Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Student Profiling and Resource Tagging Using Machine Learning for Adaptive Learning Systems

Anas El Moustamid[1], Jaber El Bouhdidi[2]
[1]*SIGL Laboratory, National School of Applied Sciences*
[2] *SIGL Laboratory, National School of Applied Sciences*

ARTICLE INFO

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Personalized learning has become a central focus in modern education, aiming to adapt content and support to individual learner needs. This study investigates how students can be grouped according to their learning characteristics, and how educational materials can be classified based on their instructional value

The objective is to enhance recommendation systems in education by providing learners with more relevant resources, while taking into account both their learning styles and the nature of the content.

To support this goal, real-world data was collected from a private secondary school, involving students from different grade levels. The study compares various classification strategies to determine which are most effective in supporting adaptive learning environments.

The findings contribute to ongoing efforts to make learning more targeted, efficient, and student-centered by leveraging data-driven approaches to profile learners and educational content.

Keywords: Recommender Systems, Personalized learning, Collaborative filtering, Content-based filtering, Hybrid Approach, KNN, SVM, Accuracy.

INTRODUCTION

With the rise of digital learning platforms, the ability to adapt educational content to learner's individual needs have become a central challenge in educational engineering. This personalization relies on leveraging data from interactions between students and learning resources. In this context, supervised classification algorithms play a strategic role by enabling both the profiling of students based on their behaviors and performance, and the categorization of educational content according to its intrinsic characteristics.

Given the diversity of existing techniques, it has become essential to conduct comparative evaluations of these algorithms to guide their integration into intelligent recommendation systems. This study aligns with that objective by examining several classical and scalable machine learning approaches, while accounting for the specificities of both collaborative filtering (Bobadilla, Serradilla, and Hernando 2009; Ekstrand, Riedl, and Konstan 2011; Su and Khoshgoftaar 2009; Tran et al. 2024) and content-based filtering (Javed et al. 2021) paradigms.

Content-based filtering

Content-Based Filtering, also known as cognitive filtering, recommends educational resources by comparing their attributes with those of a learner's profile(Tran et al. 2024). Each resource (video, exercise, text, etc.) is described using pedagogical descriptors such as topic, level, or targeted skills. The learner's profile is based on previously accessed or appreciated content (Nilla and Setiawan 2024), (Tian 2024).

The system then suggests resources similar to those the learner has mastered. This approach does not rely on other users' data and quickly adapts to evolving preferences or skill levels.

Collaborative filtering

Collaborative filtering is an educational recommendation method that analyzes the behaviors and preferences of student groups to suggest relevant resources to individual learners (Ekstrand et al. 2011), (Schafer et al. 2007) , (Bobadilla, Serradilla, and Bernal 2010). Unlike content-based filtering, it emphasizes user similarities over resource characteristics.

The core idea: if two students show similar learning behaviors—such as liking the same videos, completing similar exercises, or engaging with related content—resources used successfully by one are recommended to the other.

Recommendation System based on hybrid approach

The hybrid approach in recommendation systems represents a sophisticated strategy that combines the strengths of multiple types of recommendation engines, most commonly collaborative filtering and content-based filtering (Hazar et al. 2025). Rather than relying solely on the past interactions of similar users or the intrinsic characteristics of items, a hybrid model leverages both. This synergy helps overcome the inherent limitations of each isolated approach: the "cold start problem" (Hazar et al. 2025)(when there isn't enough data for new users or new resources) is mitigated by content analysis, while the "limited diversity" of content-based recommendations is enriched by the discovery of unexpected items via collaborative methods. By integrating contextual information such as a detailed user profile, learning objectives, or resource prerequisites, the hybrid approach significantly refines the relevance of suggestions, thereby offering a robust, rich, and dynamic personalization experience, which is essential for complex domains like education.

OBJECTIVES

The purpose of our work is to propose a rigorous comparison of several supervised classification algorithms: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, and Decision Tree. This is to identify the algorithm that demonstrates the best performance. The chosen algorithm will be integrated into a hybrid recommendation system designed to suggest personalized educational resources based on each learner's preferences, specific needs, and performance level.

This paper is structured as follows: Section 2 reviews the state of the art, Section 3 describes the methodology of our approach, Section 4 presents and discusses the obtained results, and Section 5 concludes the study while outlining future research directions.

