Developing an Adaptive Learning Recommendation Algorithm and System for MOOCs

Ying Zhang
Career Foundation Department, Changchun Polytechnic, Changchun, Jilin, China.
zy_820@126.com

Correspondence should be addressed to Ying Zhang : zy_820@126.com

Article Info
Journal of Machine and Computing (http://anapub.co.ke/journals/jmc/jmc.html)
Doi : https://doi.org/10.53759/7669/jmc202404089
Received 16 March 2024; Revised from 24 April 2024; Accepted 25 July 2024.
Available online 05 October 2024.

Abstract – Massive Open Online Courses (MOOC) based learning platform had totally changed the educational environment by providing easy and accessible learning opportunities for global learners. But even such environment display high dropout and low learner engagement which remain a significant challenge to be addressed. To handle the challenge of this study, propose an Adaptive Learning Recommendation System (ALRS) that is designed to personalize learning paths based on individual preferences and performance metrics. The study employed Open University Learning Analytics Dataset (OULAD) and build recommendation model that combine k-means Clustering, Content-based Filtering, Collaborative Filtering, and Random Forest (RF) classifiers to make course recommendations. The proposed model have shown better recommendation when compared to other models with Precision of 0.92, Recall of 0.89, F1 Score of 0.90, and AUC of 0.95. Also the proposed model had shown the lowest Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) at 0.042 and 0.205, respectively.

Keywords – Adaptive Learning Recommendation System, Online Learning, Machine Learning, Mean Squared Error, Root Mean Squared Error.

I. INTRODUCTION

In recent years the field of Massive Open Online Courses (MOOC) have grown a lot and revolutionized the field of higher education by providing a widespread access to educational resources for learners who are located across the globe [1-3]. But even achieving popularity the MOOC platforms often struggle with high dropout rates and low learner engagement that are mostly attributed to lack of personalization in course offerings [4-6]. The traditional MOOC platforms have all based on one-size-fits-all approach, however the learners come with diverse backgrounds and varying educational needs so such models fails to gather interest and commitment among the learners [7-8].

Recently many studies have focused on Adaptive Learning Systems (ALS) and their potential to enhance educational outcomes [9-10]. However, there remains a large gap in effectively integrating these systems within the MOOC platforms so that to address the challenges of learner diversity and engagement [11-12]. Many earlier works have employed Machine Learning (ML) techniques to predict course suitability and learner performance. But such models to do not totally consider the different aspects of learner feedback and progression.

This article introduces an Adaptive Learning Recommendation System (ALRS) for MOOCs with an aim to handle the above challenges. The model employ data analytics and ML models to modify dynamically the learning paths based on individual learner interactions and preferences. The model was experimented using Open University Learning Analytics Dataset (OULAD) to prove its efficiency against other traditional models.

The Contributions of The Work Are

(a) Develops a comprehensive model that integrates both content-based and collaborative filtering techniques to enhance the accuracy of course recommendations.
(b) Evaluates the impact of these personalized learning paths on user engagement and course completion rates.

The paper is organized as follows, Section 2 present the materials used in the study, Section 3 present the architecture and the proposed model, Section 4 present the analysis of the findings and Section 5 concludes the work.

II. MATERIALS

Dataset

The Open University Learning Analytics Dataset (OULAD) [13-15] is a publicly available dataset that has been widely used in the field of educational research. The dataset comprises data from about 32,593 students who were all engaged in 22 courses (modules). OULAD includes a variety of data variables that include demographic information, such as age and geographic location. It also include engagement metrics like clickstream data from the Virtual Learning Environment (VLE) [16-18], which includes daily summaries of how students interact with course materials and it also contains assessment scores, registration information that detail about the course selections and durations, and final results. The dataset contains 7 CSV files as presented in **Table 1**, and it requires pre-processing and transformation before presenting to recommendation model. The schema of the dataset is presented in **Fig 1**.

