International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-Jun-2026, ISSN (Online): 2395-566X

Privacy- Preserving Personalized Pathway Recommendation in Kenya's Competence-Based Education using Federated Learning, Cosine Similarity and Random Forest

Brian Levi Okimaru[1], Betty Mayeku[2], Humphrey Juma kilwake[2]
Department of Computer Science, School of Computing and Informatics, Kibabii University, Bungoma-Kenya

Abstract- The transition from junior to senior school under Kenya's Competency-Based Education (CBE) requires learners to select academic pathways that align with their competencies and interests. This transition presents a challenge because pathway selection requires personalized guidance while ensuring the privacy of sensitive student information. Existing educational recommender systems predominantly rely on centralized data processing, exposing learner data to privacy risks and limiting the secure exchange of information across institutions. This study proposes a privacy-preserving personalized pathway recommender system that integrates federated learning, cosine similarity, and Random Forest to support academic pathway recommendation without sharing raw student data. Cosine similarity was employed to model learner competency profiles and measure their alignment with predefined pathway requirements. The resulting similarity scores were incorporated into a Random Forest classifier through feature engineering to improve pathway prediction accuracy. A horizontal federated learning framework enabled multiple schools to collaboratively train the recommendation model by exchanging only model updates while retaining student records locally. The proposed model was evaluated using accuracy, precision, recall, and F1-score. Experimental results showed that integrating cosine similarity with Random Forest improved pathway classification performance, while the federated recommender system achieved an accuracy of 86.54%, outperforming the centralized recommender approach while preserving student privacy. The proposed framework provides an effective and privacy-preserving decision-support tool for personalized academic pathway recommendation within Kenya's Competency-Based Education. The study demonstrates that integrating federated learning with content-based filtering and machine learning can simultaneously enhance recommendation accuracy, personalization, and data privacy in educational environments.

Keywords- Competency-Based Education, Federated Learning, Random Forest, Cosine Similarity, Educational Recommender System, Privacy Preservation, Academic Pathway Recommendation.

I. INTRODUCTION

Artificial Intelligence and machine learning have revolutionized industries such as healthcare, finance, transportation, cybersecurity, and education. Traditional centralized machine learning approaches aggregate data into a single repository, enabling model training but raising serious privacy, confidentiality, and security concerns. Federated learning has emerged as a promising alternative, allowing decentralized model training across multiple institutions without sharing raw data, thereby addressing privacy and ethical challenges [1].

In education, recommender systems are increasingly used to support learners in course selection, personalized learning, and academic pathway guidance. However, centralized recommender systems often rely on vast amounts of student data, creating privacy risks. Researchers have emphasized the need for privacy-preserving approaches in educational contexts, where sensitive student information must be protected [2]. Kenya's ongoing transition from the 8-4-4 system to Competence-Based Education (CBE) underscores the importance of accurate and ethical pathway recommendations. Under CBE, students must select pathways in Social Sciences,

STEM, or Arts and Sports Science [3], a process complicated by limited guidance and the risk of misaligned choices.
Recommender systems typically employ collaborative filtering, content-based filtering, or hybrid approaches. Content-based filtering, particularly cosine similarity, has been effective in mapping student competencies to pathway requirements [4], while Random Forest has demonstrated strong performance in educational prediction tasks [5].

However, these techniques have traditionally been applied in isolation, limiting their effectiveness. This study integrates cosine similarity with Random Forest to leverage their complementary strengths, enhancing pathway prediction accuracy. To further address privacy concerns, the system is implemented within a federated learning framework, ensuring decentralized training and compliance with data protection standards.

The proposed federated recommender system contributes to Kenya's CBE objectives by offering personalized, transparent, and privacy-conscious pathway recommendations. It empowers students, educators, and policymakers to make informed decisions while safeguarding sensitive data, thereby advancing equitable learning outcomes and ethical AI adoption in education.

II. RELATED WORK

a. Privacy Preservation with Federated Learning in Education

Traditional recommender systems have achieved notable success in education, but concerns about data privacy have prompted exploration of federated learning. Fachola et al.
[6] demonstrated that federated learning could predict student dropout more effectively than centralized systems while preserving privacy. Farooq et al. [7] applied federated learning with SVM to predict student grades, outperforming decision trees, Naïve Bayes, and k-NN. Guo et al. [8] introduced FEEDAN, a federated framework that prevented direct data exchange between institutions while improving analysis quality. Zhang et al. [9] proposed FLPADPM, combining CNN and LSTM layers in a federated setting, achieving superior dropout prediction accuracy. These studies confirm that federated learning can balance privacy and performance in educational contexts.

b. Recommender System Approaches

Recommender systems generally fall into three categories: collaborative filtering, content-based filtering, and hybrid approaches. Collaborative filtering, both memory-based and model-based, identifies similarities between users or items without analyzing item content [10]. While effective, it suffers from scalability, cold-start, and sparsity issues [11]. Content-based filtering, by contrast, relies on user and item profiles, often represented as vectors. Cosine similarity is a common metric, with Mohanty et al. [12] noting that higher scores indicate stronger similarity. Hybrid systems combine multiple techniques, such as weighted, switching, or cascade methods, to improve recommendation quality [13].