Related works

In (Al-Kindi and Al-Khanjari 2022), They investigated two classification algorithms: Naive Bayes (NB) and Random Forest (RF). Their classification performance was evaluated using Weka, an open-source software that offers tools for data preparation, algorithm implementation, and visualization. The comparison results revealed that Random Forest achieved the highest accuracy, correctly predicting 97.36% of instances. They suggested that these classification techniques could be applied within a Moodle environment to forecast student performance.

In (Aher and Lobo 2012), the authors compared five classification algorithms with the aim of identifying the most effective one for integration into a course recommendation system. The algorithms studied are: ADTree, Simple CART, J48, ZeroR, and Naïve Bayes. The evaluation was conducted using the open-source data mining tool WEKA. The results obtained show that the ADTree algorithm offers the best classification performance, thus outperforming the other methods analyzed within the framework of this recommendation system. The dataset used is extracted from the Moodle platform, comprising 45 students and 15 courses. Although this dataset allows for an initial comparative evaluation of classification algorithms, its size does not permit a full generalization of the results obtained.

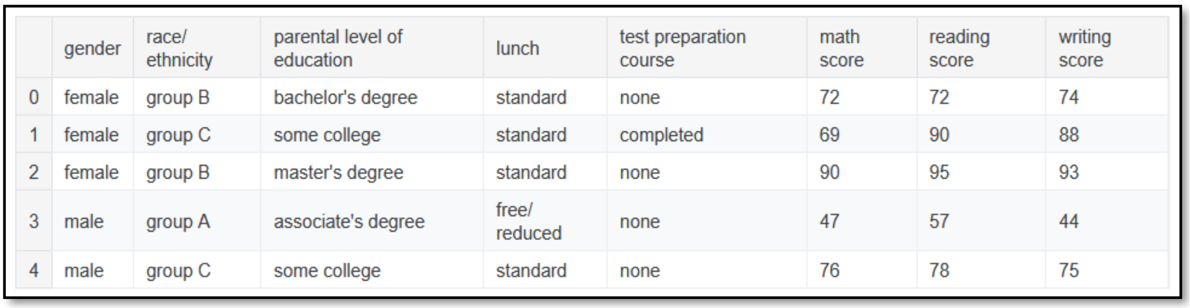
METHODS

Dataset

While many studies use the MovieLens dataset—despite its widespread use, it often falls short in representing real-world educational interactions due to its lack of diversity—our research stands apart. We've based our work on two datasets: one sourced from Kaggle and a second from an actual learning environment.

Kaggle Dataset

The Kaggle StudentsPerformance dataset comprises 1,000 student records, each described by eight variables. Following preprocessing, transformation, and feature extraction steps, we trained and validated four algorithms: k-nearest neighbors (KNN), decision tree, random forest, and support vector machine (SVM). The results demonstrate the strong performance of our model, with accuracy rates of 87% for KNN, 85% for the decision tree, 86% for the random forest, and 86% for SVM



	gender	race/ ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/ reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

Figure 1 : Dataset Kaggle StudentsPerformance

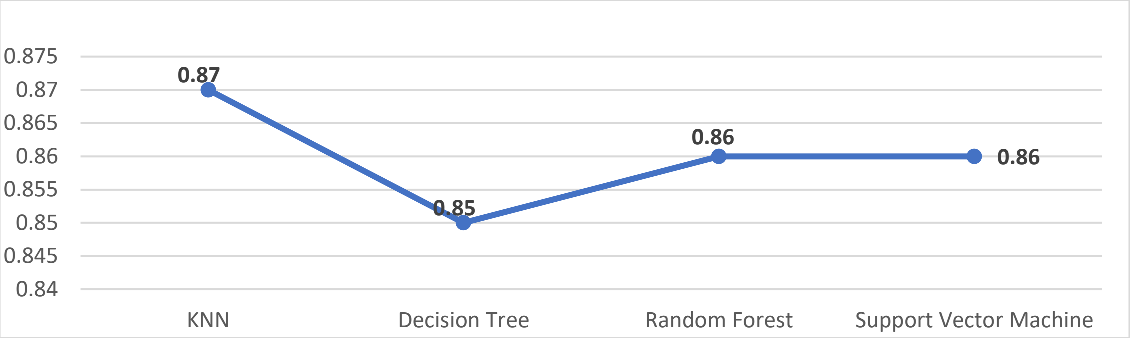
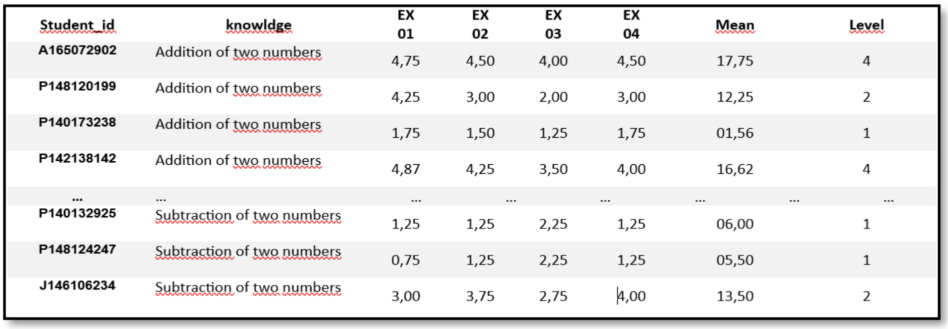