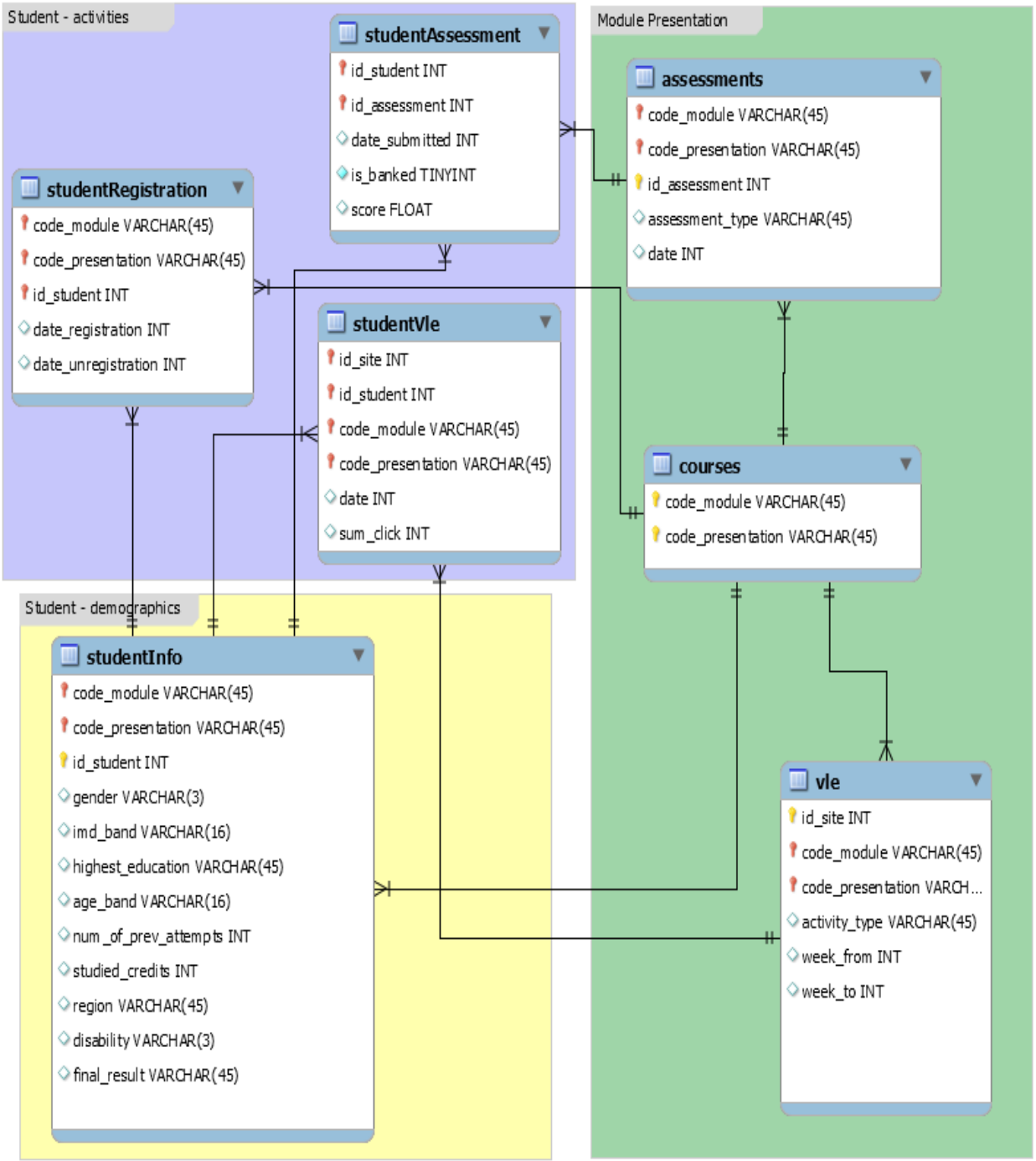

Fig 1. OULAD Database Schema.
(*Source: https://analyse.kmi.open.ac.uk/resources/images/model.png*)

Table 1. Tables Information of the Open University Learning Analytics Dataset

Table Name	Records	Table Attributes	Description
Student_Info	32.593	Code_Module, Code_Presentation, ID_Student, Gender, Region, Highest_Education, Imd_Band, Age_Band, Num_of_Prev_Attempts, Studied_Credits, Disability, Final_Result	Provides demographic and academic profiles of students.
VLE	6.365	Id_Site, Code_module, Code_Presentation, Activity_Type, Week_from, Week_To	Details VLE resources and their availability periods.
Student_VLE	1.048 .574	Code_Module, Code_Presentation, ID_Student, ID_Site, Date, Sum_Click	Tracks daily student interactions with VLE resources.
Student_Registration	32.593	lode_Module, code_Presentation, Module_Presentation_Length	Records registration details for course modules.
Assessments	196	Code_Module, Code_Presentation, ID_Assessment, Assessment_Type, Data, Weight	Lists course assessments, including types and scheduling.
Student_Assessments	173.740	ID_Assessment, ID_Student, Score Date_Submitted, IS_Banked	Contains scores and submission details for assessments.

Table 2. Features Selection for Adaptive Learning Prediction in a MOOC

Feature Type	Feature Name	Feature Data Type	Feature Values Encoding Type
MOOC Features	Code_Module	Categorical (Nominal)	(Int64)
	Module_Presentation_Length	Numerical (Discrete)	(Int64)
	Course_Start	Categorical (Ordinal)	(Int64)
	Course_End	Categorical (Ordinal)	(Int64)
Learner Features	Age_Band	Categorical (Ordinal)	(Int64)
	Highest_Education	Categorical (ordinal)	(Int64)
	Imd_Band	Categorical (ordinal)	(Int64)
	Num_of_Prev_Attempts	Numerical (discrete)	(Int64)
	Studied_Credits	Numerical (discrete)	(Int64)
	Disability	Categorical (binary)	(bool)
Outcome Features	Final_Result	Categorical (binary)	(bool)

Data Preprocessing

Feature Engineering and Dataset Preparation

Variable Adjustment

Certain variables that are more reflective of outcomes rather than predictors of engagement, such as *'Num_of_Prev_Attempts'* and *'Studied_Credits'*, are adjusted or removed to focus on factors influencing a student's initial motivation and ongoing engagement rather than their historical performance.