c. Recommender Systems in Education

Recommender systems have been widely applied in education, including elective course recommendations [14], online enrollment systems [15], and repeated course performance prediction [16]. Collaborative filtering remains the most common approach, as highlighted by Algarni and Sheldon [17] and Urdaneta-Ponte et al. [18]. Content-based filtering has also been explored, such as Mokarrama et al. [19], who recommended private universities in Bangladesh using GPA and institutional attributes. Hybrid approaches, like Yang et al. [20] and Chang et al. [21], further improved recommendation accuracy by combining collaborative filtering with other algorithms. Narrowing to content-based recommender systems (CBRS), cosine similarity has been widely used to match student profiles with course attributes [4], [22]. However, while effective in measuring similarity, cosine similarity alone cannot capture complex feature interactions or predictive relationships.

d. Machine Learning Techniques in Education

Machine learning has been extensively applied to predict student performance. Sokkhey [23] achieved over 97% accuracy in predicting mathematics achievement using Random Forest. Hashim et al. [24] applied multiple supervised learning techniques to predict final grades, with logistic regression performing best. Saleem et al. [25] demonstrated that ensemble approaches improved predictive performance, with Random Forest achieving strong F1-scores. Other studies, such as Harvey and Kumar [26], Amra and Maghari [27], Adnan et al. [28], and Tarik et al. [29], confirmed the effectiveness of Random Forest in educational prediction tasks. Despite its strengths, Random Forest lacks explicit similarity measures, limiting its ability to capture feature directionality.

e. Research Gaps

The reviewed literature highlights several gaps. First, few studies have applied recommender systems to junior school students in Kenya, despite the importance of pathway selection under CBE. Second, while cosine similarity and Random Forest have been used independently, their integration remains underexplored. Third, centralized recommender systems raise privacy concerns, underscoring the need for federated learning. This study addresses these gaps by developing a federated recommender system that integrates cosine similarity and Random Forest to provide accurate, personalized, and privacy-preserving pathway recommendations.

III. METHODOLOGY

a. Research Design

This study adopted an experimental, quantitative design, leveraging federated learning, cosine similarity, and Random Forest to recommend academic pathways. Models were trained locally at institutions, with only model updates shared to a central server, ensuring privacy preservation.

b. Data Collection & Preprocessing

Data were collected from junior schools in Bungoma County. Each student record included Grade 7–9 subject scores and pathway interests. Preprocessing involved mean imputation for missing values, normalization to prevent bias, and feature selection to eliminate redundancy.

c. Data preprocessing

The data was subjected to various pre-processing processes to guarantee quality and consistency. The missing data were filled with mean imputation, and numerical features were normalized to improve the performance of the model, and categorical variables were encoded.

There was also the use of feature selection methods to eliminate any irrelevant or redundant data and therefore increased the efficiency of the computation process.

d. Data Model

The federated recommender system was developed using a hybrid approach that combines federated learning, cosine similarity, and random forest for classification. Federated learning ensured data remains local in the institutions while only sharing model parameters and updates with the server. This approach leverages both techniques' strengths to improve recommendations' accuracy and privacy.

Federated learning

This study employed horizontal federated learning, where institutions had datasets with the same features but different students. Each school trained a local Random Forest model using student scores and cosine similarity values, then shared model updates (not raw data) with a central server. Updates were aggregated using weighted averaging, producing a global model redistributed to schools.

Diagram(conceptual)

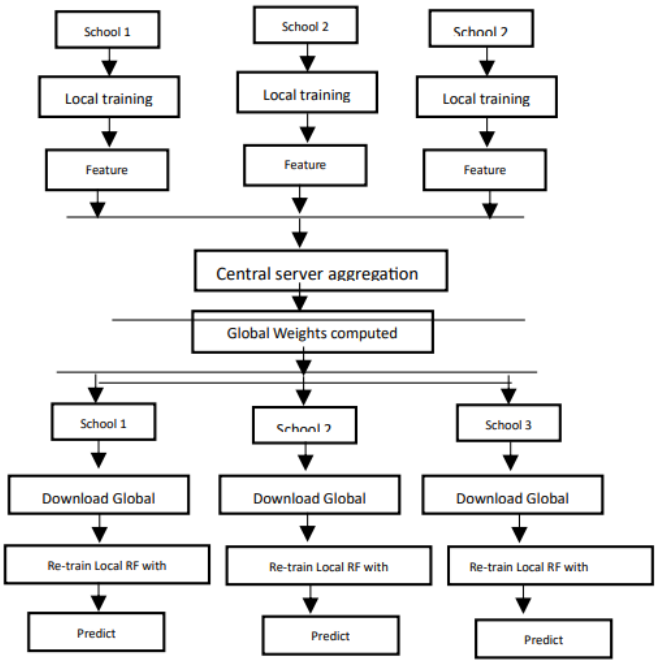

Figure 3.1: Federated learning framework

Dataset Preparation

Independent variable included grades 7–9 scores (Math_Grade7 … Pretechnical_Grade9) and Cosine similarity scores (STEM_Similarity, Social Science_Similarity, Arts and Sports_Similarity). Dependent variable was Best Pathway (from cosine similarity). Dataset was split 70% training, 30% testing and stratified by pathway.

