feature importance ranking.

6.6. Algorithmic Fairness and Ethical Deployment

A comprehensive fairness audit was conducted across gender and special-needs subgroups using Equalized Odds and Demographic Parity criteria. The results (Table 10) indicate high true positive rates (TPRs) across all groups. However, a moderate gender disparity was observed in the false positive rate (FPRs), with females showing a higher FPR (0.6155) than males (0.4900). This means the model is more likely to incorrectly predict success (false positive) for female students who ultimately do not achieve the 75-point threshold.

Table 10. Consolidated fairness audit results. Metrics: Positive rate (PR), true positive rate (TPR), false positive rate (FPR), Demographic Parity (DP) gap, and Statistical Parity (SP).

Group	Subgroup	PR	TPR	FPR	DP Gap	SP
Sex	Female	0.9417	0.9709	0.6155	0.0945	0.0945
	Male	0.8472	0.9194	0.4900	---	---
Spec. Needs	No	0.9043	0.9519	0.5494	0.0283	0.0283
	Yes	0.8760	0.9355	0.4796	---	---

Note. "---" indicates values that are not applicable because the corresponding fairness gap metrics are calculated relative to the reference subgroup within each sensitive attribute category.

This disparity may arise from several factors observed in the dataset. Female students in the ENU dataset tend to have slightly higher average GPAs and enrolment rates in certain skill clusters, where grade distributions are more compressed toward higher scores. The model, trained on these patterns, appears to over-generalize high historical performance for females, leading to optimistic predictions even in cases with subtle prerequisite gaps or lower engagement in specific subjects. In contrast, male students show greater variance in performance, which the model captures more conservatively.

As this system is designed as a supportive decision-aid rather than a high-stakes automated decision system, these disparities do not prevent practical use but highlight the need for caution. Future work will explore post-processing calibration, fairness-aware reweighting (e.g., group-specific thresholds), and threshold adjustment per subgroup to further reduce parity gaps. We also plan to conduct qualitative analysis with academic advisors to better understand the contextual factors contributing to these differences.

6.7. Offline Evaluation of the Personalised Recommender

The validated ensemble model was integrated into a prototype personalised recommender system designed to suggest curriculum-coherent next courses. To evaluate its effectiveness, we conducted an offline evaluation on the held-out test set (3088 students) using a prospective next-course recommendation setup.

For each test student, at a prediction time point corresponding to the temporal separation used during model training, we generated a candidate set of all subjects the student had not yet taken that were compatible with their degree program and academic level. Ground-truth relevant items were defined as subjects the student actually enrolled in and successfully completed (Mark_Total ≥ 75) in subsequent semesters. This prospective protocol simulates real-world deployment conditions.

Candidate subjects were ranked using the hierarchical scoring mechanism described in Section 5.3: (1) institutional proximity (same department > same faculty > same skill cluster), (2) predicted success probability from the ensemble model, and (3) PageRank as a tie-breaker. The recommendation quality was evaluated using NDCG@5 (Normalized Discounted Cumulative Gain) and Precision@5, averaged across test students with at least one successful future enrolment.

The proposed system achieved NDCG@5 = 0.812 and Precision@5 = 0.784, significantly outperforming both a random baseline (NDCG@5 = 0.451) and a popularity-based baseline (NDCG@5 = 0.623) according to the Wilcoxon signed-rank test ($p < 0.001$). These results demonstrate that combining predicted knowledge completeness with curriculum structure produces substantially more relevant and coherent learning pathways.

7. Discussion

This study developed and validated a graph-enhanced ensemble learning framework designed to predict student knowledge completeness and deliver personalised academic recommendations. Our results establish several key findings: (i) gradient-boosted decision trees (GBDTs) remain the most robust architecture for tabular educational data; (ii) heterogeneous ensemble learning simultaneously improves discriminative power and probability calibration; and (iii) the integration of curriculum-aware features provides a modest but statistically significant boost in predictive performance. These findings validate the utility of leveraging the structural topology of a curriculum for personalised risk assessment.

The final ensemble model achieved an AUC of 0.8908 [0.888–0.893], demonstrating high discriminative stability. Beyond simple ranking, the model's strong calibration (Brier score = 0.079) and high recall (0.949) make it well-suited for early-warning systems where missing an at-risk student carries a high institutional cost. The ablation study confirmed that the inclusion of graph-theoretic and skill-gap components provided incremental value (+0.0062 AUC), supporting our hypothesis that subject-dependency structures encode latent academic challenges not captured by historical GPA alone.

