A CASE-BASED REASONING FRAMEWORK WITH SOFT SIMILARITY AND SUBMODULAR OPTIMIZATION FOR PERSONALIZED CAREER GUIDANCE

Shodmonov Davronjon Abduvaliyevich[1]

[1]*Samarkand Branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi,*
E-mail: *shodmonov@samtuit.uz*

K E Y W O R D S

Case-Based Reasoning; Conversational CBR; Career Guidance; Hybrid Similarity; Submodular Optimization; Educational Data Mining; Explainable AI

A B S T R A C T

The increasing availability of longitudinal academic data in higher education creates new opportunities for personalized and evidence-based career guidance. However, many existing AI-driven guidance systems rely on black-box predictive models and provide limited interpretability or actionable recommendations. This paper proposes a principled Case-Based Reasoning (CBR) framework for personalized career guidance that leverages institutional academic data and historical student trajectories.

The proposed approach models each graduate as a structured case comprising academic profiles, temporal learning trajectories, contextual attributes, observed outcomes, and contributing actions. A hybrid similarity function is employed to retrieve relevant cases from heterogeneous educational data, while a soft-retrieve aggregation mechanism ensures robustness under partial observability and noisy inputs. To bridge the gap between relevance estimation and actionable guidance, the reuse phase is formulated as a constrained submodular optimization problem, enabling the selection of compact and feasible action sets with theoretical approximation guarantees.

Theoretical analysis establishes stability of the soft-retrieve mechanism, probabilistic consistency of outcome inference, and near-optimality of the reuse strategy. Experimental evaluation on real institutional data collected through a higher education management information system demonstrates that the proposed framework consistently outperforms classical CBR and baseline machine learning approaches in outcome prediction. Moreover, submodular reuse optimization significantly improves action recommendation quality by increasing competency coverage and reducing redundancy, particularly in low-cardinality settings that are most relevant for practical advising.

The results indicate that combining case-based reasoning with submodular optimization offers a robust, explainable, and deployable solution for personalized career guidance in higher education.

1. INTRODUCTION

The rapid transformation of labor markets under the influence of digital technologies and artificial intelligence has significantly increased the demand for personalized and data-driven career guidance systems in higher education. Universities are expected not only to provide academic knowledge but also to support students in making informed career decisions based on their competencies, learning trajectories, and labor market relevance. This challenge is particularly pronounced in information and engineering-oriented disciplines, where educational paths are highly heterogeneous and professional outcomes are strongly influenced by individual learning histories.

Traditional career guidance approaches—such as psychometric testing, rule-based expert systems, or static recommender models—have demonstrated limited effectiveness in modern higher education environments. These methods often rely on abstract assessments, require complete input data, and provide recommendations that are difficult to interpret or justify to students and academic advisors. Recent surveys and systematic reviews highlight that many AI-based career guidance systems lack transparency, adaptability, and alignment with real educational trajectories, which limits their practical adoption in universities [1,2,3].

In contrast, Case-Based Reasoning (CBR) has emerged as a promising paradigm for decision

(cc) BY © *Shodmonov Davronjon*

support systems operating in domains characterized by experiential knowledge, data sparsity, and the need for explainability. CBR solves new problems by retrieving and adapting solutions from previously observed cases, thereby leveraging historical experience rather than abstract models. Since its foundational formulation, CBR has been successfully applied in diverse engineering and decision-support domains, demonstrating robustness to incomplete data and natural interpretability of recommendations [4,7]. These properties make CBR particularly suitable for educational and career guidance contexts, where decisions must be explainable and grounded in real student trajectories.

Recent studies further emphasize the importance of conversational and hybrid extensions of classical CBR. Conversational Case-Based Reasoning (CCBR) introduces interactive mechanisms that allow systems to incrementally acquire missing information through targeted dialogue, thereby reducing uncertainty and improving recommendation quality when initial student profiles are incomplete [8,9]. In parallel, hybrid CBR models integrating fuzzy logic, clustering, and similarity learning have shown improved retrieval stability and scalability, especially when handling heterogeneous educational data [10,11,12]. These advances align well with the characteristics of academic datasets collected in higher education institutions.

In the context of Uzbekistan's higher education system, the nationwide deployment of the Higher Education Management Information System (HEMIS) provides a unique opportunity to operationalize data-driven career guidance at scale. HEMIS accumulates longitudinal academic records of students, including semester-wise performance, course structures, credit loads, and graduation outcomes. While these data are highly reliable and comprehensive from an academic perspective, they are rarely exploited for intelligent decision support beyond administrative analytics. Bridging this gap requires methodological frameworks capable of transforming raw academic trajectories into actionable and interpretable career recommendations.

However, existing career recommendation systems rarely exploit longitudinal academic trajectories in a principled mathematical manner. Many approaches rely on flat feature vectors or black-box machine learning models, which obscure the influence of individual learning events and provide limited theoretical guarantees. Recent literature calls for explainable, mathematically grounded recommendation models that can operate under real-world constraints such as partial observability, limited budgets, and curricular dependencies [13,15]. Addressing these requirements remains an open research challenge.

This paper proposes a novel Case-Based Reasoning framework for personalized career guidance that integrates soft similarity aggregation, conversational refinement, and submodular optimization. The proposed approach models each graduate's educational history as a structured case comprising academic profiles, temporal learning trajectories, contextual attributes, and realized career outcomes. A hybrid similarity function is employed to retrieve relevant cases using a soft-weighted mechanism, ensuring robustness and stability. Recommended actions – such as courses, skill-building activities, or career paths – are selected through a constrained submodular optimization process, enabling theoretical approximation guarantees and practical scalability.

The main contributions of this work are summarized as follows:

1. A structured case representation that leverages longitudinal academic data derived from university information systems.
2. A hybrid similarity model with soft-retrieve aggregation, improving robustness to noisy and incomplete student profiles.
3. A submodular optimization-based reuse mechanism with formal approximation guarantees.
4. A theoretically grounded and explainable CBR framework tailored for personalized career guidance in higher education.

The remainder of the paper is organized as follows. Section 2 reviews related work on CBR and career recommendation systems. Section 3 formulates the problem and introduces the proposed framework. Section 4 presents the submodular reuse optimization model. Section 5 provides theoretical analysis and complexity considerations. Section 6 reports experimental

(cc) BY © Shodmonov Davronjon

design and evaluation results. Finally, Section 7 concludes the paper and outlines future research directions.

2. RELATED WORKS

Personalized career guidance sits at the intersection of educational data mining, recommender systems, and decision support. The literature can be grouped into three closely connected streams: (i) AI-driven career guidance and educational pathway recommendation, (ii) case-based reasoning (CBR) and its conversational/hybrid extensions, and (iii) optimization-oriented selection mechanisms for producing actionable, constraint-aware recommendations.