Defining Dependent and Independent Variables

In alignment with our objective to increase user engagement, the dependent variable (target) is identified as *'Final_Result'*, representing successful course completion. Independent variables are derived from both course characteristics (*e.g.,* module presentation length, course start and end dates) and learner demographics (*e.g.,* age band, highest education).

Transformation of Categorical Features

Many machine learning models require numerical input; therefore, categorical variables such as *'Gender', 'Region',* and *'Highest_Education'* are encoded using one-hot encoding. This process converts categories into a binary matrix, facilitating their use in predictive modeling.

Data Sampling

Given the imbalanced nature of class distribution in *'Final_Result'* (*e.g.,* more pass instances than fail), we apply sampling techniques to balance the dataset. This ensures that the predictive model is not biased towards the majority class and can accurately predict less frequent outcomes.

Data Cleaning

Steps are taken to clean the data thoroughly, which includes handling missing values either by imputation—where it makes sense—or by removing records that are incomplete to a degree that could skew the analysis significantly.

A summarization of the processed features, their types, and the modifications made are presented in **Table 2**.

III. SYSTEM ARCHITECTURE

The proposed ALRS is a III-Tier architecture (**Fig 2**) that consists of a data layer, an application layer, and a presentation layer. The Data Layer manages data storage that include user profiles, course information, and interaction logs. It employs both SQL and NoSQL databases to ensure robust data availability and rapid access. The Application Layer employs the proposed ALRS that processes data to generate personalized course recommendations. It includes modules for clustering using K-means to segment users, preference modeling that combines content-based and collaborative filtering, and a machine learning module that utilizes Random Forest (RF) algorithms to predict and adapt user preferences. The Presentation Layer, or user interface, is where users interact directly with the system that display the personalized course recommendations and provides a feedback mechanism for users.