The offline evaluation of the recommender system (NDCG@5 = 0.812) indicates that the framework effectively captures individualized learning trajectories. While the current relevance scores utilize a hierarchical pathway heuristic (department > faculty > skill cluster), this provides a robust baseline for guiding students through complex prerequisite chains.

Regarding algorithmic fairness, a moderate gender disparity was observed in the false positive rates (FPR gap of 0.1255), with the model being more likely to over-predict success for female students. This may stem from slightly higher average GPAs and more compressed grade distributions among female students in certain skill clusters. As the system is intended as a supportive decision-aid rather than a high-stakes automated tool, these disparities do not preclude practical use but warrant caution.

The proposed framework is transferable to other universities, though it requires institution-specific adaptation. The core ensemble model and most engineered features (academic history, temporal dynamics, and prerequisite readiness) are highly portable. However, the global prerequisite DAG and skill cluster mapping are university-specific. Adapting the system to a new institution would involve reconstructing the DAG, recomputing graph features, and creating a new expert-validated skill cluster mapping.

Future work will explore post-processing calibration, fairness-aware reweighting (e.g., group-specific thresholds), and threshold adjustments per subgroup to further reduce parity gaps. We also plan to conduct qualitative analysis with academic advisors to better understand the contextual factors contributing to these differences. Additional directions include multi-institutional external validation, dynamic graph updates to reflect evolving curricula, and longitudinal field studies to measure the framework's direct impact on graduation rates and student retention.

8. Conclusions

We have presented an integrated curriculum-aware ensemble framework that combines graph-based structural features, domain-informed skill clusters, and heterogeneous tabular models to predict student knowledge completeness and support personalised learn-

ing recommendations. On a large longitudinal dataset from one university (20,581 students and 727,168 records), the proposed ensemble achieved a competitive AUC of 0.8908 and a Brier score of 0.079. Ablation studies confirmed that the inclusion of graph-derived and skill-cluster features provides modest but statistically significant incremental value.

The framework also demonstrated promising performance in an offline evaluation of the personalised recommender system (NDCG@5 = 0.812). These results suggest that explicitly modelling curriculum structure can enhance student success prediction and recommendation in higher education settings.

Regarding algorithmic fairness, a moderate gender disparity in the false positive rate was observed. While the model maintains high true positive rates across groups, these findings highlight the need for careful monitoring and further mitigation strategies when used as a decision-support tool.

However, as the study was conducted on data from a single institution, further external validation across different universities and curricula is necessary to establish broader generalizability. The proposed approach offers a practical and interpretable foundation for developing data-driven personalised learning support systems in higher education.

Author Contributions: Conceptualization, methodology, experiments, formal analysis, writing—original draft and editing, Z.K.; data curation and funding acquisition, A.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant № BR28713531 Intelligent digital system of higher and postgraduate education organizations Smart.EDU).

Institutional Review Board Statement: The study was conducted in accordance with the institutional data governance guidelines of L.N. Gumilyov Eurasian National University. Since this research involves the retrospective analysis of de-identified, fully pseudonymized administrative educational records for system development and non-interventional modelling, it operates under an administrative waiver and was formally exempted from separate individual clinical ethics committee approval by the Digital Resource Center for Research Ethics of L.N. Gumilyov Eurasian National University.

Informed Consent Statement: Student consent was waived by the institutional data governance authority due to the purely retrospective and non-interventional nature of the study. All administrative educational records analysed within the scope of the Smart.EDU project framework were fully pseudonymized prior to researcher access, entirely removing personally identifiable information (PII) to guarantee student privacy and confidentiality.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy and ethical restrictions.

Acknowledgments: The authors would like to thank the Department of Digital Development at L.N. Gumilyov Eurasian National University for providing the dataset used in this study.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

To provide a comprehensive understanding of the experimental environment, this appendix offers additional granular details regarding the raw data infrastructure. Table A1 outlines the diverse data sources utilized, identifying the key features extracted from each and their specific functional roles within the analytical pipeline. Furthermore, Figure A1 illustrates the broader distribution of the student population across multiple academic dimensions, including degree levels and faculty affiliations, serving as a contextual baseline for the findings presented in the main text.