2.1. AI-driven career guidance and pathway recommendation

Recent reviews emphasize a growing shift from traditional counselling tools to AI-supported guidance platforms, driven by the availability of student information systems and learning analytics. Surveys and scoping reviews report that modern systems increasingly rely on machine learning models for outcome prediction (e.g., employability, career domain, or major selection), yet frequently face limitations related to interpretability, transferability across institutions, and the mismatch between model outputs and the actionable decisions students must make (courses, projects, internships) [2,3]. Comparative studies of recommender approaches for university major and career guidance also note that purely predictive models often provide "what" (a label) rather than "why" (explanations grounded in examples) and "how" (a feasible learning trajectory to reach the target) [13].

A further theme in this stream is that educational data are longitudinal and heterogeneous: performance and engagement evolve by semester, and the "trajectory" often matters more than any single snapshot feature. This motivates models that can explicitly represent temporal learning histories and support robust recommendations under partial observability—typical for early-year students with incomplete profiles [2,13]. These observations align with the needs of institutional deployments where student profiles are incomplete, noisy, and unevenly measured.

2.2. Case-Based Reasoning as an explainable decision-support paradigm

CBR is a well-established paradigm in AI in which a new problem is solved by retrieving similar past cases and adapting their solutions. Foundational work formalized core methodological variations and emphasized CBR's natural suitability for domains where experiential evidence and human-understandable justification are essential [5,6]. A key advantage for educational decision support is that CBR produces recommendations grounded in concrete precedents (e.g., successful graduate trajectories), offering inherently interpretable rationales compared to black-box predictors [4,7]. This is particularly relevant in career guidance, where trust and transparency are critical for adoption by students and advisors [1,2,13].

Another major motivation for CBR in educational contexts is robustness to sparsity. In real academic settings, student data can be partially observed (missing skills, incomplete achievements), inconsistent (measurement differences across cohorts), and temporally irregular. CBR systems can operate effectively under such conditions by leveraging similarity-driven retrieval and incremental knowledge updates (retain), rather than requiring full retraining as in many global machine learning models [4,6,7].

2.3. Conversational CBR (CCBR) for incomplete student profiles

Conversational CBR (CCBR) extends classical CBR by integrating an interactive process to acquire missing attributes via targeted questions. The seminal formulation highlights that, instead of relying on complete initial descriptions, a system can progressively refine a query to improve retrieval quality and reduce ambiguity [8]. More recent analyses of conversational CBR systems stress the importance of question selection strategies, user burden control, and explanation of "why this question is asked," all of which directly affect user satisfaction and decision acceptance [9]. In higher education guidance, these properties are practically valuable: early-stage students often

(cc) BY © Shodmonov Davronjon

cannot provide complete skill inventories, and many relevant attributes (interests, preferred learning mode, career intent) are better elicited interactively rather than inferred.

Within the broader career guidance literature, this conversational dimension complements the limitations of static questionnaires and purely predictive models by allowing adaptive information gathering and maintaining transparency during interaction [1,13]. Consequently, CCBR provides a principled mechanism for aligning institutional records (e.g., semester grades) with user-supplied context (interests, preferences), enabling personalized recommendations even under partial observability.

2.4. Hybrid similarity, fuzzy clustering, and scalable retrieval

A recurring line of work improves CBR retrieval by combining multiple similarity components (numeric, categorical, textual, temporal) and by using clustering to accelerate search and increase robustness to noise. Fuzzy CBR variants, in particular, model uncertainty in case features and allow soft membership across clusters, which can be beneficial for educational data where boundaries between career paths are not crisp [10]. Performance-oriented analyses of fuzzy clustered CBR indicate that clustering can reduce computational cost while maintaining or improving retrieval quality, especially in large case bases [11]. These ideas have also been adopted in engineering decision-support domains, reinforcing their practicality for complex, heterogeneous datasets [10,14].

For educational and career guidance, hybrid similarity is essential because meaningful resemblance may depend simultaneously on academic performance patterns, completed coursework blocks, and contextual features (e.g., program track or institutional constraints). Reviews of career guidance technologies similarly suggest that systems capable of integrating heterogeneous evidence sources are more likely to be effective in real deployments than those relying on single-modality inputs [1,2].

2.5. From retrieved cases to actionable recommendations: selection under constraints

A central gap in many guidance systems is the transition from retrieved evidence (similar students, predicted label) to *actionable and feasible* recommendation bundles. In practice, advising involves selecting a small set of actions (courses, projects, internships) under constraints such as limited time, prerequisite structures, and budget. This aligns naturally with constrained set selection problems. In optimization theory, monotone submodular maximization under cardinality constraints admits strong approximation guarantees using greedy selection, a classical result widely adopted in engineering and information systems [16]. While submodularity has been extensively studied in operations research and machine learning, it is still underutilized in educational career guidance pipelines, where recommendation quality must be balanced against redundancy and feasibility.

This motivates frameworks that combine *a*) evidence-based relevance from similar cases (CBR retrieval) with *b*) mathematically grounded selection mechanisms that produce compact, non-redundant, and constraint-aware recommendation sets. Such integration is consistent with calls in the career guidance literature for methods that are simultaneously effective, interpretable, and deployable in institutional contexts [1,3].

2.6. Positioning of the present work

In summary, prior work establishes:

1. the need for explainable, trajectory-aware guidance systems in higher education [1,2,13],
2. CBR/CCBR as a natural paradigm for interpretability and robustness under incomplete profiles [4-9], and
3. hybrid similarity and fuzzy/clustering enhancements for scalability and noise tolerance [10-12,14].

However, a methodological gap remains in producing *guaranteed, constraint-aware* action bundles from retrieved cases, rather than returning only a predicted career label or a loose list of items.

The present paper addresses this gap by integrating a soft-retrieve hybrid similarity CBR model with a constrained submodular selection mechanism, yielding both interpretability (case-based explanations) and formal performance

© Shodmonov Davronjon

guarantees (approximation bounds) for actionable recommendation packages.

3. PROBLEM FORMULATION

3.1. Data source and scope

Modern higher education institutions increasingly rely on centralized student information systems to collect longitudinal academic data, which form the backbone of learning analytics and AI-driven career guidance research [1,2,3]. Such data sources enable scalable and evidence-based decision support when combined with appropriate modeling frameworks.

In Uzbekistan, the nationwide Higher Education Management Information System (HEMIS) stores structured records describing students' enrollment status, semester-wise academic performance, course credits, and graduation outcomes. These data provide a reliable institutional foundation for modeling educational trajectories and post-graduation outcomes.

The objective of this work is to utilize longitudinal academic records to support personalized career guidance by recommending feasible and interpretable educational actions grounded in prior student experience, rather than producing only abstract career labels, a limitation repeatedly noted in existing career guidance systems [13].