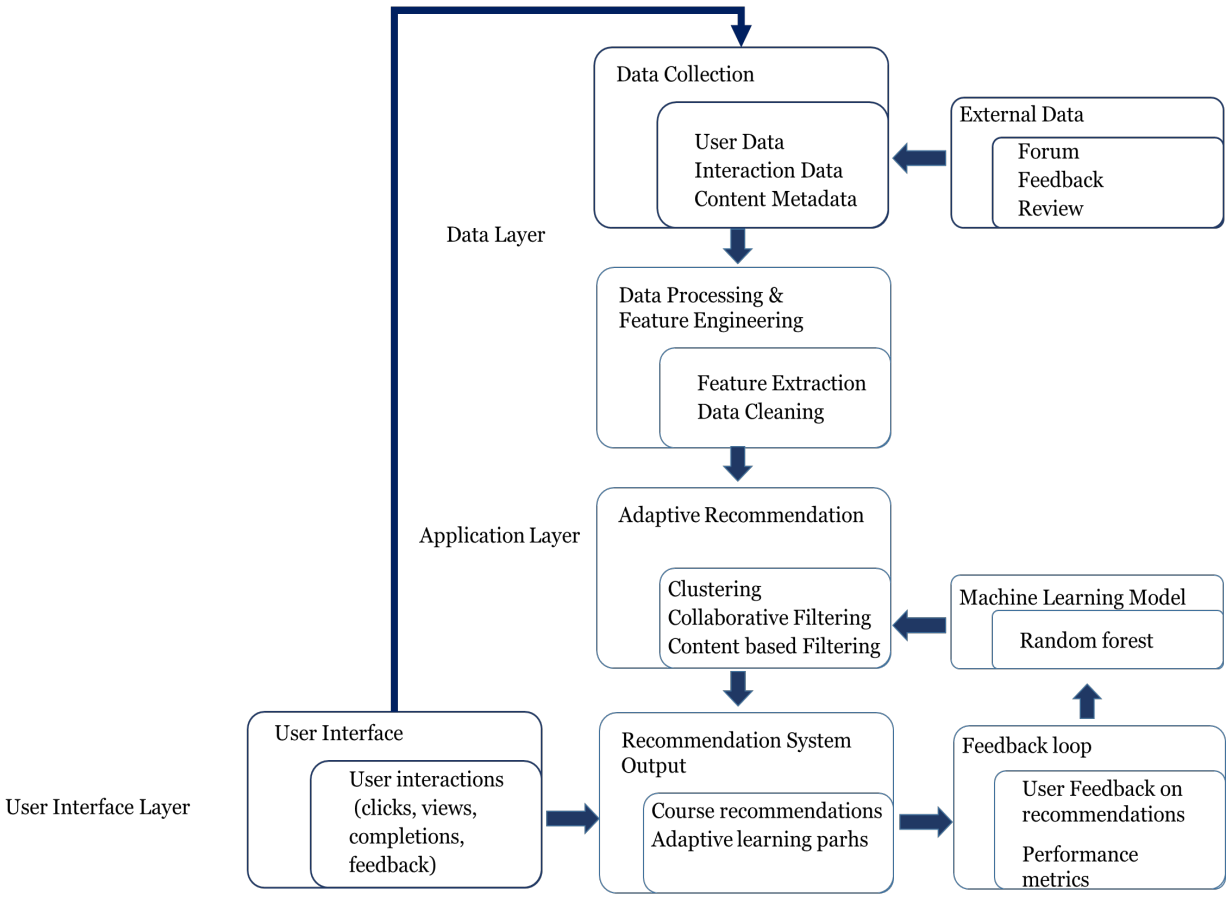

Fig 2. Architecture of the ALRS.

Proposed ALRS

The proposed ALRS integrates three key phases: clustering, preference modeling, and ML algorithms. In the first phase the K-means clustering model is employed to segment the users into groups based on activity patterns and learning outcomes, creating targeted cohorts. In the next phase a hybrid system that combine both the content-based and collaborative filtering is employed for course recommendations based on aligning the user preferences with that of course content. Finally the RF classifiers was used to predict user preferences and generate personalized course suggestions.

Clustering

For clustering we employ K-means clustering to segment MOOC users into distinct groups based on their activity patterns and learning outcomes. Let $X = \{x_1, x_2, \dots, x_n\}$ represent the set of n feature vectors, where each vector x_i corresponds to a user's interaction profile within the MOOC platform. The interaction profile includes variables such as course access frequency, time spent on learning materials, quiz scores, and forum participation. The objective of K -means clustering is to partition the n users into k distinct clusters $= \{C_1, C_2, \dots, C_k\}$, such that the within-cluster sum of squares (WCSS) is minimized. The WCSS is defined as follows EQU (1).

$$\text{WCSS} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (1)$$

where μ_i is the mean of points in C_i. The K -means algorithm iteratively updates the cluster assignments based on the following steps:

Initialization
Select k initial cluster centers randomly from the data points.

Assignment Step
Assign each data point x_i to the nearest cluster by minimizing the Euclidean distance between x_i and each cluster center μ_i.

Update Step
Recalculate the new cluster centers μ_i as the mean of all points assigned to C_i.

The algorithm repeats the assignment and update steps until the cluster assignments no longer change, indicating convergence.

Preference Modeling
Following the clustering phase, the recommendation system employs preference modeling approach that integrates both content-based and collaborative filtering techniques.

Content-Based Filtering
Content-based filtering focuses on the characteristics of the courses themselves. Each course is represented by a feature vector $\mathbf{f}_c$ that encapsulates attributes such as course topics, difficulty levels, and learning outcomes. For each user, a preference profile $\mathbf{p}_u$ is constructed based on their interactions with course content. The similarity between the user's profile and each course's features is computed using the cosine similarity measure EQU (2).

$$\text{similarity}\ (\mathbf{p}_u, \mathbf{f}_c) = \frac{\mathbf{p}_u \cdot \mathbf{f}_c}{\|\mathbf{p}_u\| \|\mathbf{f}_c\|} \quad (2)$$

Courses with higher similarity scores are recommended to the user, assuming these offerings align more closely with their established interests.

Collaborative Filtering
Collaborative filtering, on the other hand, leverages user behavior data to predict preferences. It operates under the premise that users who agreed in the past will agree in the future about course preferences. Using the matrix factorization technique, user-item interactions are decomposed into latent factors representing underlying characteristics EQU (3)

$$\mathbf{R} \approx \mathbf{U}\mathbf{V}^T \quad (3)$$

where $\mathbf{R}$ is the user-item interaction matrix, $\mathbf{U}$ is the user-factor matrix, and $\mathbf{V}$ is the item-factor matrix. The model predicts unknown entries in $\mathbf{R}$, which represent unobserved user-item interactions.

Machine Learning Model for Prediction
In the final phase we deploy Random Forest classifiers to predict user preferences and generate accurate course recommendations. The Random Forest model operates by constructing multiple decision trees during training and output the class that is the mode of the classes predicted by individual trees. Each tree in the forest is built from a random sample of data points and a subset of features which reduce the overfitting.