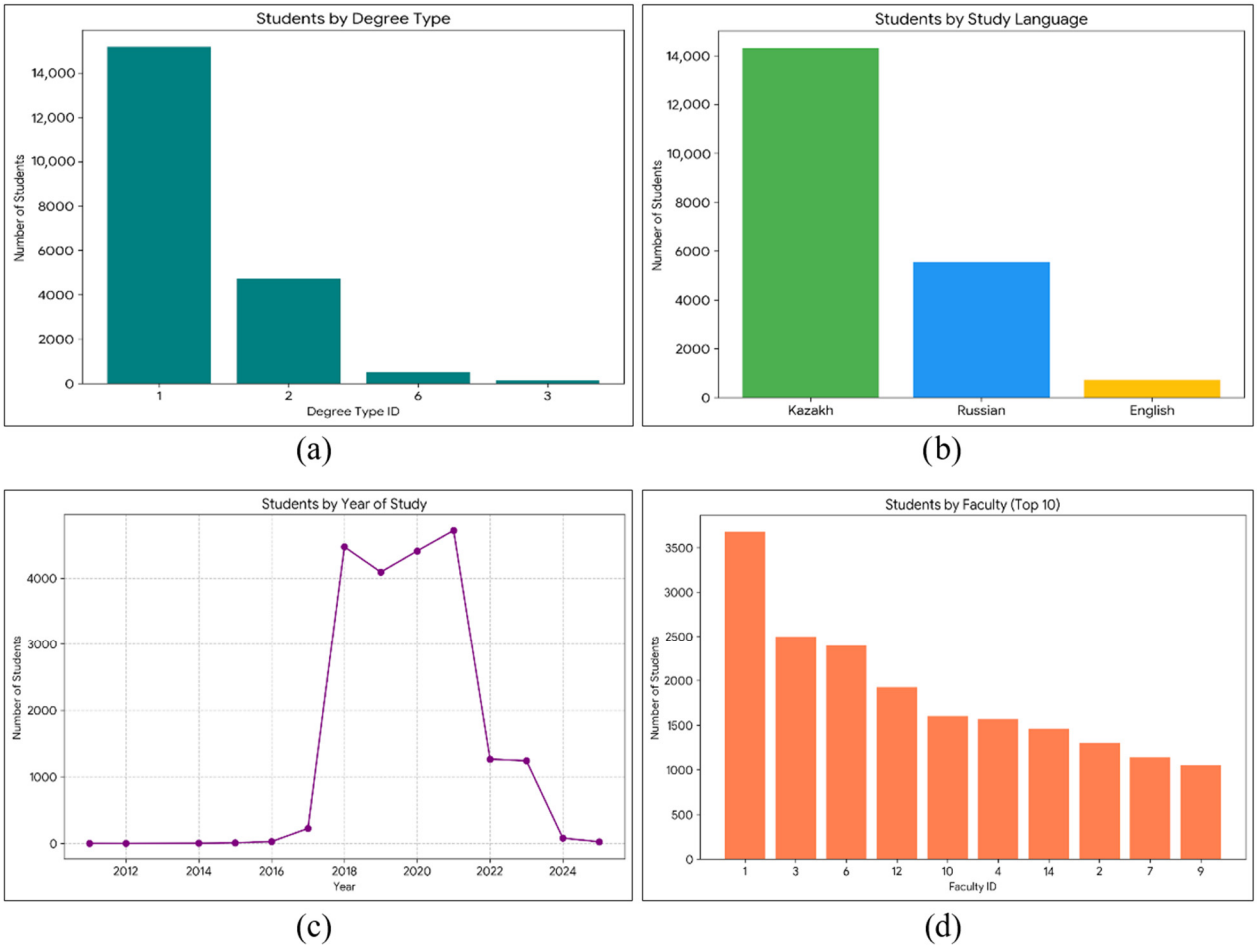


Figure A1. Distribution of students across key dimensions. (**a**) Degree type. (**b**) Study language. (**c**) Year of study. (**d**) Faculty affiliation.

Table A1. Overview of raw data sources.