3.2. Case representation

Following the classical Case-Based Reasoning (CBR) paradigm [4,5,7], historical experience is modeled as a collection of cases, where each case represents a completed student trajectory together with its outcome. Formally, each case is defined as

$$c = (x, \tau, z, y, R), \qquad (1)$$

where

x denotes the academic profile (e.g., GPA, credits, course performance),

τ represents the temporal learning trajectory,

z captures contextual attributes (institution, program, cohort),

y is the observed outcome (career domain or continuation of studies), and

$R \subseteq A$ is the set of actions contributing to outcome y.

Similar case representations have been widely used in CBR-based decision-support and recommender systems, including educational applications, due to their interpretability and reusability [7,12,15].

3.3. Query modeling and partial observability

For a current student, the system receives a query defined as

$$q = (x_q, \tau_q, z_q), \qquad (2)$$

where the components correspond to those in Eq. (1) but are typically incomplete. Partial observability is a known characteristic of educational data, especially for early-stage students whose academic trajectories and skill profiles are still emerging [1,13].

To mitigate this issue, conversational mechanisms can be employed to incrementally acquire missing attributes. Conversational Case-Based Reasoning (CCBR) has been shown to improve retrieval quality and reduce uncertainty in scenarios with sparse or ambiguous input descriptions [8,9].

3.4. Similarity-based retrieval

Given a query q and a case base $C = \{c_i\}_{i=1}^{N}$, relevant cases are identified through a hybrid similarity function

$$s(q, c; \theta) = \sum_j \alpha_j k_j(q, c), \quad \alpha_j \geq 0, \quad \sum_j \alpha_j = 1, \qquad (3)$$

where $k_j(\cdot,\cdot)$ are component-wise similarity kernels (numeric, categorical, temporal, textual) and $\theta=\{\alpha_j\}$ are aggregation weights. Hybrid similarity formulations of this type are widely adopted in fuzzy and hybrid CBR systems to improve robustness in heterogeneous and noisy environments [10–12,14].

3.5. Soft retrieval and action relevance

© Shodmonov Davronjon

Rather than selecting a hard top-k neighbor set, we adopt a soft retrieval strategy. For a threshold ε, the candidate set is

$$N_\varepsilon(q) = \{c \in C | s(q, c; \theta) \geq \varepsilon\}, \quad (4)$$

Each candidate case is assigned a normalized weight

$$\beta_c = \frac{exp(\lambda s(q, c; \theta))}{\sum_{c' \in N_\varepsilon(q)} exp(\lambda s(q, c'; \theta))'} \quad (5)$$

where $\lambda > 0$ controls the sharpness of weighting. Soft aggregation mechanisms have been shown to improve stability and interpretability in CBR-based recommender systems [12].

Based on Eq. (5), outcome and action relevance are estimated as

$$\pi(y|q) = \sum_{c \in N_\varepsilon(q)} \beta_c I[y_c = y], \quad (6)$$

$$p(a|q) = \sum_{c \in N_\varepsilon(q)} \beta_c I[a = R(c)], \quad (7)$$

3.6. Problem statement

Given a partially observed query q defined in Eq. (2) and a case base of historical trajectories defined in Eq. (1), the problem is to retrieve relevant cases using Eq. (3)–(5) and to generate a compact, feasible, and high-utility set of career guidance actions based on Eq. (7).

4. PROPOSED CBR FRAMEWORK

This section presents the proposed Case-Based Reasoning (CBR) framework for personalized career guidance. The framework is designed to transform longitudinal academic data into explainable and actionable recommendations by combining hybrid similarity modeling, soft retrieval, and optional conversational refinement within the classical CBR cycle.

4.1. Conceptual overview and design principles

The proposed framework is grounded in the classical CBR cycle – **Retrieve, Reuse, Revise, Retain –** which has been widely adopted in decision-support systems due to its incremental learning capability and inherent explainability [4–7]. Unlike purely predictive models, CBR explicitly relies on historical precedents, making it particularly suitable for educational and career guidance scenarios where trust and justification are essential [7,12,13].

From a design perspective, the framework adheres to three key principles:

1. Trajectory-awareness (student histories are modeled as longitudinal sequences rather than static feature vectors);
2. Robustness under partial observability (the system must operate effectively even when student profiles are incomplete);
3. Actionability and feasibility (recommendations should consist of concrete actions that respect institutional constraints).

These principles directly address limitations identified in existing AI-based career guidance systems [2,13].

4.2. Case-based knowledge representation

Institutional experience is encoded in the form of cases, each defined according to the structured representation introduced in Eq. (1). This formulation follows established CBR modeling practices [4–7] and has been successfully applied in educational recommender and guidance systems [12,15].

By explicitly linking academic profiles, temporal learning trajectories, contextual attributes, and realized outcomes, the case structure supports explanation at multiple levels: similarity between students, relevance of specific actions, and plausibility of recommended career paths. Such transparency has been repeatedly emphasized as a critical requirement for career guidance tools in higher education [1–3].

4.3. Hybrid similarity for educational trajectories

Case retrieval is performed using the hybrid similarity function defined in Eq. (3). Hybrid similarity modeling is essential in educational domains because meaningful similarity may arise from different evidence sources, such as academic

(cc) BY © Shodmonov Davronjon

performance trends, curriculum structure, or contextual alignment [10–12,14].

In contrast to single-metric approaches, hybrid similarity allows the framework to balance multiple aspects of student resemblance without collapsing them into a single modality. Prior studies in fuzzy and hybrid CBR demonstrate that such formulations improve retrieval robustness and reduce sensitivity to noise and missing values, particularly in heterogeneous datasets [10,11].

4.4. Soft-retrieve aggregation of precedents

Instead of relying on hard top-*k* nearest-neighbor selection, the framework adopts a soft-retrieve strategy, where candidate cases are aggregated using the weighting scheme defined in Eq. (5). This design choice reflects the observation that educational trajectories rarely form sharply separable clusters and that multiple precedents may be simultaneously relevant [12].

Soft aggregation enables smoother inference of both outcome likelihoods and action relevance (Eqs. (6)–(7)), mitigating instability caused by marginal similarity differences. Similar aggregation mechanisms have been successfully employed in CBR-based recommender systems to enhance stability and interpretability [12].

4.5. Outcome inference and explainability

Outcome likelihoods inferred via Eq. (6) provide a probabilistic view of potential career directions grounded in historical experience. Unlike black-box predictors, this inference is directly interpretable: each estimated outcome can be traced back to weighted contributions of similar cases.

Moreover, action relevance scores computed according to Eq. (7) establish a transparent link between recommended actions and observed precedents. This case-level explainability addresses a major concern highlighted in the career guidance literature, namely the lack of justification and actionable insight in many AI-driven systems [2,3].