Figure 2: Accuracy KNN, Decision Tree, Random Forest and SVM

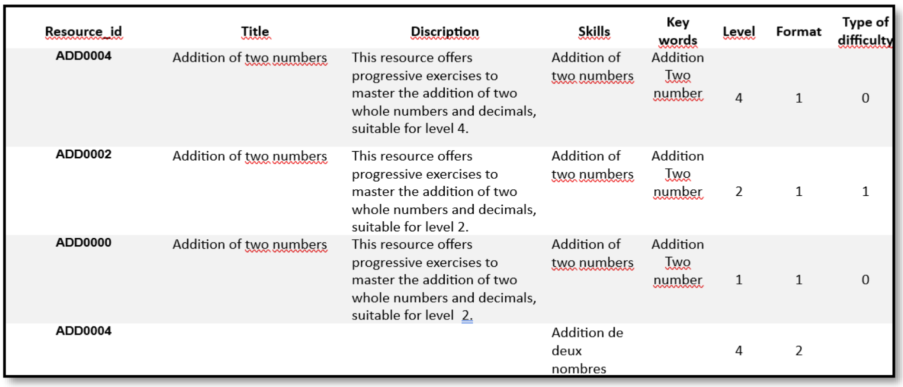
Real-world Dataset

The dataset stems from a real-world experiment conducted in a private school, ensuring authenticity, relevance, and direct applicability to educational settings. It encompasses responses from 233 students across multiple grade levels, specifically 98 in 1st grade, 77 in 2nd grade, and 58 in 3rd grade, covering an age range of 12 to 15 years old. This real-world composition allows for a better understanding of student behaviors, engagement patterns, and learning preferences, making the dataset far more representative of educational contexts compared to standardized datasets built primarily on artificially structured interactions. By capturing genuine student responses, our dataset provides valuable insights into how learners of different grades engage with educational content, paving the way for more effective and adaptable recommendation systems tailored to real student needs rather than theoretical assumptions.



Student_id	knowldge	EX 01	EX 02	EX 03	EX 04	Mean	Level
A165072902	Addition of two numbers	4,75	4,50	4,00	4,50	17,75	4
P148120199	Addition of two numbers	4,25	3,00	2,00	3,00	12,25	2
P140173238	Addition of two numbers	1,75	1,50	1,25	1,75	01,56	1
P142138142	Addition of two numbers	4,87	4,25	3,50	4,00	16,62	4
...	...	...	...	...	...	...	...
P140132925	Subtraction of two numbers	1,25	1,25	2,25	1,25	06,00	1
P148124247	Subtraction of two numbers	0,75	1,25	2,25	1,25	05,50	1
J146106234	Subtraction of two numbers	3,00	3,75	2,75	4,00	13,50	2

Figure 3 : The samples of each dataset represent the student's performance



Resource_id	Title	Discription	Skills	Key words	Level	Format	Type of difficulty
ADD0004	Addition of two numbers	This resource offers progressive exercises to master the addition of two whole numbers and decimals, suitable for level 4.	Addition of two numbers	Addition Two number	4	1	0
ADD0002	Addition of two numbers	This resource offers progressive exercises to master the addition of two whole numbers and decimals, suitable for level 2.	Addition of two numbers	Addition Two number	2	1	1
ADD0000	Addition of two numbers	This resource offers progressive exercises to master the addition of two whole numbers and decimals, suitable for level 2.	Addition of two numbers	Addition Two number	1	1	0
ADD0004			Addition de deux nombres		4	2	

Figure 4 : Representation of educational resources

After collecting the data, we cleaned it by handling missing values, correcting errors, and removing duplicates to ensure the quality of our dataset. Next, we performed exploratory data analysis (EDA) to identify trends, outliers, and correlations between variables. Then, we carried out preprocessing, which included normalization or standardization of the data, as well as dimensionality reduction (PCA) or selecting the most relevant features (feature selection). Finally, we split the data into training and test sets to evaluate our model's performance.