Model Training

Trained local Random Forest on each school's data. Extracted feature importance

Uploaded weights to central server with dataset size. Aggregated global weights on server:

Global Weight*i*

$$= \frac{\sum_{k=1}^{n\_schools} (\text{Local Weight}_i^{(k)} \times \text{Samples}^{(k)})}{\sum_{k=1}^{n\_schools} \text{Samples}^{(k)}}$$

Downloaded global weights and used them as sample weights in local retraining:

$$\text{Sample Weight}_i = \frac{X_{\text{train}} \cdot \text{Global Weights}}{\sum X_{\text{train}} \cdot \text{Global Weights}}$$

Number of trees in Random Forest used was 100. Random seed was 42 for reproducibility

Cosine Similarity

Cosine similarity is one of the many techniques in content-based filtering. In determining how similar a student's competencies were to the ideal pathway subject performance, this study used cosine similarity matrix. Cosine similarity is a mathematical technique used to measure how similar two vectors are, regardless of their magnitude. In this study, it was employed to quantify the alignment between a student's competency profile (e.g., academic scores in subjects like Mathematics, Integrated Science, and Creative Arts) and the ideal competency profiles predefined for each senior school pathway (STEM, Social Science, Arts/Sports).

Dataset Description

The dataset used in this study was stored in a CSV file and imported into Python using the panda's library. The dataset contains student academic performance records across three academic years: Grade 7, Grade 8 and Grade 9. Each subject score is structured as:

Table 3.1: structure of csv file

Math_ Grade7	math_ Grade8	math_ Grade9	Eng_ Grade 7	Eng_ Grade 8	Eng_ Grade 9

Pathway-Specific Subjects

Each pathway is defined by a set of core subjects:

Table 3.2: pathway and their core subjects

Pathway	Subjects Considered
STEM	Math, Integrated, Agriculture, Pretechnical
Social Science	English, Kiswahili, Social, CRE
Arts and Sports	English, Kiswahili, Creative_Arts_and_Sports

Additionally, the dataset includes an actual Pathway column used as the ground truth label for evaluation, which was based on student interest.

Thus, the system operated primarily on structured academic performance data spanning three years.

Data Preprocessing

Before computing similarity scores, preprocessing was performed to ensure data consistency and fairness.

Subject Weight Normalization

Although all subject scores are on a 0–100 scale, average performance trends differ across subjects. Some subjects may have higher or lower overall means. To prevent certain subjects from disproportionately influencing similarity computation, subject-level normalization weights were computed.

For each subject:

$$Weight_{subject} = \frac{\max(\text{Mean Subject Score})}{100}$$

This produced a scaling factor between 0 and 1.
These weights were later used to scale both student performance vectors and pathway profile vectors.

Construction of Ideal Pathway Profiles

Each pathway was assigned a predefined ideal performance profile across Grade 7–9.
For example, the STEM pathway ideal profile is structured as:

[65,90,90,70] × 3

This means the ideal STEM student was expected to demonstrate strong performance in Mathematics (65), Integrated Science (90), Agriculture (90), and Pretechnical (70) consistently across three academic years.

Similarly, Social Science and Arts and Sports pathways were assigned ideal subject expectations reflecting pathway-specific competence patterns.

Student Performance Vector Construction

For each student: Relevant subject scores for a pathway were extracted, Scores were arranged sequentially by grade (Grade 7 → Grade 8 → Grade

9) and the values were combined into a single numerical vector.

For example, in the STEM pathway (4 subjects across 3 grades), the student vector contains 12 values. This structured representation allows direct comparison with the corresponding pathway profile vector.

Weight Application
Before computing cosine similarity, subject-level weights were applied to both the student vector and the ideal pathway profile vector.
Mathematical Formulation
Element-wise multiplication was applied:
$Weighted\_Student = Student\_Vector \times Weight\_Vector$
$Weighted\_Profile = Profile\_Vector \times Weight\_Vector$

Where:
Student_Vector = actual student performance values, Profile_Vector = ideal pathway performance values and Weight_Vector = subject normalization weights
Weights were repeated across grades to maintain vector alignment.

This ensures balanced subject contribution and reduced dominance of high-magnitude subjects.

Cosine Similarity Computation
Cosine similarity measures the angle between two vectors and is defined as:

$$Similarity = \frac{A \cdot B}{||A|| \ \ ||B||}$$

Where:
A= Weighted Student Vector, B= Weighted Pathway Profile Vector, $A \cdot B$= dot product and $||A||$and $||B||$= Euclidean norms
The similarity score ranges between: $0 \leq Similarity \leq 1$
Where: 1 indicates perfect alignment and 0 indicates no alignment
Pathway Selection
For each student, similarity scores were computed for: STEM, Social Science and Arts and Sports
The predicted pathway was determined as:
$Best\_Pathway = \arg\max(Similarity\_Scores)$
The pathway with the highest similarity score was selected as the student's competence-aligned pathway.

Random Forest
In this study, the Random Forest classifier was used to predict the most suitable pathway for students by analyzing both their raw competency data and precomputed cosine similarity scores.