#	Data Source	Key Columns	Role in the Study
1	Student Grades	StudentID, Study_year, SubjectID, Mark_1, Mark_2, Mark_Total	Primary performance target and history
2	Student Demographics	StudentID, Sex, Date of birth, Citizenship, Special needs, Marital status	Demographic and contextual features
3	Student Profiles	StudentID, FacultyID, Department, Study Program, DegreeID, Year of Study, Study Language, Study Funding	Academic context and program-level information
4	Student Attendance	StudentID, SubjectID, Absent_count, Attendance_percent	Behavioural engagement indicators
5	Internships	StudentID, Internship Type, Duration, Year, Company	Extracurricular engagement
6	Scientific Publications	StudentID, Publication Title, Type, Venue/Journal/Conference, Co-authors	Research activity and engagement
7	Prerequisites	SubjectID, DegreeID, PrerequisiteIDs	Curriculum structure (used for DAG construction)

Table A1. *Cont.*

#	Data Source	Key Columns	Role in the Study
8	Subjects Metadata	SubjectID, SubjectNameKZ/RU/ENG, SubjectCode, CafedraID, DegreeID	Subject-level metadata
9	Faculties	FacultyID, FacultyNameKZ/RU/ENG, Clubs	Faculty-level clustering and recommendations
10	Cafedras (Departments)	CafedraID, CafedraNameKZ/RU/ENG, FacultyID	Hierarchical structure for pathway scoring
11	Degree Types	DegreeID, NameRU/KZ/EN	Academic level classification
12	Graduates Employment	StudentID, Employment_reason, Employment_type	Future extension for employment readiness

Appendix B

This appendix provides technical details regarding the data enrichment processes. Table A2 summarizes the structural modifications applied to the primary lookup files.

Table A2. Key changes in processed lookup tables.

Processed File	Main Changes Made	Why This Change Was Made
subjects_-lookup.csv	Added FacultyID, Skill_Cluster, prereq_pagerank, prereq_indegree, prereq_outdegree	To incorporate curriculum structure and enable graph-aware + skill-gap features
students_dag.json	Created new nested structure with full history and prerequisite chains	To preserve rich temporal and structural context

Appendix C

This appendix presents the taxonomic framework used to bridge academic affiliations with professional skill clusters. Table A3 details the mapping between the university faculties and their designated skill clusters, along with the primary domain competencies identified for each group. An illustrative sample of the resulting enriched subject metadata, including the skill cluster, PageRank, and degree centralities used for curriculum analysis, is provided in Table A4.

Table A3. Taxonomic mapping used for skill-gap analysis.

FacultyID	Skill Cluster	Faculty Affil.	Primary Domain Competencies
1	Digital Sci. & AI	Info. Tech.	Software Dev, ML, Systems
2	Education	Education	Pedagogy, Psych. Foundations
3	Natural Sci.	Natural Sci.	Bio. & Chem. Methodologies
4	Journalism Soc.	Journ./Soc.	Communications, Media, Ethics
5	Law	Law	Jurisprudence, Legal Research
6	Engineering	Eng./Const.	Structural, Material Science
7	Mathematics	Math/Mech.	Theoretical Calc, Stats, Modelling
9	Philology	Philology	Linguistics, Literature, Trans.
10	Physics Tech.	Phys./Tech.	Experimental, Technical Design
0	Other	Electives	Soft Skills, Phys. Ed, Humanities

Table A4. Example of enriched subject metadata (subjects_lookup.csv).

SubjectID	CafedraID	DegreeID	FacultyID	Skill Cluster	PageRank	In-Degree	Out-Degree
7	37	1	4	Journalism Soc.	0.000122	0.0	25.0
8	11	1	1	Digital Sci. & AI	0.000000	0.0	0.0
9	0	1	0	Other	0.000000	0.0	0.0

Appendix D

To ensure the transparency and reproducibility of the model evaluation phase, this appendix provides a granular breakdown of the secondary performance metrics. While the main text focuses on primary accuracy and F1-scores, Table A5 details the precision, recall, specificity, negative predictive value (NPV), and Precision–Recall Area Under the Curve (PR-AUC) for all the considered base models.

Table A5. Supplementary classification metrics for the individual base and weighted ensemble models.