Data Preprocessing

Data preprocessing is a foundational component in the development of machine learning systems, particularly in educational domains where data heterogeneity is common. Real-world datasets often contain missing entries, inconsistencies, and noisy variables that, if left untreated, compromise model reliability (Arya 2025). To mitigate these challenges, robust preprocessing techniques—such as:

- **Data cleaning**: Handle missing quiz scores, duplicated student records, or noisy log entries.
- **Normalization**: Align metrics like time-on-platform or completion rates to a consistent scale.
- **Imputation**: Fill gaps in interaction data using cohort-based averages or KNN-based strategies.
- **Encoding**: Convert categorical attributes like “course level” or “resource type” into numerical formats.

are employed to transform raw data into a structured format suitable for model training (Osman 2025). These steps not only enhance algorithmic performance but also contribute to the interpretability and reproducibility of predictive models in educational research (Mohd 2025)

Data Transformation

Data transformation aims to make variables compatible and relevant to learning models. It includes, standardizing numerical values, encoding categorical variables, and restructuring temporal dimensions. These operations improve the quality of inputs and facilitate the detection of exploitable learning patterns.

- **Standardization**: Rescale time-on-task, quiz scores.
- **Normalization**: Align engagement metrics.
- **Temporal Structuring**: Extract session durations, study periods.

Feature Extraction

Feature extraction, generating interaction-based indicators (e.g., access frequency, average dwell time), contextual markers (e.g., session hour, grade level), and pedagogically relevant variables (e.g., skill level, topic domain). These steps collectively produced a robust and informative dataset tailored for machine learning tasks such as student profiling and resource classification.

Algorithms Metrics:

To validate the performance of our model, we used several evaluation metrics to measure its robustness and generalization capability. Specifically, we employed Accuracy (1), Recall (2) and F1-score (3):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1 - score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (3)$$

K-Nearest Neighbors

Classification algorithms are essential tools for categorizing learners based on their characteristics, behaviors, preferences, and performance (El Moustamid, El Mokhtar, and Bouhdidi 2017). These algorithms enable the personalization of the learning experience by identifying specific profiles and adapting educational resources accordingly.

The K-Nearest Neighbors (KNN) algorithm, is a non-parametric classification method that relies on the proximity of data points in the feature space. To classify a new instance, KNN identifies the k closest instances in the training set and assigns the majority class among these neighbors(Debnath, Sinha, and Bhowmik 2022). The distance between points is typically calculated using the Euclidean distance(Debnath et al. 2022) (1).

$$d(x,y) = \sqrt{\sum_{i=1}^{n} \pi (x_i - y_i)^2} \qquad (1)$$

The choice of K in the K-Nearest Neighbors (KNN) algorithm is a crucial step that directly influences the model's performance. A good practice is to test different K values using cross-validation, such as K-fold cross-validation, to identify the one that maximizes metrics like precision, recall, and the F1-score, while maintaining a good balance between bias and variance. In our case, we tested the model with K values of 5, 10, 15, 20, 25, and 30, and found that K=15 yields the best performance with an accuracy of 0.95 Figure 5.