Random Forest Classifier was initialized with random_state=42 for reproducibility. A stratified 70/30 train-test split ensured balanced representation. Number of trees (n_estimators) used was 100, this number provides a good balance between stability and training time. Each decision tree in the forest was trained on a bootstrap sample of the training data. Splits features using the Gini impurity criterion to maximize class separation. Majority voting across all trees determines the final predicted pathway.

e. Model Evaluation
To assess the performance of the models the following metrics were used:

$$\text{Accuracy} = \frac{Number\ of\ Correct\ predictions}{Total\ Number\ of\ predictions}\ x\ 100 = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100$$

$$\text{Precision} = \frac{correct\ pathway\ prediction}{All\ prediction\ for\ that\ pathway} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{correct\ pathway\ prediction}{All\ Actual\ students\ in\ that\ pathway} = \frac{TP}{TP+FN}$$

$$\text{F1-score} = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall}$$

IV. RESULTS AND ANALYSIS

a. Compute Similarity with Pathway Requirements using Cosine Similarity.
Cosine similarity approach generated similarity scores for each student across STEM, Social Science, Arts and Sports as shown in table 4.1. The pathway with the highest similarity score was selected as the recommended pathway. Actual pathway represented learner interested pathway based on the questionnaire.

Table 4.2: Computed similarity with pathway requirements using cosine similarity

Name	STEM_Similarity	Social Science_Similarity	Arts and Sports_Similarity	Best pathway	Actual pathway
April	0.9402	0.9645	0.987	Arts and Sports	STEM

June	0.9255	0.8181	0.8091	STE M	STEM
Levi	0.8936	0.8808	0.8956	Arts and Spor ts	Social science
Ann	0.9052	0.8709	0.8383	STE M	Arts and sports
Mitc hem	0.9075	0.9282	0.9426	Arts and sports	STEM

The results indicated that the cosine similarity was an effective measure of similarity between student competencies and pathway requirements. Higher STEM similarity showed a student performed well in science-related subjects, while those who excelled in languages aligned with Social Science and those with higher marks in creative arts aligned more with Arts and Sports pathways. Similarity scores closer to 1 indicated that the student's academic profile aligned well with the pathway requirements.

In Table 4.1, the model correctly identified pathway for example June was interested in STEM and was correctly predicted in STEM; this correct prediction showed a True Positive (TP). However, the model also had mismatches between actual and best pathways in both April and Mitchem pathways, when the predicted pathway differed from the actual pathway, resulting in False Positives (FP) or False Negatives (FN). Such misclassifications implied that cosine similarity could measure direction, but failed to capture complex relationships between subjects and overlapping competencies within pathways. This resulted to ambiguity in classification, when a student achieved good performance in both science and creative subjects.

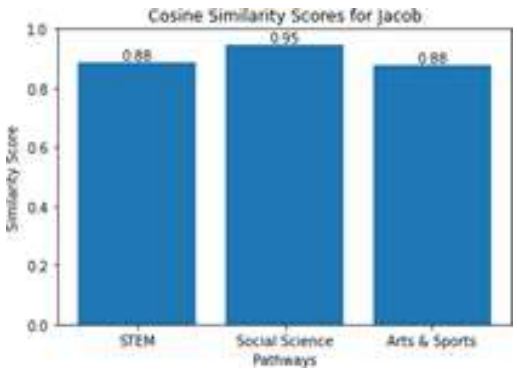

Figure 4.2: Cosine similarity for individual student

The cosine similarity scores for one student for the three pathways are shown in Figure 4.1. The highest similarity score was in the social science pathway; that is, the student's competencies were more similar to the social science pathway than to STEM or Arts and Sports. The difference in pathway alignment is plotted on the graph using cosine similarity score. The differences between these bars were highly indicative of the model's effectiveness in distinguishing between competency patterns and recommending the optimal pathway.

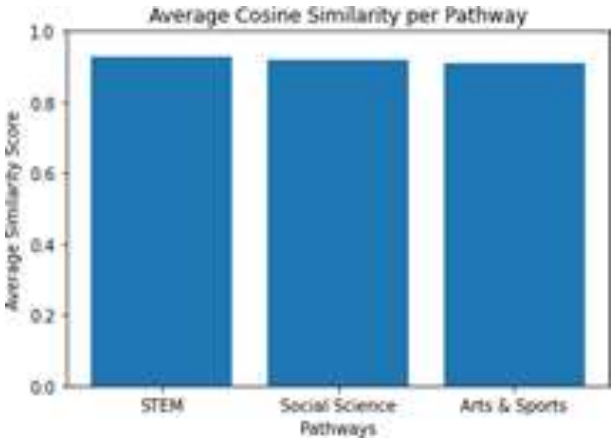

Figure 4.3: Average cosine similarity per Pathway

The average Cosine similarity scores for all students are presented in Figure 4.2 for the three pathways. Results showed overall alignment between students' competencies and pathway requirements. The domain with the highest average similarity score was the one with the strongest alignment of students' competencies. This validated the use of cosine similarity to reflect the overall performance trends in the data set.

b. Integrating Cosine Similarity with Random Forest.

Table 4.2 presents a comparison of the performance of two Random Forest models. The baseline model utilized only students' competence performance scores as input features, whereas the hybrid model incorporated both competence scores and cosine similarity values. The inclusion of similarity scores resulted in substantial improvement across all evaluation metrics, namely accuracy, precision, recall and F1-score.

Table 4.3: Evaluation metrics for Random Forest + Cosine similarity

Model variant	Accuracy	Precision	Recall	F1-score
Baseline Random Forest	43.91%	43.10%	43.91%	37.54%

Random Forest + Cosine similarity score	85.90%	86.32%	85.90%	85.54%

The accuracy, which reflects the percentage of correctly predicted pathways out of the total number of pathways, improved significantly from 43.91% in the baseline model to 85.90% in the hybrid model. The baseline model was able to correctly classify less than 50% of students, whereas the hybrid model classified the majority of students. This improvement showed the value of similarity information to greatly improve the model's discrimination between pathways.