Let $\mathbf{X}$ represent the feature matrix where each row $\mathbf{x}_i$ corresponds to a user's profile, including both demographic and interaction data derived from the preprocessing steps. The target variable $\mathbf{y}$ represents user course preferences, categorized into classes such as 'highly interested', 'moderately interested', and 'not interested'. The decision function for a single tree can be represented as EQU (4).

$$\text{decision}\ (\mathbf{x}) = \sum_{t=1}^{T} \text{tree}_t\ (\mathbf{x}, \Theta_t) \quad (4)$$

where T is the number of trees, tree $_t$ is the prediction of the t-th tree, and Θ_t are the parameters (i.e., split points) of that tree. The final prediction $\hat{y}$ of the Random Forest is obtained by averaging the predictions of all the individual trees or taking a majority vote in the case of classification EQU (5).

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} \text{tree}_t\ (\mathbf{x}, \Theta_t) \quad (5)$$

The training of the Random Forest model involves optimizing parameters such as the number of trees in the forest, the maximum depth of trees, and the minimum number of samples required to split an internal node.

IV. EXPERIMENTAL SIMULATION

The proposed adaptive learning-based MOOC recommender system utilizes learner's ID to access historical interaction data. It then filters out courses that don't align with the learner's needs and preferences. For example consider a learner who is interested in doing a professional development course alongside a full-time job, who can allocate up to 6 hours each weekend for studying. This learner is particularly interested in advancing their skills in digital marketing and data analysis. Once this learner's profile is selected in the recommendation system, it assesses their previous course interactions and current learning goals. It then generates a tailored list of the top-10 MOOC (**Fig 3**) that match the learner's time constraints based on their professional aspirations.

course_id	recommended	discipline	grade_requirements	course_requirements	course_length (days)
832945145	True	Humanities	True	True	60
832945515	True	Mathematics & Statistics	True	True	35
832945665	True	Education	True	True	77
832945891	True	Applied Sciences	True	True	365
832960448	True	Interdisciplinary	True	True	122
832960714	True	Business Management	True	True	365
832960719	True	Applied Sciences	True	True	35
832960271	True	Applied Sciences	True	True	300
832960758	True	Humanities	True	True	42
832960903	True	Interdisciplinary	True	True	47

Fig 3. Recommendations from the ALRS.

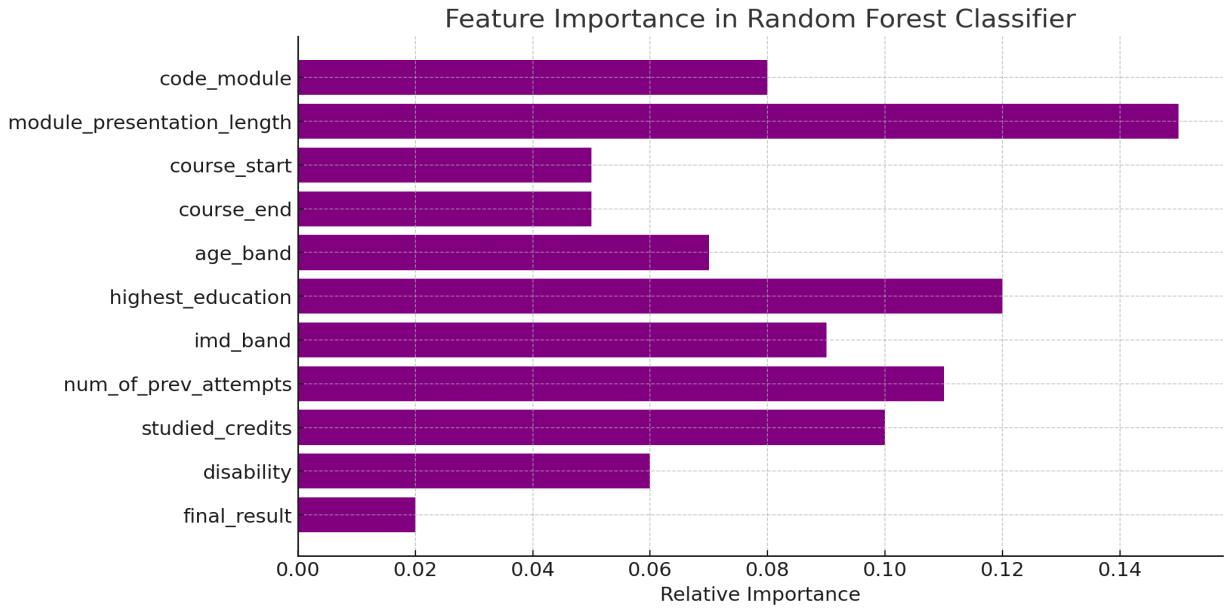

Fig 4. Feature Importance Analysis.