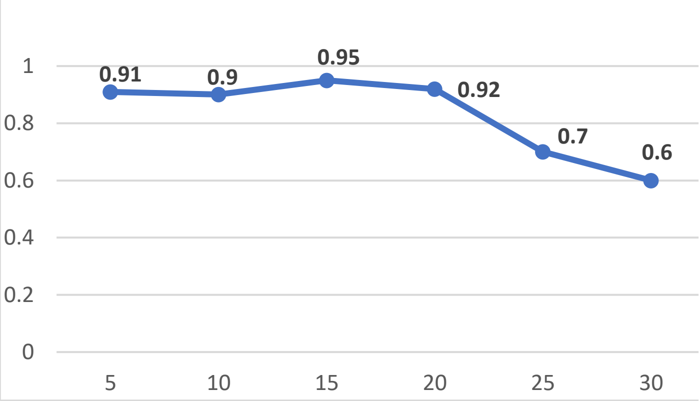


Figure 5: Impact of the of the value of k on accuracy

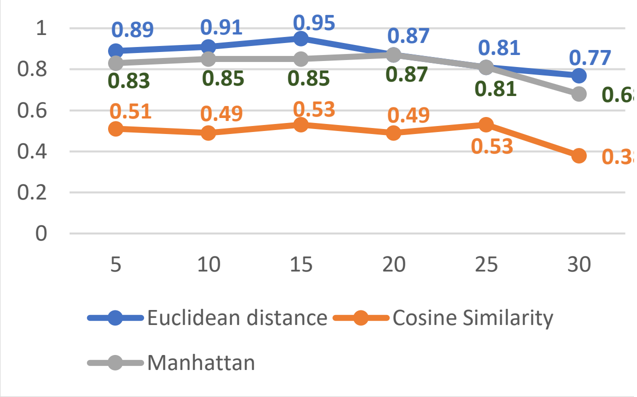


Figure 6: Comparison of different methods of calculating similarity

Support Vector Machines (SVM)

Support Vector Machines (SVM) are grounded in the principle of Structural Risk Minimization, a key concept in statistical learning theory. Recognized as a powerful and versatile machine learning algorithm, SVM has been extensively applied to tasks such as classification, estimation, and tracking. The algorithm identifies the support vectors—the training data points closest to the decision boundary—and uses them exclusively to classify new input vectors. The goal of structural risk minimization is to select a hypothesis *h* that ensures the lowest possible true error, balancing model complexity with generalization ability. To address real-world scenarios involving noisy data, Vapnik introduced the concept of a soft margin, allowing the model to tolerate certain classification errors while still maintaining robustness (Shanghai Second Polytechnic University, Shanghai, 201209, China and Xia 2016).

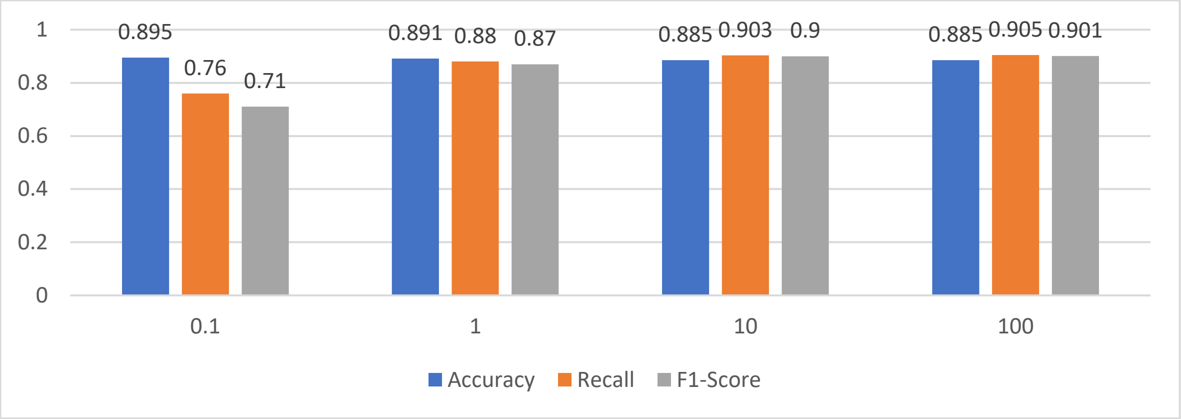


Figure 7: Impact of the of the value of C on Accuracy, Recall, F1-Score

Random Forest

The random forest, part of the ensemble learning family, stands out as a robust and adaptable classifier for educational contexts. It constructs multiple decision trees, each generated independently from a bootstrap sample of the training data (Thomas and J 2020) and a random subset of features. This architectural independence across trees fosters structural diversity, which enhances the model's generalization capability and mitigates overfitting. Such design proves effective, particularly when classifying heterogeneous learner profiles or diverse educational resources. Additionally, the use of out-of-bag data for internal evaluation eliminates the need for a separate test set and strengthens validation reliability. The algorithm also ranks features by importance, offering pedagogical insights into variable relevance. With its tolerance to noise, resilience to missing data, and ability to scale efficiently across large or high-dimensional datasets, the random forest emerges as a powerful driver of personalized learning pathways.