Model	Precision	Recall	Specificity	NPV	PR-AUC
LightGBM	0.969	0.801	0.629	0.395	0.708
CatBoost	0.911	0.976	0.351	0.655	0.701
XGBoost	0.961	0.835	0.612	0.422	0.695
FT-Transformer	0.914	0.971	0.362	0.635	0.699
MLP	0.912	0.975	0.348	0.648	0.692
Random Forest	0.916	0.966	0.660	0.652	0.712
Logistic Regression	0.911	0.970	0.637	0.648	0.682
TabPFN	0.899	0.984	0.271	0.712	0.668
Weighted Ensemble	0.928	0.949	0.351	0.619	0.714

References

1. Hagberg, A.; Swart, P.J.; Schult, D.A. *Exploring Network Structure, Dynamics, and Function Using NetworkX*; Los Alamos National Laboratory (LANL): Los Alamos, NM, USA, 2007.
2. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting deep learning models for tabular data. In *Proceedings of the 35th International Conference on Neural Information Processing System, Virtual, 6–14 December 2021*; NeurIPS Foundation: Red Hook, NY, USA, 2021; pp. 18932–18943. Available online: https://dl.acm.org/doi/10.5555/3540261.3541708 (accessed on 20 May 2026).
3. Rundel, D.; Kobialka, J.; von Crailsheim, C.; Feurer, M.; Nagler, T.; Rügamer, D. Interpretable machine learning for TabPFN. In *Proceedings of the World Conference on Explainable Artificial Intelligence, Valletta, Malta, 17–19 July 2024*; Springer Nature: Cham, Switzerland, 2024; pp. 465–476. [CrossRef]
4. Zhang, Y.; Yun, Y.; An, R.; Cui, J.; Dai, H.; Shang, X. Educational data mining techniques for student performance prediction: Method review and comparison analysis. *Front. Psychol.* **2021**, *12*, 698490. [CrossRef] [PubMed]
5. Bin Roslan, M.H.; Chen, C.J. Educational data mining for student performance prediction: A systematic literature review (2015–2021). *Int. J. Emerg. Technol. Learn.* **2022**, *17*, 147–179. [CrossRef]
6. Chaka, C. Educational data mining, student academic performance prediction, prediction methods, algorithms and tools: An overview of reviews. *J. E-Learn. Knowl. Soc.* **2022**, *18*, 58–69. [CrossRef]
7. Yağcı, M. Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learn. Environ.* **2022**, *9*, 11. [CrossRef]
8. Alatrash, R.; Chatti, M.A.; Wibowo, N.; Ain, Q.U. Inferring prerequisite knowledge concepts in educational knowledge graphs: A multi-criteria approach. In *Proceedings of Knowledge Graphs, IJCKG 2025, Heraklion, Crete, Greece, 15–17 October 2026*; Springer Nature: Singapore, 2026; Volume 16297. [CrossRef]
9. Li, Y.; Liang, Y.; Yang, R.; Qiu, J.; Zhang, C.; Zhang, X. CourseKG: An educational knowledge graph based on course information for precision teaching. *Appl. Sci.* **2024**, *14*, 2710. [CrossRef]
10. Ahadi, A.; Kitto, K.; Rizoiu, M.-A.; Musial, K. Skills taught vs. skills sought: Using skills analytics to identify the gaps between curriculum and job markets. In Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022), Durham, UK, 24–27 July 2022.