Precision went up from 43.10% to 86.32% and measures the percentage of predicted pathway assignments that were correct out of the total number of assignments predicted for a given pathway. The low precision of the baseline model led to many false positives, resulting in students being placed in inappropriate pathways. The hybrid model, on the other hand, significantly reduced these errors, resulting in more accurate forecasts.

The model's ability to correctly identify actual pathway members (recall) also improved significantly from 43.91% to 85.90%. Many students were not identified in the correct pathway by the baseline model, indicating many false negatives. The hybrid model identified the most relevant cases, with higher sensitivity for pathway identification.

In the baseline model, the F1-score was 37.54%, whereas, in the hybrid model, it was 85.54% which is the Harmonic Mean of precision and recall. The F1 score for the baseline model was low, suggesting an imbalance between precision and recall, whereas the hybrid model had a higher F1 score, suggesting better performance and fewer classification errors.

Several factors may explain the lower performance of the baseline Random Forest model: the model used only raw academic scores, which do not reflect the extent to which student competencies are aligned to pathway requirements; This led the model to generate many false positives and false negatives, indicating that it was unlikely to make decisions based on a broad understanding of the relationship between competencies and pathways and more likely to make decisions based on isolated subject competency metrics. To overcome this limitation, cosine similarity was incorporated into the hybrid model to provide an explicit competency-pathway alignment competency measure. This improvement to the model's feature space provided more structure and context to the predictions, resulting in a significant decrease in false-positive and false-negative rates.

The feature importance analysis presented in Figures 4.3 and 4.4 provided additional understanding of how the model made its predictions. Figure 4.3 shows the model without cosine similarity, it depended mostly on individual subject scores, particularly Creative Arts and Integrated Science. This meant that the recommendation process was strongly influenced by performance in specific subjects rather than the student's overall competency profile. After cosine similarity was introduced into the model, a different pattern was observed as illustrated in Figure 4.4. The importance of similarity-based features increased and became more dominant in the prediction process. This indicated, model began to place greater emphasis on how closely a student's competencies matched a given pathway. Consequently, the recommendations reflected a more holistic assessment of student suitability instead of focusing mainly on separate subject scores.

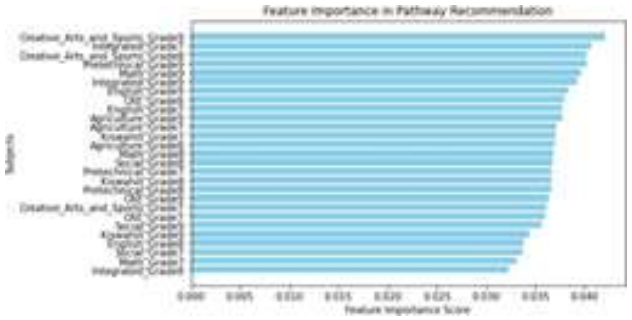

Figure 4.4: Feature importance score for Random Forest with no cosine similarity

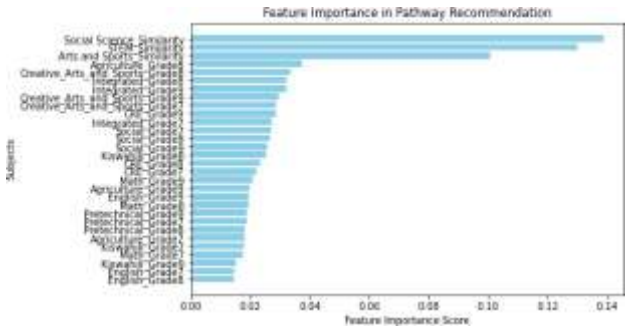

Figure 4.5: Feature importance score for Random Forest with Cosine similarity

The results showed that cosine similarity was a composite feature that showed the relation between the scores of the various subjects in a single form. This enabled the model to

detect patterns that were not easily observable on a subject-by-subject basis. This minimized the impact of irrelevant differences in the data and enhanced the model's ability to accurately predict across various cases. This resulted in better classification accuracy and more reliable decision-making.

a. Confusion Matrix Comparison

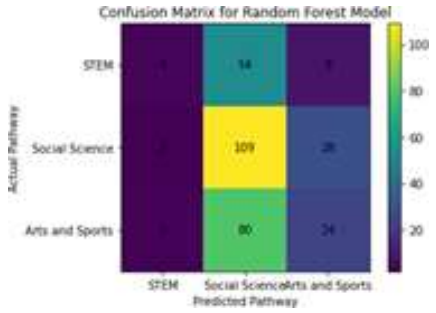

Figure 4.6: Confusion Matrix for baseline Random Forest

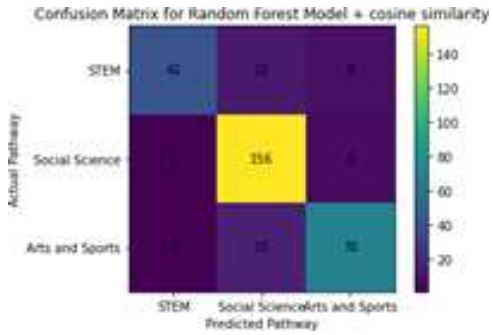

Figure 4.7: Confusion Matrix for Random Forest + Cosine similarity

The confusion matrix for the baseline Random Forest model and the Random Forest model with cosine similarity are displayed in Figures 4.5 and 4.6, respectively. The baseline Random Forest model correctly classified 4 STEM students, 109 Social Science students and 24 Arts and Sports Science students. By contrast, the Random Forest and cosine similarity model correctly classified 42 students in STEM, 156 in Social Science, and 70 in Arts and Sports Science. The diagonal cells in the second matrix were markedly brighter, indicating a substantial increase in correct classifications across all pathways.