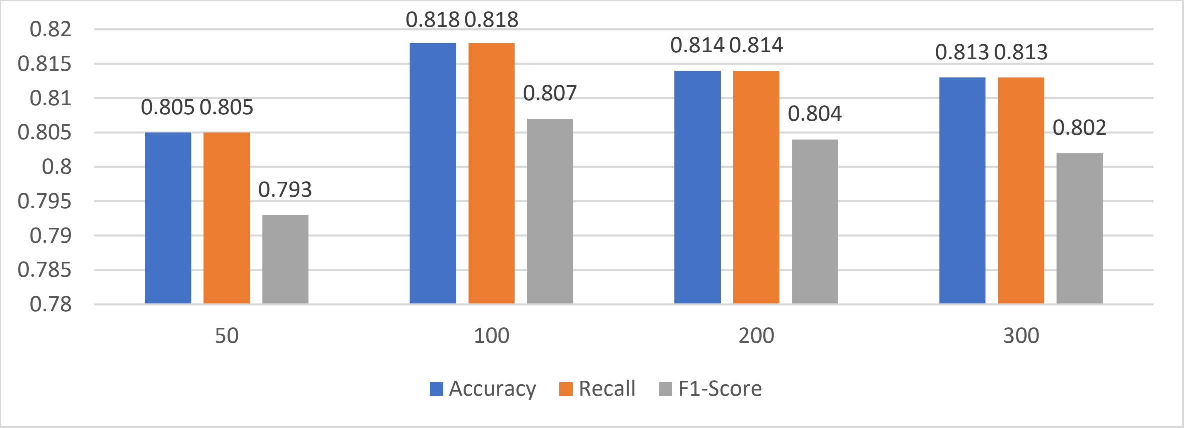


Figure 8: Impact of the of the value of C on Accuracy, Recall, F1-Score

Decision Tree

A decision tree is a hierarchical model in which each internal node represents a selection among various alternatives, while each terminal (leaf) node corresponds to a specific outcome or decision (Dunham and Seshadri 2006), (Lakshmi et al. 2013). Widely employed as a tool for extracting actionable insights, decision trees assist in the decision-making process (Lakshmi et al. 2013). The structure begins with a root node, where users initiate actions, and proceeds through recursive node splitting guided by a decision tree learning algorithm. The final structure maps out multiple decision paths, with each branch depicting a scenario and its corresponding result. Among the most commonly used algorithms for constructing decision trees are ID3, C4.5, and CART.

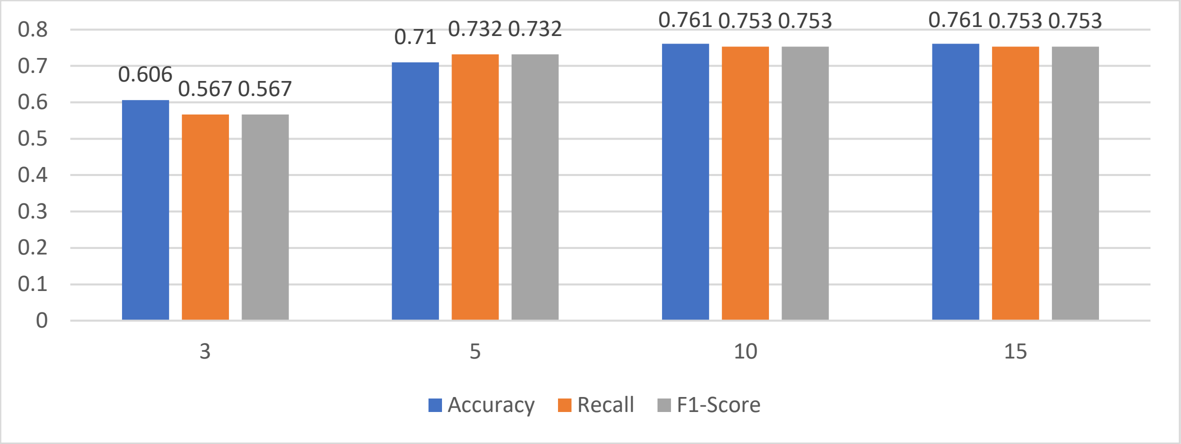


Figure 9: Impact of the of the value of max_depth on Accuracy, Recall, F1-Score

RESULTS

The table 1 provides a comparative overview of four classification algorithms KNN, Decision Tree, Random Forest, and Support Vector Machine (SVM) evaluated in the context of student and pedagogical resource classification.