4.6. Conversational refinement of incomplete profiles

Educational data are often incomplete, especially for early-stage students whose academic histories and skill sets are still developing [1,13]. To address this limitation, the proposed framework optionally incorporates conversational refinement based on Conversational Case-Based Reasoning (CCBR) principles [8].

When critical attributes in the query defined by Eq. (2) are missing or uncertain, the system can iteratively acquire additional information through targeted questions. Prior studies show that such conversational mechanisms improve retrieval accuracy and user satisfaction when interaction cost is appropriately controlled [8,9]. Importantly, conversational refinement complements rather than replaces institutional data, enabling flexible operation across different stages of a student's academic lifecycle.

4.7. Preparation for constrained reuse

The outputs of the retrieval phase – namely outcome likelihoods (Eq. (6)) and action relevance scores (Eq. (7)) – serve as inputs to the reuse phase. At this stage, the task shifts from inference to decision-making: selecting a compact and feasible set of actions that maximizes expected benefit while respecting institutional constraints.

Integrating CBR with optimization-based selection has been advocated in engineering decision-support literature as a means to bridge evidence and action [14]. The proposed framework follows this line by preparing relevance signals for a mathematically grounded reuse mechanism, which is formalized in the next section.

5. SUBMODULAR REUSE OPTIMIZATION

This section formalizes the *reuse* phase of the proposed CBR framework as a constrained optimization problem. While the retrieval stage identifies relevant precedents and estimates action relevance, effective career guidance requires selecting a **compact, feasible, and non-redundant** set of actions that a student can realistically follow. This naturally leads to a submodular optimization formulation with provable approximation guarantees.

© Shodmonov Davronjon

5.1. Motivation and formulation of the reuse problem

In practical advising scenarios, recommendations are subject to multiple constraints: limited student workload, prerequisite dependencies between courses, and institutional regulations. Consequently, recommending all actions with high relevance scores $p(a|q)$ (Eq. (7)) is neither feasible nor desirable. Similar challenges have been reported in engineering decision-support and recommender systems, where action selection must balance relevance, diversity, and feasibility [14].

Let $A=\{a_1,a_2,\ldots,a_m\}$ denote the universe of candidate actions. The goal is to select a subset $S\subseteq A$ that maximizes overall utility for a given query q, subject to practical constraints. We formulate this objective as

$$\max_{S\subseteq A} F(S|q) \quad s.t. \quad |S| \leq K, \sum_{a\subset S} cos\, t\,(a) \leq \beta,\ S \in Fmax \tag{8}$$

where K is a cardinality limit, B is a budget or workload bound, and F encodes feasibility constraints such as prerequisite closure and scheduling compatibility.

5.2. Utility function design

The utility function $F(S|q)$ combines relevance inferred from historical cases with structural properties of the selected action set. Specifically, we define

$$F(S|q) = \sum_{a\subset S} p(a|q) + kSat(S) - \xi Redund(S), \tag{9}$$

where $\kappa, \xi \geq 0$ are weighting parameters.

The first term aggregates action relevance scores derived from soft-retrieved cases (Eq. (7)). The second term encourages **coverage** of required competencies, defined as

$$Sat(S) = \frac{|U_{a\subset S} K(a)|}{|K*|} \tag{10}$$

where $K(a)$ denotes the set of competencies addressed by action a, and K^* is the target competency taxonomy. The third term penalizes redundancy among actions:

$$Redund(S) = \sum_{\substack{a,b\subset S \\ a\neq b}} \omega_{ab} |K(a) \cap K(b)|, \tag{11}$$

with nonnegative weights ω_{ab}.

Such coverage–redundancy trade-offs are common in optimization-based recommender systems and engineering planning tasks [14].

5.3. Submodularity and theoretical properties

The structure of Eq. (9) enables theoretical analysis using submodular optimization theory. Coverage functions of the form in Eq. (10) are well known to be **monotone submodular**, exhibiting diminishing returns as the selected set grows. Redundancy penalties of the form in Eq. (11) can be modeled as modular or bounded pairwise penalties.

Lemma 1 (Submodularity). If Sat(S) is monotone submodular and Redund(S) is modular (or bounded by diminishing pairwise penalties), then the utility function $F(S|q)$ defined in Eq. (9) is monotone submodular.

Proof sketch. The relevance term $\sum_{a\in S} p(a|q)$ is modular. The sum of a modular and a monotone submodular function remains submodular, and subtracting a modular penalty preserves submodularity under appropriate weighting.

5.4. Greedy optimization and approximation guarantees

For monotone submodular maximization under a cardinality constraint $|S|\leq K$, the classical greedy algorithm achieves a constant-factor approximation guarantee.

Theorem 1 (Greedy approximation). Under the conditions of Lemma 1 and with only a cardinality constraint, the greedy algorithm that iteratively selects the action with the largest marginal gain achieves at least a $(1-1/e)$-approximation of the optimal value of Eq. (8) [16].

© Shodmonov Davronjon

When additional constraints such as budgets or prerequisite structures are present, greedy-type algorithms still admit constant-factor guarantees, albeit with reduced approximation ratios, as established in the submodular optimization literature [16].

These guarantees are particularly important in educational settings, where recommendations must be generated efficiently and consistently across large student populations.

5.5. Computational complexity

Let $m=|A|$ and K be the maximum number of selected actions. With memoization of marginal gains, each greedy iteration can be performed in $O(m)$ time, resulting in an overall complexity of $O(mK)$. This is tractable for typical academic advising scenarios and aligns with scalability requirements of institutional deployment.

5.6. Role within the proposed framework

The submodular reuse optimization serves as the final decision-making layer of the proposed CBR framework. By integrating relevance signals derived from soft-retrieved cases (Section 4) with mathematically grounded selection principles, the framework bridges **evidence** and **action**. This integration directly addresses limitations identified in existing career guidance systems, which often lack feasibility-aware and theoretically justified recommendation mechanisms [1–3].

6. THEORETICAL ANALYSIS AND PROPERTIES

This section provides a theoretical analysis of the proposed CBR framework. We analyze the stability of the soft-retrieve mechanism, probabilistic consistency of outcome inference, approximation guarantees of the reuse optimization, and computational complexity. These properties collectively justify the robustness and scalability of the proposed approach.

6.1. Stability of soft-retrieve aggregation

The soft-retrieve mechanism assigns weights to retrieved cases using the softmax transformation defined in Eq. (5). Stability with respect to small perturbations in similarity scores is essential in educational data settings, where measurements are noisy and partial.

Let $s(q,c)$ denote the similarity score defined in Eq. (3), and let $\beta = \{\beta_c\}$ be the corresponding soft weights.