To optimize the recommendations a feature importance analysis is conducted within the Recommendation model. This analysis as shown in **Fig 4** highlights the significance of specific MOOC features such as course length, content depth, start/end times and learner preferences such as preferred study times and subject interest areas. To assess the effectiveness of this ALRS the following metrics are used to measure the system's effectiveness:

Precision

Precision $= \frac{TP}{TP+FP}$, Where TP is True Positives and FP is false positives. This metric evaluates the accuracy of the positive predictions this model makes.

Recall (Sensitivity)

Recall $= \frac{TP}{TP+FN}$, Where FN is false negatives. This metric assesses the model's ability to correctly identify all relevant instances.

F1 Score

$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$, The F1-score balances precision and recall, providing a comprehensive measure of model accuracy.

Area Under the Curve (AUC) - RoC Curve

$AUC = \int_0^1 TPR(t)dFPR(t)$, Where TPR is the true positive rate (recall) and FPR is the false positive rate. AUC measures the model's ability to distinguish between classes across different thresholds.

To evaluate our system's performance, we benchmark it against three established recommendation models such as: LightFM Hybrid Model, Singular Value Decomposition (SVD) and Neural Collaborative Filtering (NCF).

Table 3. Comparative Performance of Recommendation Models

Model	Precision	Recall	F1-score	AUC
Proposed ALRS	0.92	0.89	0.90	0.95
LightFM Hybrid Model	0.87	0.85	0.86	0.90
SVD	0.83	0.80	0.81	0.88
NCF	0.90	0.88	0.89	0.93

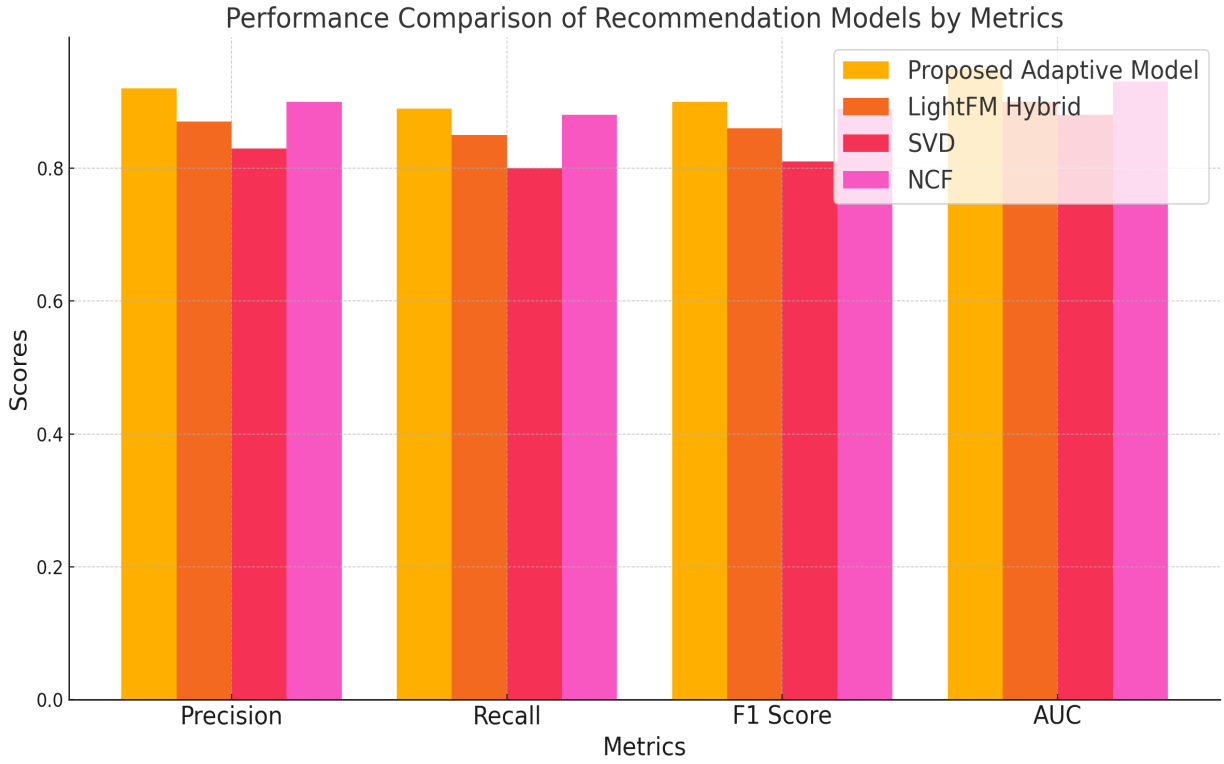

Fig 5. Performance Results of The Compared Models for Metrics.