KNN achieves consistently high performance across all three metrics (Accuracy, Recall, F1-score: 0.95), making it a strong candidate for balanced and reliable predictions. SVM also delivers competitive results, with a particularly strong Recall (0.91) and F1-score (0.91), indicating its robustness in identifying varied learner profiles. The Random Forest model maintains solid and stable performance (0.82), which, coupled with its interpretability, makes it an appealing option when model transparency is essential. Decision Tree, while generally robust in many domains, underperforms here (Accuracy: 0,76, Recall: 0.75, F1-score: 0.75), suggesting limited sensitivity in detecting diverse educational needs. Overall, KNN and SVM emerge as the most effective models for adaptive learning systems requiring both precision and inclusiveness.

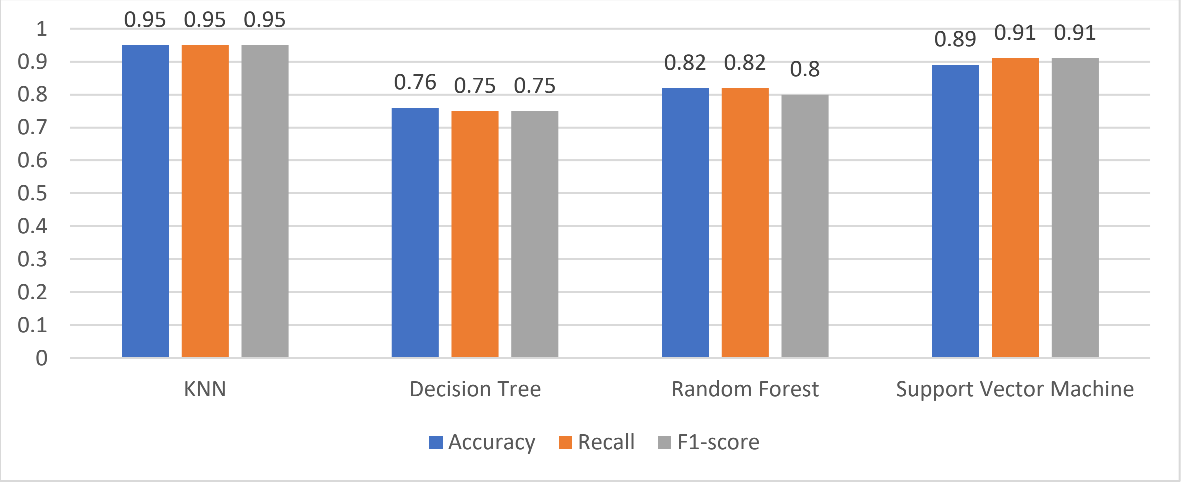


Figure 10: Comparison of the performance of KNN, SVM, Random Forest and Decision tree algorithms.

	Accuracy	Recall	F1-score
KNN	0,95	0,95	0,95
Decision Tree	0,76	0,75	0,75
Random Forest	0,82	0,82	0,80
Support Vector Machine	0,89	0,91	0,91

Table 1: comparative table between the different classification algorithms

DISCUSSION

In summary, the comparative analysis of classification algorithms applied to educational contexts identified KNN as the most performant model, both in terms of accuracy and its ability to capture the diversity of learner profiles. These results highlight the value of equipping intelligent pedagogical platforms with reliable models capable of guiding content adaptation decisions.

Building on this, our next work involves designing the architecture of a recommendation system based on a hybrid approach, combining collaborative filtering and content-based filtering, with the goal of seamless integration into a personalized learning environment. This direction aims to deliver more dynamic, contextualized, and adaptive learning pathways, ultimately enhancing educational success

REFRENCES

[1] Aher, Sunita B., and L. M. R. J. Lobo. 2012. “COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS.”

[2] Al-Kindi, Iman, and Zuhoor Al-Khanjari. 2022. “A Comparative Study of Classification Algorithms of Moodle Course Logfile Using Weka Tool.” International Journal of Computers and Their Applications 29:202–11.

[3] Arya, Ashish. 2025. “Comprehensive Analysis of Machine Learning and Deep Learning Models for Fake News Detection on Twitter.” Journal of Information Systems Engineering and Management 10(34s):1044–58. doi:10.52783/jisem.v10i34s.5909.

[4] Bobadilla, J., F. Serradilla, and J. Bernal. 2010. “A New Collaborative Filtering Metric That Improves the Behavior of Recommender Systems.” Knowledge-Based Systems 23(6):520–28. doi:10.1016/j.knosys.2010.03.009.