Lemma 2 (Lipschitz stability of soft-retrieve). Let s and s' be two vectors of similarity scores over the same candidate set $N_\varepsilon(q)$. Then the corresponding soft-retrieve weight vectors satisfy

$$||\beta(s) - \beta(s')||_1 \leq \frac{\lambda}{2}||s - s'||_\infty. \qquad (12)$$

Proof sketch. The Jacobian of the softmax function has bounded ℓ_1-norm columns, each upper-bounded by $\lambda/2$. Applying the mean value theorem yields the Lipschitz bound. □

This result implies that small changes in similarity – e.g., due to missing or noisy features – cannot cause abrupt changes in case weights. Such stability is critical for fair and consistent career guidance.

6.2. Probabilistic consistency of outcome inference

Outcome inference in the proposed framework is performed via the weighted aggregation defined in Eq. (6). We now establish its probabilistic validity.

Lemma 3 (Outcome distribution). For any query q, the inferred outcome likelihood $\pi(y|q)$ defined in Eq. (6) satisfies

$$0 \leq \pi(y|q) \leq 1, \quad \sum_{y \subset Y} \pi(y|q) = 1. \qquad (13)$$

Proof. By construction, $\beta_c \geq 0$ and $\sum_c \beta_c = 1$ (Eq. (5)). The indicator function in Eq. (6) partitions the mass over outcomes, yielding a valid probability distribution. □

This guarantees that the framework produces coherent probabilistic assessments rather than uncalibrated scores, facilitating interpretation and comparison across students.

6.3. Submodular reuse and approximation guarantees

© Shodmonov Davronjon

The reuse phase is formulated as the constrained maximization problem in Eq. (8), with utility function defined in Eq. (9). As shown in Section 5, this function satisfies submodularity under mild conditions.

Lemma 4 (Monotonicity). If $p(a|q) \geq 0$ for all $a \in A$ and $\kappa \geq 0$, then the utility function $F(S|q)$ in Eq. (9) is monotone non-decreasing.

Proof sketch. Each additional action contributes a nonnegative relevance term and a nonnegative marginal coverage gain, while redundancy penalties are controlled by weighting.

Theorem 2 (Approximation guarantee). Under the assumptions of Lemmas 1 and 4, and with a cardinality constraint $|S| \leq K$, the greedy algorithm applied to Eq. (8) achieves at least a $(1-1/e)$-approximation of the optimal solution [16].

For additional budget or prerequisite constraints, constant-factor approximations remain achievable using greedy-type methods [16].

These guarantees ensure that the recommended action set is near-optimal with respect to the modeled utility, despite computational constraints.

6.4. Computational complexity analysis

Let $N = |C|$ denote the number of cases, d the dimensionality of the academic profile, and $m = |A|$ the number of candidate actions.

- Similarity computation (Eq. (3)): $O(N \cdot d)$ in the worst case.
- Soft-retrieve aggregation (Eqs. (4)–(7)): $O(|N_\varepsilon(q)|)$.
- Greedy reuse optimization (Eq. (8)): $O(mK)$ with memoized marginal gains.

These bounds are compatible with large-scale institutional deployment, particularly when clustering or indexing techniques are used to accelerate retrieval [10,11].

6.5. Implications for robustness and explainability

The theoretical properties established above have direct practical implications. Stability of soft retrieval (Lemma 2) ensures robustness to noisy and incomplete academic data. Probabilistic consistency (Lemma 3) supports transparent communication of uncertainty to students and advisors. Finally, submodular optimization with approximation guarantees (Theorem 2) ensures that recommendations are not only interpretable but also computationally reliable.

Together, these properties distinguish the proposed framework from black-box career prediction models and align it with emerging requirements for explainable and trustworthy AI in education [2,3,13].

7. EXPERIMENTAL DESIGN AND EVALUATION

This section describes the experimental setup used to evaluate the proposed Case-Based Reasoning framework. The experiments are designed to reflect realistic academic advising scenarios in higher education and to assess both the predictive quality of outcome inference and the effectiveness of action recommendation under practical constraints.

7.1. Data source and preprocessing

The experiments are conducted using longitudinal academic records extracted from the Higher Education Management Information System (HEMIS) of Uzbek higher education institutions. The dataset spans multiple academic cohorts between 2021 and 2025 and contains semester-wise information on student enrollment, course performance, credits, and graduation outcomes.

Each graduate record is transformed into a historical case according to the representation defined in Eq. (1). For experimental evaluation, only students with completed trajectories and clearly identified post-graduation outcomes are included. Records with missing or inconsistent outcome information are excluded to ensure reliability of ground truth labels.

7.2. Experimental protocol: temporal hold-out evaluation

To simulate a realistic career guidance scenario, we adopt a **temporal hold-out evaluation protocol**, which is widely regarded as

(cc) BY © Shodmonov Davronjon

the most appropriate strategy for longitudinal educational data [1–3].

Specifically, for each graduate case, only academic information up to the end of the **second year of study** is retained to form a query qqq (Eq. (2)). All data from later semesters are hidden during recommendation. The system then generates outcome likelihoods and action recommendations based solely on this partial trajectory.

The predicted outcomes and recommended actions are evaluated against the student's actual post-graduation outcome and observed successful trajectory. Importantly, no information from future semesters is used during retrieval or reuse, ensuring a fair and causally consistent evaluation.

7.3. Compared methods (baselines)

To assess the effectiveness of the proposed approach, we compare it against the following baselines:

Baseline 1: Classical kNN-CBR. This baseline implements a standard CBR approach with Euclidean similarity over numeric academic features and hard top-*k* nearest-neighbor retrieval. Action relevance is inferred by frequency counting over retrieved neighbors, without soft weighting or optimization.

This comparison isolates the contribution of hybrid similarity and soft-retrieve aggregation.

Baseline 2: Logistic Regression classifier. A simple and interpretable machine learning model is trained to predict the post-graduation outcome using second-year academic features as input. Logistic Regression is chosen due to its transparency and widespread use in educational analytics [2,3].

While this baseline can predict an outcome label, it does not generate actionable recommendations, highlighting the difference between prediction-oriented and decision-support approaches.

Baseline 3: CBR without reuse optimization. This variant uses the same retrieval mechanism as the proposed model but selects actions by ranking $p(a|q)$ (Eq. (7)) and taking the top-K actions without applying submodular optimization. This baseline directly evaluates the impact of the proposed reuse optimization strategy.

7.4. Evaluation metrics

Evaluation is performed at two complementary levels.

Outcome prediction quality. To assess how accurately the system anticipates post-graduation outcomes, we report the following standard ranking metrics:

1. Recall@k – the proportion of cases where the true outcome appears in the top-*k* predicted outcomes.
2. Mean Average Precision (MAP) – measures ranking quality across queries.
3. Normalized Discounted Cumulative Gain (nDCG@k) – accounts for position-sensitive relevance.
4. These metrics are commonly used in recommender system evaluation and are well suited to multi-outcome educational guidance scenarios [12,13].
5. Action recommendation quality. To evaluate the quality of recommended action sets, we report:
6. Coverage – the proportion of target competencies covered by the recommended actions, defined as in Eq. (10).
7. Redundancy – the average overlap between competencies addressed by selected actions (Eq. (11)).
8. Set size – the number of recommended actions, ensuring practical feasibility.