The comparative analysis of recommendation models as shown in the **Table 3** and **Fig 5** highlights the effectiveness of the Proposed ALRS across all metrics. The proposed model achieves the highest scores in every class, including a Precision of 0.92, Recall of 0.89, F1-score of 0.90, and an AUC of 0.95. In contrast, the LightFM Hybrid Model and the SVD exhibit lower performance across the board, with the SVD showing the weakest performance among the models evaluated. The LightFM model, with scores of 0.87 in Precision, 0.85 in Recall, and an F1-score of 0.86, performs adequately but lacks the effectiveness of the proposed model. The NCF model stands out as the second most effective model after the proposed model, with scores close to the top performer in all metrics (Precision of 0.90, Recall of 0.88, and F1-score of 0.89).

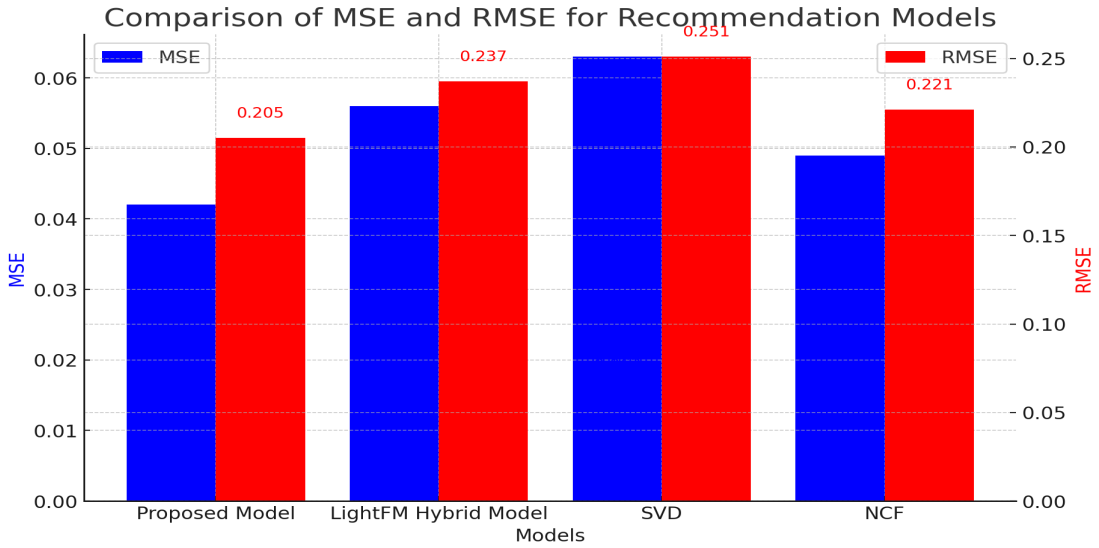

Fig 6. MSE and RMSE results.

Table 4. MSE and RMSE for the Different Models

Model	MSE	RMSE
Proposed ALRS	0.042	0.205
LightFM Hybrid Model	0.056	0.237
SVD	0.063	0.251
NCF	0.049	0.221

The **Table 4** and **Fig 6** compares the MSE and RMSE for different recommendation models. The Proposed ALRS demonstrates the best performance among the models, with the lowest MSE at 0.042 and RMSE at 0.205. These low error rates suggest that it is most effective at predicting user preferences and making accurate course recommendations. The LightFM Hybrid Model, which combines collaborative and content-based filtering, has higher error rates with an MSE of 0.056 and RMSE of 0.237. The SVD model shows the highest error metrics, with an MSE of 0.063 and RMSE of 0.251. NCF, which uses a neural network to model complex user-item interactions, records intermediate values of MSE and RMSE at 0.049 and 0.221, respectively.

V. CONCLUSION AND FUTURE WORK

This study proposes an ALRS to addresses the critical challenge of high dropout rates and low learner engagement in Massive Open Online Courses (MOOC). The proposed model integrates K-means clustering, content-based filtering, collaborative filtering, and RF classifiers to personalize learning paths based on individual learner profiles and preferences. To analyse the models performance the Open University Learning Analytics Dataset (OULAD) is employed and compared with traditional models such as LightFM Hybrid Model, SVD, and NCF. The model was compared for Precision, Recall, F1 Score, and AUC and for which the proposed model had achieved a Precision of 0.92, Recall of 0.89, F1 Score of 0.90, and AUC of 0.95. Additionally, it recorded the lowest MSE and RMSE at 0.042 and 0.205.

Future research focus on exploring additional data sources, enhancing model complexity, and incorporating real-time feedback mechanisms to further refine and optimize personalized learning experiences.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding agency is associated with this research.

Competing Interests

There are no competing interests

References

[1]. N. Voudoukis and G. Pagiatakis, "Massive Open Online Courses (MOOCs): Practices, Trends, and Challenges for the Higher Education," European Journal of Education and Pedagogy, vol. 3, no. 3, pp. 288–295, Jun. 2022, doi: 10.24018/ejedu.2022.3.3.365.