[5] Bobadilla, J., F. Serradilla, and A. Hernando. 2009. “Collaborative Filtering Adapted to Recommender Systems of E-Learning.” Knowledge-Based Systems 22(4):261–65. doi:10.1016/j.knosys.2009.01.008.

[6] Debnath, Suman, Nitish Sinha, and Bishanka Brata Bhowmik. 2022. “ML Based Modulation Format Identifier Using K-NN Algorithm.” Materials Today: Proceedings 65:2626–30. doi:10.1016/j.matpr.2022.04.880.

[7] Dunham, Margaret, and Dr. Sridhar Seshadri. 2006. Data Mining- Introductory and Advanced Topics.

[8] Ekstrand, Michael D., John T. Riedl, and Joseph A. Konstan. 2011. “Collaborative Filtering Recommender Systems.” Foundations and Trends® in Human–Computer Interaction 4(2):81–173. doi:10.1561/1100000009.

[9] El Moustamid, Anas, En-Naimi El Mokhtar, and Jaber Bouhdidi. 2017. Integration of Data Mining Techniques in E-Learning Systems: Clustering Profil of Lerners and Recommender Course System.

[10] Hazar, Manar Joundy, Samawel Jaballi, Mohsen Maraoui, Mounir Zrigui, and Henri Nicolas. 2025. “A Hybrid E-Learning Recommendation System Incorporating User Reviews and Ratings for Enhanced Course Selection.”

[11] Javed, Umair, Kamran Shaukat, Ibrahim A. Hameed, Farhat Iqbal, Talha Mahboob Alam, and Suhuai Luo. 2021. “A Review of Content-Based and Context-Based Recommendation Systems.” International Journal of Emerging Technologies in Learning (iJET) 16(03):274–306. doi:10.3991/ijet.v16i03.18851.

[12] Lakshmi, T. Miranda, A. Martin, R. Mumtaj Begum, and V. Prasanna Venkatesan. 2013. “An Analysis on Performance of Decision Tree Algorithms Using Student’s Qualitative Data.” International Journal of Modern Education and Computer Science 5(5):18–27. doi:10.5815/ijmecs.2013.05.03.

[13] Mohd, Ayesha. 2025. “Data Analytics in Machine Learning.” Journal of Information Systems Engineering and Management 10:114–22. doi:10.52783/jisem.v10i31s.5015.

[14] Nilla, Arliyanna, and Erwin Budi Setiawan. 2024. “Film Recommendation System Using Content-Based Filtering and the Convolutional Neural Network (CNN) Classification Methods.” Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika 10(1):17. doi:10.26555/jiteki.v9i4.28113.

[15] Osman, Ahmed Abdelgader Fadol. 2025. “A New Approach to Machine Learning Algorithms in Adaptive E-Learning Systems.” Journal of Information Systems Engineering and Management 10(15s):419–32. doi:10.52783/jisem.v10i15s.2480.

[16] Schafer, J. Ben, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. “Collaborative Filtering Recommender Systems.” Pp. 291–324 in The Adaptive Web: Methods and Strategies of Web Personalization, edited by P. Brusilovsky, A. Kobsa, and W. Nejdl. Berlin, Heidelberg: Springer.

[17] Shanghai Second Polytechnic University, Shanghai, 201209, China, and Tian Xia. 2016. “Support Vector Machine Based Educational Resources Classification.” International Journal of Information and Education Technology 6(11):880–83. doi:10.7763/IJIET.2016.V6.809.

[18] Su, Xiaoyuan, and Taghi M. Khoshgoftaar. 2009. “A Survey of Collaborative Filtering Techniques.” Advances in Artificial Intelligence 2009(1):421425. doi:10.1155/2009/421425.

[19] Thomas, Benny, and Chandra J. 2020. “Random Forest Application on Cognitive Level Classification of E-Learning Content.” International Journal of Electrical and Computer Engineering (IJECE) 10(4):4372–80. doi:10.11591/ijece.v10i4.pp4372-4380.

[20] Tian, Xinhua. 2024. “Content-Based Filtering for Improving Movie Recommender System.” Pp. 598–609 in. Atlantis Press.

[21] Tran, Phong, Khang Vu, Minh Doan, Quynh Dang, Quang Dang, and Thanh Ho. 2024. “Personalized Learning Paths Recommendation System with Collaborative Filtering and Content-Based Approaches.” Science & Technology Development Journal - Economics - Law and Management 8. doi:10.32508/stdjelm.v8i2.1370.