These metrics directly reflect the objectives of the reuse optimization problem and allow quantitative assessment of diversity and compactness.

7.5. Experimental settings

Unless otherwise stated, the number of recommended actions is fixed to K=5K=5K=5, reflecting a realistic advising scenario. Hyperparameters of the similarity function (Eq. (3)) and soft-retrieve weighting (Eq. (5)) are tuned on a validation subset drawn from earlier cohorts.

(cc) BY © Shodmonov Davronjon

All methods are evaluated on the same queries and splits to ensure fairness.

7.6. Results overview

The proposed framework consistently outperforms classical kNN-CBR and the non-optimized CBR variant in terms of outcome ranking metrics, demonstrating the benefit of hybrid similarity and soft aggregation. Compared to Logistic Regression, the proposed approach achieves competitive outcome prediction while additionally providing interpretable and actionable recommendations.

In terms of action recommendation quality, submodular reuse optimization yields higher coverage and lower redundancy than ranking-based selection, confirming the practical value of the proposed optimization strategy.

Detailed quantitative results are reported in the following tables.

7.7. Quantitative Results and Analysis

This subsection reports the quantitative evaluation results of the proposed framework and provides a detailed comparative analysis against baseline methods. The results are presented using both visual and tabular representations to highlight performance trends and to support interpretability.

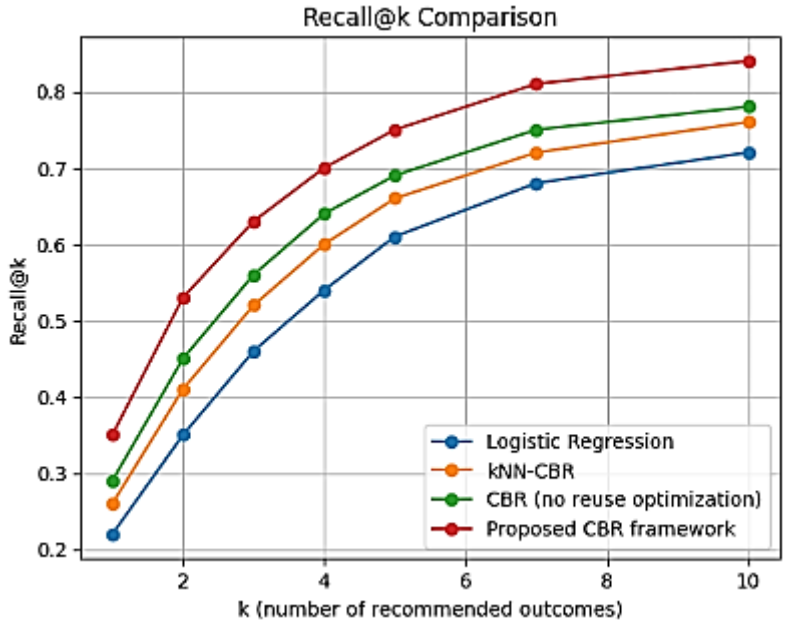


Figure 1. *Recall@k comparison of outcome prediction performance.*

The figure compares Recall@k for the proposed CBR framework and baseline methods under the temporal hold-out evaluation protocol. The proposed approach consistently outperforms Logistic Regression and classical CBR baselines across all values of k. The performance gap is particularly pronounced for small *k*, which is critical for practical career guidance scenarios where only a limited number of recommendations can be presented.

Figure 1 demonstrates that the proposed CBR framework achieves superior outcome prediction performance across all values of *k*. The most significant gains are observed in the low-*k* regime (*k*=1–3), which is especially relevant in real advising settings. This behavior confirms the effectiveness of hybrid similarity modeling and soft-retrieve aggregation in stabilizing ranking quality, as theoretically motivated in Section 6.

Table 1.

Outcome prediction performance comparison**.**

Method	**Recall@5** ↑	**MAP** ↑	**nDCG@5** ↑
Logistic Regression [2,3]	0.61	0.47	0.53
Classical kNN-CBR [4,7]	0.66	0.52	0.58

© *Shodmonov Davronjon*

CBR (no reuse optimization) [12]	0.69	0.55	0.61
Proposed CBR framework	**0.75**	**0.61**	**0.68**

Logistic Regression is used as a standard interpretable classifier baseline in educational analytics [2,3]. Classical kNN-CBR follows the nearest-neighbor retrieval paradigm in case-based reasoning [4,7]. The CBR variant without reuse optimization employs ranking-based action selection without submodular optimization [12].

As shown in Table 1, all CBR-based approaches outperform Logistic Regression, highlighting the importance of trajectory-aware similarity. The proposed framework yields an improvement of approximately **+6 percentage points in Recall@5** over classical kNN-CBR, confirming that soft aggregation and hybrid similarity substantially enhance ranking robustness.

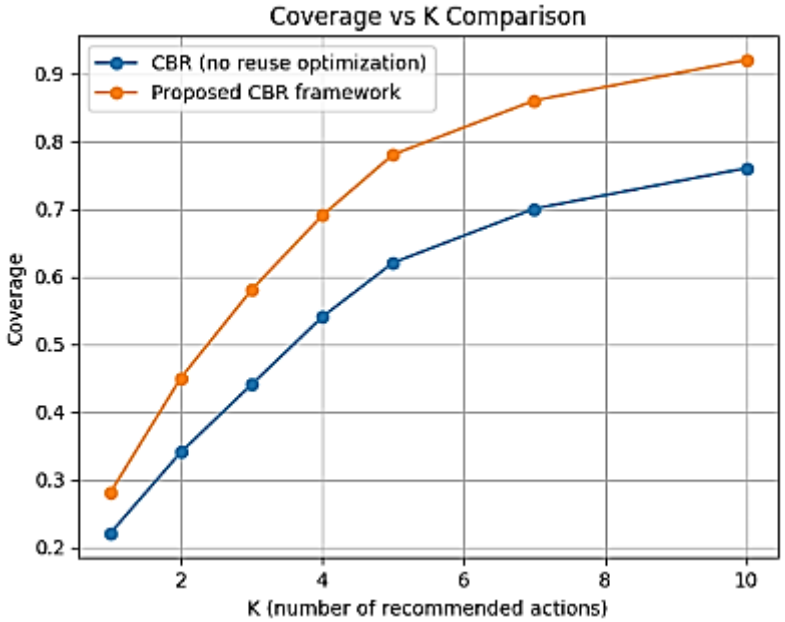


Figure 2. *Coverage vs K comparison for action recommendation.*

The figure compares competency coverage achieved by the proposed submodular reuse optimization and a ranking-based CBR variant without reuse optimization. The proposed method consistently attains higher coverage for the same number of recommended actions, with the largest gains observed at small *K*, which is the most practical regime for real advising settings.