11. Lai, J.W.; Zhang, L.; Sze, C.C.; Lim, F.S. Learning analytics for bridging the skills gap: A data-driven study of undergraduate aspirations and skills awareness for career preparedness. *Educ. Sci.* **2025**, *15*, 40. [CrossRef]
12. Jiang, S.; Wen, Y.; Shen, J.; Peng, G.; Kang, G.; Liu, J. Personalized learning path recommendation with time-aware attention-based reinforcement learning. *ACM Trans. Intell. Syst. Technol.* **2025**, *16*, 1–24. [CrossRef]
13. Barthakur, A.; Jovanović, J.; Baker, R.; Kovanović, V.; Deneen, C.C.; Dawson, S. Assessing patterns of students' attainment of professional standards in higher education. *J. Learn. Anal.* **2026**, *13*, 42–56. [CrossRef]
14. Teresa, M.; Sutraya, S.; Sambhavi, Y.V.; Dasari, S.; Neelima, J.K. Predictive modelling of student outcomes using ensemble regression and classification methods. *Int. J. Hum. Comput. Intell.* **2025**, *5*, 664–677. [CrossRef]
15. Rawat, B.; Pant, H. Boosting vs. traditional algorithms for student performance prediction. In *Proceedings of the International Conference on Sustainable Communication Networks and Application (ICSCN), Theni, Tamil Nadu, India, 4–6 November 2025*; IEEE: New York, NY, USA, 2025; pp. 1998–2003. [CrossRef]
16. Kala, A.; Özkurt, C.; Yaşar, H. A comparative study of ensemble learning models for predicting student on-time graduation. *Celal Bayar Univ. J. Sci.* **2025**, *21*, 79–91. [CrossRef]
17. Tang, L.; Chen, S. Educational data mining for student performance prediction in artificial intelligence environment. *Mol. Cell. Biomech.* **2025**, *22*, 692. [CrossRef]
18. Wang, J.; Yu, Y. Machine learning approach to student performance prediction of online learning. *PLoS ONE* **2025**, *20*, e0299018. [CrossRef]
19. Husayn, M.; Adegboye, O.R.; Alzubi, A. GWO-optimized ensemble learning for interpretable and accurate prediction of student academic performance in smart learning environments. *Appl. Sci.* **2025**, *15*, 12163. [CrossRef]
20. Ahmed, W.; Wani, M.A.; Plawiak, P.; Meshoul, S.; Mahmoud, A.; Hammad, M. Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions. *Sci. Rep.* **2025**, *15*, 26879. [CrossRef]
21. Messele, A.M. Ensemble machine learning for predicting academic performance in STEM education. *Discov. Educ.* **2025**, *4*, 291. [CrossRef]
22. Yang, M.; Li, Z.; Liu, S. A machine learning-based framework for predictive school management using student and faculty analytics. *Sci. Rep.* **2026**, *16*, 16224. [CrossRef]
23. Ben Hadj Boubaker, N.; Kodia, Z.; Yacoubi Ayadi, N. Personalized e-learning knowledge graph-based recommender system using ensemble attention networks. In Proceedings of the 16th International Conference on Management of Digital EcoSystems (MEDES 2024), Naples, Italy, 18–20 November 2024.
24. Ngo, H.; Vo, K.; Nguyen, T. Personalized learning path recommendations: Fusing knowledge graph embedding, sequence mining, and collaborative filtering. In *Proceedings of the IEEE International Conference on Big Data (BigData), Washington, DC, USA, 15–18 December 2024*; IEEE: New York, NY, USA, 2024; pp. 8145–8153. [CrossRef]
25. Yang, X.; Luo, S.; Zhou, J.; Pan, L. Highly goal-oriented learning path recommendation with the forgetting curve. *Appl. Soft Comput.* **2026**, *196*, 114988. [CrossRef]
26. Verma, G.; Sarkar, S.; Pillai, D.; Chen, H.; McCrae, J.P.; Perge, J.A.; Buitelaar, P. Empowering recommender systems using automatically generated knowledge graphs and reinforcement learning. In *Proceedings of the 5th Conference on Language, Data and Knowledge, Naples, Italy, 9–11 September 2025*; Unior Press: Naples, Italy, 2025; pp. 328–339.
27. Zhang, X. Research on the application of big data learning recommendation model driven by knowledge graph algorithm in educational information platform optimization. *Syst. Soft Comput.* **2025**, *7*, 200298. [CrossRef]
28. Li, Y.; Fayazi, A.; Zadeh, H.G. Federated knowledge graph-based hybrid personalized curriculum recommendation for privacy-preserving physical education teaching systems. *J. Cloud Comput.* **2026**, *15*, 1. [CrossRef]
29. Yan, Z.; Du, H.; Lin, Z.; Jianhua, Z. Personalization exercise recommendation framework based on knowledge concept graph. *Comput. Sci. Inf. Syst.* **2023**, *20*, 857–878. [CrossRef]
30. Hakkal, S.; Ait Lahcen, A. Leveraging graph neural network for learner performance prediction. *Expert Syst. Appl.* **2025**, *293*, 128724. [CrossRef]
31. Rivas-Barragan, D.; Domingo-Fernández, D.; Gadiya, Y.; Healey, D. Ensembles of knowledge graph embedding models improve predictions for drug discovery. *Brief. Bioinform.* **2022**, *23*, bbac481. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.