[2]. M. Nascimento Cunha, T. Chuchu, and E. T. Maziriri, "Threats, Challenges, And Opportunities for Open Universities and Massive Online Open Courses in The Digital Revolution," International Journal of Emerging Technologies in Learning (iJET), vol. 15, no. 12, p. 191, Jun. 2020, doi: 10.3991/ijet.v15i12.13435.

[3]. Y. Xiong, Q. Ling, and X. Li, "Ubiquitous e-Teaching and e-Learning: China's Massive Adoption of Online Education and Launching MOOCs Internationally during the COVID-19 Outbreak," Wireless Communications and Mobile Computing, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/6358976.

[4]. W. Wang, Y. Zhao, Y. J. Wu, and M. Goh, "Factors of dropout from MOOCs: a bibliometric review," Library Hi Tech, vol. 41, no. 2, pp. 432–453, Aug. 2022, doi: 10.1108/lht-06-2022-0306.

[5]. R. Wang, J. Cao, Y. Xu, and Y. Li, "Learning engagement in massive open online courses: A systematic review," Frontiers in Education, vol. 7, Dec. 2022, doi: 10.3389/feduc.2022.1074435.

[6]. H. Aldowah, H. Al-Samarraie, A. I. Alzahrani, and N. Alalwan, "Factors affecting student dropout in MOOCs: a cause and effect decision-making model," Journal of Computing in Higher Education, vol. 32, no. 2, pp. 429–454, Oct. 2019, doi: 10.1007/s12528-019-09241-y.

[7]. S. Reinhard, S. Serth, T. Staubitz, and C. Meinel, "From One-Size-Fits-All to Individualisation: Redefining MOOCs through Flexible Learning Paths," Proceedings of the Eleventh ACM Conference on Learning @ Scale, vol. 15, pp. 154–164, Jul. 2024, doi: 10.1145/3657604.3662037.

[8]. D. F. O. Onah, E. L. L. Pang, J. E. Sinclair, and J. Uhomoibhi, "An innovative MOOC platform: the implications of self-directed learning abilities to improve motivation in learning and to support self-regulation," The International Journal of Information and Learning Technology, vol. 38, no. 3, pp. 283–298, Apr. 2021, doi: 10.1108/ijilt-03-2020-0040.

[9]. I. Gligorea, M. Cioca, R. Oancea, A.-T. Gorski, H. Gorski, and P. Tudorache, "Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review," Education Sciences, vol. 13, no. 12, p. 1216, Dec. 2023, doi: 10.3390/educsci13121216.

[10]. F. Martin, V. P. Dennen, and C. J. Bonk, "A synthesis of systematic review research on emerging learning environments and technologies," Educational Technology Research and Development, vol. 68, no. 4, pp. 1613–1633, Jul. 2020, doi: 10.1007/s11423-020-09812-2.

[11]. M. L. Bernacki, M. J. Greene, and N. G. Lobczowski, "A Systematic Review of Research on Personalized Learning: Personalized by Whom, to What, How, and for What Purpose(s)?," Educational Psychology Review, vol. 33, no. 4, pp. 1675–1715, Apr. 2021, doi: 10.1007/s10648-021-09615-8.

[12]. S. E. Werners, R. M. Wise, J. R. A. Butler, E. Totin, and K. Vincent, "Adaptation pathways: A review of approaches and a learning framework," Environmental Science & Policy, vol. 116, pp. 266–275, Feb. 2021, doi: 10.1016/j.envsci.2020.11.003.

[13]. J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," Scientific Data, vol. 4, no. 1, Nov. 2017, doi: 10.1038/sdata.2017.171.

[14]. J. Shen, M. Ye, Y. Wang, and Y. Zhao, "Massive open online course (MOOC) in China: Status quo, opportunities, and challenges," 2016 IEEE Global Engineering Education Conference (EDUCON), Apr. 2016, doi: 10.1109/educon.2016.7474692.

[15]. D. E. Fatumo and S. Ngwenya, "Online learning platforms and their roles in influencing pass rate in rural communities of South Africa: Massive Open Online Courses(MOOCs)," 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC), Nov. 2020, doi: 10.1109/imitec50163.2020.9334135.

[16]. B. C. Padilla Rodriguez, "Success Indicators for Massive Open Online Courses (MOOCs)," 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT), Jul. 2020, doi: 10.1109/icalt49669.2020.00018.

[17]. B. Yulianto, G. Prajena, and M. T. Zulfikar, "GreatNusa: Fostering and Empowering the Society through Massive Open Online Course (MOOC)," 2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), Oct. 2021, doi: 10.1109/iceeie52663.2021.9616757.

[18]. K. Soraya, P. Purnawarman, and D. Suherdi, "Revisiting Massive Open Online Courses Concept in The 21st Century Era," 2019 2nd International Conference of Computer and Informatics Engineering (IC2IE), Sep. 2019, doi: 10.1109/ic2ie47452.2019.8940849.