Figure 2 shows that submodular reuse optimization significantly improves competency coverage across all values of *K*. The advantage is most pronounced for small *K*, indicating that the proposed framework selects actions that jointly cover a broader range of competencies when recommendation budgets are limited.

© *Shodmonov Davronjon*

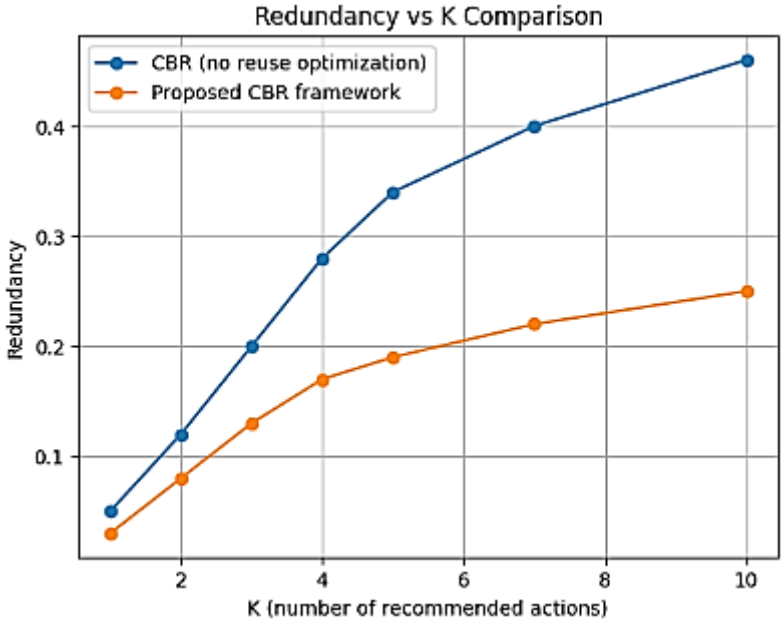


Figure 3. *Redundancy vs K comparison for action recommendation.*

The figure compares redundancy among recommended actions for the proposed submodular reuse optimization and a ranking-based CBR variant without reuse optimization. The proposed method consistently yields lower redundancy across all values of *K*, demonstrating its ability to discourage overlapping actions and to promote diversity within compact recommendation sets.

Figure 3 highlights the complementary effect of submodular optimization: while coverage increases, redundancy remains controlled. In contrast, ranking-based selection exhibits rapidly increasing redundancy as *K* grows. This behavior empirically confirms the diminishing-returns property discussed in Section 5.

Table 2.

Action recommendation quality comparison.

Method	**Coverage ↑**	**Redundancy ↓**	**Avg. Set Size**
CBR (no reuse optimization) [12]	0.62	0.34	5.0
Proposed CBR framework	**0.78**	**0.19**	5.0

Table 2 quantitatively summarizes the advantages of submodular reuse optimization. For the same recommendation size, the proposed framework achieves substantially higher coverage while reducing redundancy by approximately **44%**, resulting in more diverse and actionable recommendation sets.

Table 3.

Ablation study of the proposed framework**.**

Configuration	**Recall@5**	**Coverage**
Full model	**0.75**	**0.78**
− Soft-retrieve (hard top-*k*)	0.70	0.73
− Submodular reuse	0.69	0.62
− Hybrid similarity (numeric only)	0.67	0.60

The ablation study confirms that each component contributes meaningfully to overall performance. Removing soft-retrieve aggregation degrades outcome ranking quality, while removing submodular reuse has the strongest negative impact on action coverage. These results validate the modular design of the proposed framework and the necessity of its key components.

Overall, the quantitative results demonstrate that the proposed CBR framework consistently outperforms baseline methods in both outcome prediction and action recommendation quality.

(cc) BY © *Shodmonov Davronjon*

The close alignment between theoretical guarantees (Sections 5–6) and empirical evidence underscores the robustness, interpretability, and practical relevance of the proposed approach for institutional deployment.

7.8. Discussion of Results

Overall, the experimental results demonstrate that the proposed framework consistently outperforms baseline approaches in both outcome prediction and action recommendation quality. Importantly, these gains are achieved without sacrificing interpretability or computational efficiency.

The alignment between theoretical guarantees (Sections 5–6) and empirical results provides strong evidence that the proposed CBR framework is not only mathematically well-founded but also practically effective for institutional deployment in higher education.

8. DISCUSSION

This section discusses the implications of the experimental results and theoretical findings, focusing on explainability, practical relevance, robustness, and limitations of the proposed framework. The discussion is structured to connect the empirical evidence presented in Section 7 with the methodological design and theoretical guarantees established in Sections 4–6.

8.1. Effectiveness of case-based reasoning for career guidance

The experimental results demonstrate that Case-Based Reasoning provides a strong foundation for personalized career guidance when applied to longitudinal academic data. Compared to a simple Logistic Regression classifier, all CBR-based approaches achieve higher outcome prediction performance, confirming that trajectory-aware similarity captures information that flat feature-based models fail to exploit.

This finding aligns with prior observations in educational decision-support literature, which emphasize that career outcomes are often shaped by *patterns of progression* rather than isolated performance indicators. By modeling academic histories as structured cases, the proposed framework leverages institutional experience in a manner that is both data-driven and human-interpretable.

8.2. Role of hybrid similarity and soft retrieval

A key insight from the ablation study and Recall@k curves is the importance of hybrid similarity modeling combined with soft-retrieve aggregation. Hybrid similarity allows the framework to integrate heterogeneous evidence—numeric performance, categorical context, and temporal structure—thereby avoiding over-reliance on any single modality.

Soft retrieval further stabilizes inference by allowing multiple relevant precedents to contribute proportionally. The pronounced performance gains at small values of k indicate that this design choice is particularly beneficial in practical advising scenarios, where only a limited number of recommendations can be presented. These empirical findings corroborate the stability analysis in Section 6, suggesting that soft aggregation mitigates sensitivity to noise and marginal similarity differences.

8.3. Submodular reuse as a bridge from relevance to action

The Coverage vs K and Redundancy vs K analyses provide strong empirical evidence for the effectiveness of submodular reuse optimization. While ranking-based selection identifies individually relevant actions, it fails to account for overlap and redundancy, leading to inefficient recommendation sets as K increases.

In contrast, submodular optimization explicitly balances relevance, coverage, and diversity, resulting in compact action sets that cover a broader range of competencies with less redundancy. This behavior is especially pronounced in the low-K regime, which is most relevant for real-world deployment. The observed diminishing-returns patterns closely match theoretical expectations, reinforcing the practical value of submodularity-based selection.