The baseline Random Forest model recorded a high number of misclassifications, particularly between Arts and Sports Science and Social Science, as well as between STEM and Social Science. This suggested that the model had difficulty distinguishing pathways with related or overlapping competencies when only raw subject scores were used. After cosine similarity was incorporated, the number of misclassifications reduced considerably. Errors involving STEM students being classified as Social Science students decreased from 54 to 12, while misclassification between Arts and Sports Science and Social Science reduced from 80 to 15. At the same time, the number of correct predictions increased across all pathways. These results indicated that cosine similarity improved the representation of relationships between student competencies and pathway characteristics. Consequently, the hybrid model combined the pattern-learning capability of Random Forest with cosine-based similarity, resulting in better classification performance.

The confusion matrix comparison showed that the proposed model's performance has improved significantly. The baseline Random Forest model handled the patterns in the data, but it failed to capture the path-requirement and competency relationship well. Adding this new attribute, cosine similarity, gave a more accurate representation of these relationships in the model. The hybrid approach, therefore, enabled more accurate predictions and fewer pathway misclassifications. A reduction in confusion of pathways indicated that the profile of competency-based distinctions among students became more consistent across the model. The model improved prediction accuracy and provided more explainable, interpretable recommendations. Results suggested that using similarity measures as a component of feature engineering can make ensemble learning models, such as Random Forests, more efficient.

Develop Federated Recommender System and Evaluate its Effectiveness

The study aimed to improve the accuracy of pathway recommendations while ensuring that student data remained private. This was done by enabling multiple schools to access the Random Forest model training without sharing individual student information. To enhance the quality of recommendations, cosine similarity scores were added to the input features. These scores enabled the model to address the relationships between students' competencies and the demands of various pathways and to provide more suitable pathway recommendations.

Table 4.4: Federated recommender system output

Stu den't Name	STEM_ Similarity	Social_ Science_ Similari ty	Arts_ and_S ports_ Simila rity	Reco mme nded Path way	Best Pathway
Brian	0.972	0.9823	0.96	Socia l scien ce	Social science
Mitchel	0.908 7	0.9239	0.8571	Socia l Scien ce	Social Science
June	0.9749	0.9439	0.946	STE M	STEM

Table 4.5: Federated recommender system vs centralized recommender system performance

Model variant	Accurac y	Precision	Recall	F1-score
Centralized recommender system	85.90%	86.32%	85.90 %	85.54 %
Federated recommender system	86.54%	87.03%	86.54 %	86.04 %

Table 4.3 presents the results from the federated recommender system, while Table 4.4 compares the federated and centralized recommender approaches. The centralized recommender system was trained on a single combined dataset, whereas the federated recommender system was trained on data stored across different local institutions, with only model updates shared and aggregated during training. Both approaches used academic performance scores and cosine similarity features as input variables for pathway recommendation.

All evaluation metrics showed that the federated recommender system was better than the centralized recommender system. It had a significant improvement in accuracy, with more students correctly classified than in the centralized approach. This suggests that the federated model was more effective at data generalisation from multiple data sources.

When it came to the precision, the federated model made fewer false-positive predictions; that is, fewer students were misclassified into pathways that did not align with their competencies. The recall was also improved, meaning the model identified students who were actually in each pathway more effectively, thereby lowering the number of false negatives. The high F1 score additionally validated the model's overall performance, indicating that the model had a more balanced precision and recall. The improvement was due to several factors. More diversity in student performance patterns was introduced by the use of data from multiple schools, further enhancing the learning process of the model. Furthermore, training using data from multiple distributions improved the model's ability to generalize and avoid over-fitting to a single distribution. Further, the model was weighted average from the model updates, meaning larger samples had more weight in the global model which resulted in greater stability and robustness.

The results show that federated learning is not only a privacy-preserving method but also a performance-enhancing method. It enhanced the robustness and fairness of models between institutions, while ensuring data privacy. This made the federated recommender system ideal for real-world educational settings, where data is typically distributed, and sensitive.

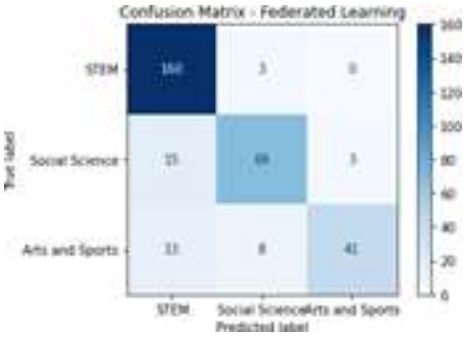

Figure 4.8: Federated Learning confusion matrix

Figure 4.7 showed stronger diagonal dominance, particularly in STEM as compared to figure 4.6. STEM had correct classification of 160, Social science 69 and Arts and Sports 41. Misclassifications reduced overall with only small spill overs between STEM and social sciences 3 cases and between arts and sports and social science 8 cases. Figure 4.7 revealed higher diagonal dominance, especially for STEM, in comparison with figure 4.6. STEM had correct classification of 160, Social science 69 and Arts and Sports 41. The number of misclassifications decreased overall, with only minor numbers of spill overs between STEM and social sciences 3 cases and between arts and sports and social science 8 cases. This suggested that federated learning enhanced generalization and diminished confusion along pathways particularly among STEM students. The federated model achieved better results in terms of overall accuracy and robustness compared to the hybrid Random Forest + Cosine Similarity. The improvement is due to the aggregation of model updates generated by multiple nodes, leading to better generalization and less