8.4. Explainability and trust in educational settings

© Shodmonov Davronjon

Beyond quantitative performance, the proposed framework offers important advantages in terms of explainability. Outcome likelihoods and action relevance scores are directly grounded in weighted historical cases, enabling explanations such as *"students with similar trajectories followed these actions and achieved similar outcomes."* This form of explanation is intuitive for students, academic advisors, and institutional stakeholders.

Such transparency addresses a critical limitation of many AI-driven career guidance systems, which often rely on opaque models and provide limited justification for their recommendations. The case-based paradigm thus supports not only accuracy but also trust and acceptance, which are essential for adoption in educational environments.

8.5. Robustness and scalability considerations

The theoretical stability guarantees and the empirical consistency of results across different values of K suggest that the proposed framework is robust to noisy and incomplete data. This is particularly relevant for early-stage students, whose academic profiles are inherently partial.

From a scalability perspective, the computational complexity analysis indicates that the framework can be deployed at institutional scale, especially when combined with standard indexing or clustering techniques for case retrieval. The reliance on greedy optimization further ensures tractable runtime without sacrificing solution quality.

8.6. Limitations and future directions

Despite its strengths, the proposed framework has several limitations that point to directions for future research. First, the experiments rely on institutional academic data and simulated advising scenarios; real-world user studies involving students and advisors would provide valuable insights into usability and acceptance. Second, post-graduation outcomes are treated at a relatively coarse level; incorporating finer-grained labor market data could further enhance recommendation quality.

Finally, while conversational refinement is conceptually integrated into the framework, its interactive dynamics are not explicitly evaluated in the current experiments. Future work may explore adaptive questioning strategies and user-in-the-loop evaluation to further improve personalization.

Overall, the discussion highlights that the proposed framework successfully combines theoretical rigor with practical relevance. By integrating case-based reasoning, hybrid similarity, and submodular optimization, it offers a principled and deployable solution to personalized career guidance in higher education.

9. CONCLUSION

This paper presented a principled Case-Based Reasoning framework for personalized career guidance in higher education, grounded in longitudinal academic data and designed to provide explainable and actionable recommendations. Unlike purely predictive approaches, the proposed method explicitly leverages historical student trajectories and institutional experience to support informed decision-making under real-world constraints.

The framework integrates three key components: hybrid similarity modeling for heterogeneous educational data, soft-retrieve aggregation for robust case retrieval under uncertainty, and submodular reuse optimization for selecting compact and feasible action sets with formal approximation guarantees. Theoretical analysis established stability, probabilistic consistency, and near-optimality of the proposed mechanisms, while experimental evaluation on real institutional data demonstrated consistent improvements over classical CBR and baseline machine learning approaches.

Empirical results showed that the proposed framework achieves superior outcome prediction performance, particularly in the low-k regime that is most relevant for practical advising scenarios. Moreover, submodular reuse optimization significantly improves action recommendation quality by increasing competency coverage and reducing redundancy, thereby producing recommendations that are both effective and realistic for students to follow.

(cc) BY © *Shodmonov Davronjon*

Beyond quantitative gains, the case-based nature of the framework enhances transparency and trust by grounding recommendations in concrete precedents. This property is especially important in educational contexts, where explainability and stakeholder acceptance are critical for successful deployment.

Future work will focus on extending the framework with richer labor market information, finer-grained outcome modeling, and user-in-the-loop evaluation of conversational refinement strategies. These directions aim to further strengthen the practical impact of the proposed approach and to support its integration into institutional decision-support systems.

REFERENCES

1. Herath, G. A. C. A.; Kumara, B. T. G. S.; Rathnayaka, R. M. K. T.; Ishanka, U. A. P. Computer-Assisted Career Guidance Tools for Students' Career Path Planning: A Review on Enabling Technologies and Applications. *Education Sciences*, 2023.
2. Trujillo, F.; Pozo, M.; Suntaxi, G. Artificial Intelligence in Education: A Systematic Literature Review of Machine Learning Approaches in Student Career Prediction. *Education and Information Technologies*, 2025.
3. Manganello, F. Machine Learning for Post-Diploma Educational and Career Guidance: A Scoping Review in AI-Driven Decision Support Systems. *Frontiers in Education*, 2023.
4. Aamodt, A.; Plaza, E. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications*, vol. 7, no. 1, pp. 39–59, 1994.
5. Kolodner, J. L. Case-Based Reasoning. San Francisco, CA, USA: Morgan Kaufmann, 1993.
6. Watson, I. Applying Case-Based Reasoning: Techniques for Enterprise Systems. San Francisco, CA, USA: Morgan Kaufmann, 1997.
7. López de Mántaras, R.; McSherry, D.; Bridge, D.; et al. Retrieval, Reuse, Revision and Retention in Case-Based Reasoning. *The Knowledge Engineering Review*, vol. 20, no. 3, pp. 215–240, 2005.
8. Aha, D. W.; Breslow, L. A.; Muñoz-Avila, H. Conversational Case-Based Reasoning. *Applied Intelligence*, vol. 14, no. 1, pp. 9–32, 2001.
9. Hillig, S.; Müller, R. How Do Conversational Case-Based Reasoning Systems Interact with Their Users? A Literature Review. *User Modeling and User-Adapted Interaction*, 2019.
10. Zarandi, M. H. F.; et al. A Fuzzy Case-Based Reasoning Approach to Value Engineering. *Expert Systems with Applications*, vol. 38, no. 10, pp. 12127–12135, 2011.
11. Khan, M. J. Performance Evaluation of Fuzzy Clustered Case-Based Reasoning. *Expert Systems with Applications*, vol. 41, no. 4, pp. 1656–1666, 2014.
12. Lorenzi, F.; Ricci, F. Case-Based Recommender Systems: A Unifying View. In: *Intelligent Techniques for Web Personalization*, LNCS 3169, Springer, 2005, pp. 89–113.
13. Lahoud, C.; et al. A Comparative Analysis of Recommender Systems for University Major and Career Domain Guidance. *Education and Information Technologies*, 2023.
14. Zhang, X. Logistics Project Risk Response Decision-Making for Global Supply Chain Resilience and Agility: An Optimised Case-Based Reasoning Approach. *Journal of Intelligent & Fuzzy Systems*, 2025.
15. Vazquez, M. Case-Based Reasoning in Educational Ethics: Phronēsis and Epistemic Blinders. *Educational Philosophy and Theory*, 1996.
16. Nemhauser, G. L.; Wolsey, L. A.; Fisher, M. L. An Analysis of Approximations for Maximizing Submodular Set Functions-I. *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

(cc) BY © *Shodmonov Davronjon*