overfitting. The Random Forest + Cosine Similarity model ensured interpretability and had good baseline performance, but the federated approach increased precision and recall while maintaining the privacy of data.

d. Privacy, Convergence, Scalability, and Fairness Federated learning preserved privacy by keeping raw student data local. The model converged stably after four communication rounds, scaled easily to new institutions, and improved fairness by incorporating diverse student groups. Trade-offs included communication overhead and reliance on network stability.

Discussion

This study's results showed that cosine similarity was effective in recommender systems, especially content-based recommender systems, as confirmed by other studies. Previous studies like [4] and [22] have shown cosine similarity to be an effective similarity measure between user profiles and item attributes, especially in education recommending applications. However, the limitation observed in this study was that cosine similarity does not capture complex relationships supporting the observations in [4] and [22] , where similarity-based approaches alone were insufficient for predictive tasks. This justified the integration with machine learning model.

When cosine similarity is combined with the Random Forest, it achieves very good performance. This was corroborated by the results of previous research in the area of hybrid recommender systems. Based on their research [11] and [13], hybrid systems using several methods are generally more effective than those using only one, as they benefit from the best of all the methods. In this study, cosine similarity provided alignment and similarity information, while Random Forest captured complex, non-linear relationships in the data. The results were also consistent with findings by Adilaksa and Musdholifah [22], who demonstrated that cosine similarity enhanced recommendation quality when used as part of a broader system. Likewise, [5] pointed out that Random Forest is effective in education-related data with complex data, but it has a drawback of not having similarity measures explicitly. The current study was an attempt to combine both techniques to fill this gap.

Furthermore, the observed improvement in accuracy, precision, recall and F1-score confirmed the importance of feature engineering in machine learning models, as emphasized by Saleem et al. [25], where combining multiple features and models improved predictive performance significantly. Even with the performance boost, the hybrid model remains centralized, however, and that pose privacy issues. This restriction is consistent with what was found in [6] and [8], where centralized systems were found to be susceptible to data privacy issues. This highlighted the need for privacy-preserving approaches such as federated learning.

The results of this study showed that federated learning not only maintained the privacy of the data but also slightly enhanced the performance of the models compared to centralized learning. This aligned with the findings of other studies, like the work by Fachola et al. [6] and Guo et al. [8], which demonstrated that federated learning can achieve similar or even better performance without requiring the centralization of data.

This boost in performance is likely due to the ability of the model to aggregate knowledge across many decentralized data sets, improving its generalization across different populations. However, the improvement in performance, although consistent, was relatively small. This is consistent with the literature, which argues that the major benefit of federated learning is privacy over significant performance improvements. Even if only a minor improvement, however, getting that with maintaining privacy was a big step forward in the field of educational recommender systems.

V. CONCLUSION

This study demonstrated that integrating cosine similarity with Random Forest within a federated learning framework provides accurate, interpretable, and privacy-preserving pathway recommendations for students transitioning under Kenya's Competence-Based Education (CBE) system. Cosine similarity proved effective for mapping student competencies to pathway requirements, while Random Forest captured complex, non-linear relationships. The hybrid model significantly improved accuracy, precision, recall, and F1-score compared to baseline models. Federated learning further enhanced performance while ensuring privacy, scalability, and fairness, making the system suitable for real-world educational deployment.

Recommendations

Adoption in Education: Junior schools, KNEC, and the Ministry of Education should adopt federated recommender systems to balance accuracy with privacy preservation.

Feature Engineering: Machine learning models in education should incorporate similarity-based features to improve predictive performance.

Policy Integration: Educational policymakers should integrate federated learning frameworks into national strategies for data-driven decision-making, ensuring compliance with privacy regulations.

Future Work

Future research should explore extending federated learning to deep learning models for improved accuracy and generalization, and incorporate explainable AI (XAI) techniques to enhance transparency and interpretability of pathway recommendations. Additionally, cross-institutional fairness metrics should be investigated to ensure equitable recommendations across diverse student populations.

Acknowledgment

I extend my sincere gratitude to my academic supervisors whose guidance and intellectual support have been instrumental throughout the development of this work. Their insights challenged my thinking, sharpened my focus and strengthened the ethical foundations of this work.

I am deeply grateful also to my family: my parents, sisters and brothers for their unwavering encouragement, patience and belief in power of education.

Most importantly, I want to thank Almighty God for His steadfast love in my life that has enable me to study well and also for providing me with knowledge and understanding.

REFERENCE

1. V. Chugani, "Federated Learning: A Thorough Guide to Collaborative AI," datacamp.com, 04 october 2024. [Online]. Available: https://www.datacamp.com/blog/federated-learning. [Accessed 17 november 2025].
2. R. Tiwari, "The Integration of AI and Machine Learning in Education and its Potential to Personalized and Improve Student Learning Experiences," International Journal of Scientific Research in Engineering and Management (IJSREM), vol. 07, no. 02, 2023.
3. K. I. o. C. D. KICD, "Basic Education curriculum framework," Kenya Institute of Curriculum Development, Nairobi, Kenya, 2017.
4. Y. L. Sukestiyarno, H. A. Sapolo and H. Sofyan, Application of Recommendation System on E-Learning Platform Using Content-Based Filtering with Jaccard Similarity and Cosine Similarity Algorithms, 2023.
5. Y. Ma and M. Gan, "A Random Forest Regression-based Personalized Recommendation Method," Pacific Asia Conference on Information Systems, p. 170, 2018.
6. A. T. P. B. ,. G. C. E. I. F. Christian Fachola, "Federated Learning for Data Analytics in Education," MDPI, pp. 1234-1243, 2023.
7. S. N. M. J. L. A. R. T. S. L. M. Umer Farooq, "Transforming educational insights: strategic integration of federated learning for enhanced prediction of student learning outcomes," SPRINGERNATURE Link, vol. 80, pp. 16334-16367, 2024.
8. D. Z. D. Song Guo, "Pedagogical data analysis via federated learning toward education 4.0," American Journal of Education and Information Technology, vol. 4, no. 2, pp. 56-65, 2020.
9. H. L. T. W. Y. C. Y. Tiancheng Zhang, "Enhancing Dropout Prediction in Distributed Educational Data Using Learning Pattern Awareness: A Federated Learning Approach," MDPI (Multidisciplinary Digital Publishing Institute), vol. 11, 2023.
10. Y. Koren, R. Bell and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," IEEE, pp. 30-37, 2009.
11. A.-B. Hael, A. M. Abdullateef, R. Awanis and H. Fadhl, "Collaborative filtering recommender system: overview and challenges," Advanced Science Letters, vol. 23, no. 9, pp. 9045-9049, 2017.
12. S. N. Mohanty, J. M. Chatterjee and A. A. E. &. P. G. Sarika Jain, Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural and Other Industries, Wiley-Scrivener, 2020.
13. K. P. a. S. Solanki, "Hybrid Recommender Systems: An Overview," Journal of Advanced Computing, vol. 10, 2023.
14. A. A.-B. &. J. Alsakran, "An Automated Recommender System for Course selection," International Journal of Advanced Computer Science and Applications,, 2016.

15. M. P. O'Mahony and B. Smyth, "A Recommender System for On-line Course Enrolment: An Initial Study," ACM Digital Library, 2007.
16. A. Bozyiğit, F. Bozyiğit and D. K. &. E. Nasiboğlu, "Collaborative Filtering based Course Recommender using OWA operators," ieeexplore.ieee.org, 2018.
17. S. Algarni and F. Sheldon, "Systematic Review of Recommendation Systems for Course Selection," Machine Learning and Knowledge Extraction, vol. 5, pp. 560-596, 2023.
18. M. C. Urdaneta-Ponte, A. M.-Z. 1 and I. Oleagordia-Ruiz, "Recommendation Systems for Education: Systematic Review," electronics, 2021.
19. M. J. Mokarrama, S. Khatun and M. S. Arefin, "A content-based recommender system for choosing universities," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 28, p. 2174–2186, 2020.
20. J. Yang, X. Liu, Z. Zhao and Z. He, "A hybrid recommendation system based on an improved Apriori algorithm and user-based collaborative filtering," IEEE Access, vol. 6, pp. 57712-57722, 2018.
21. C.-H. L.-H. C. Pei-Chann Chang, "A Hybrid Course Recommendation System by Integrating Collaborative filtering and artificial immune system," MDPI, vol. 31, pp. 1517-1527, 2016.
22. Y. Adilaksa and A. Musdholifah, "Recommendation system for elective courses using content-based filtering and weighted cosine similarity," ieeexplore, pp. 51-55, 2021.
23. S. N. L. T. a. O. T. P. Sokkhey, "Multi-models of educational data mining for predicting student performance in mathematics: A case study on high schools in Cambodia," IEIE Transactions on Smart Processing and Computing, vol. 9, pp. 217-229, 2020.
24. W. A. A. a. A. K. H. Ali Salah Hashim, "Student performance prediction model based on supervised machine learning algorithms," IOP Conference Series: Materials Science and Engineering, 2020.
25. F. Saleem, Z. Ullah and B. F. &. F. Kateb, "Intelligent decision support system for predicting student's e-learning performance using ensemble machine learning," p. 2078, 2021.
26. J. L. H. a. S. A. P. Kumar, "A practical model for educators to predict student performance in K-12 education using machine learning," IEEE, vol. 56, pp. 3004-3011, 2019.
27. I. A. A. A. a. A. Y. A. Maghari, "Students performance prediction using KNN and Naïve Bayesian," International Conference on Information Technology, vol. 167, pp. 909-913, 2017.
28. M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza and M. A. M. B. &. S. U. Khan, "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," IEEE Access, pp. 7519-7539, 2019.
29. A. Tarik and H. A. &. F. Yousef, "Artificial intelligence and machine learning to predict student performance during the COVID-19," in Procedia Computer Science, Amsterdam, Netherlands., 2021.