RESEARCH ARTICLE

Check for updates

A Data-Driven Framework for Personalized Learning Using Machine Learning Techniques in Rwanda

Tuyisenge Ephrem[1,*], **Rogers Bhalalusesa**[1] and **Juliana Kamaghe**[1]

[1] Open University of Tanzania, Dar es Salaam, Tanzania

Abstract

Personalized learning has become a popular term in education to address learner diversity and enhance student performance. However, its implementation remains a practical challenge in many developing countries due to the lack of fine-grained learning data and platforms. This study proposes a data-driven machine learning approach for personalized learning using routine administrative education data from Rwanda. The approach is scalable, interpretable, and aligned with existing national education information systems. Following a design science research approach, the study combines unsupervised learner profiling via clustering and supervised performance prediction via regression models. The proposed framework enables differentiated instructional planning and informed policy-making at the system level. Evaluated using administrative data from the Rwandan Ministry of Education, the results show that meaningful learner profiles can be derived from aggregate data, while regression models provide useful directional performance insights despite limited explanatory power. This demonstrates the technical and conceptual feasibility of supporting personalized learning in data-poor education systems without requiring detailed learning interaction data. The framework requires further practical testing and large-scale validation to confirm its impact on learning performance. This study contributes to the Artificial Intelligence in Education field by proposing a context-aware approach that addresses interpretability, scalability, and policy alignment.

Keywords: personalized learning, machine learning framework, learning analytics, administrative education data, Rwanda education system.

1 Introduction

The challenge of diversity in education systems all over the world is a global challenge that needs to be addressed for the achievement of equity, efficiency and quality education that results in improved learner outcomes. Education systems all over the world are currently under pressure to provide quality education in an equitable and efficient manner to diverse learners. Learner diversity has increased in education systems due to differences in prior knowledge, different rates of learning, different types of intelligence, socio-economic status and learning environment etc. Unfortunately, the current instructional practices in our education systems which need improvement especially in low- and middle-income countries are very rigid and do not allow for much variation to meet the diversity

Submitted: 24 February 2026
Accepted: 23 March 2026
Published: 31 March 2026

Vol. 2, **No.** 1, 2026.
10.62762/TEDM.2026.319371

***Corresponding author:**
✉ Tuyisenge Ephrem
tuyisenga@gmail.com

Citation
Ephrem, T., Bhalalusesa, R., & Kamaghe, J. (2026). A Data-Driven Framework for Personalized Learning Using Machine Learning Techniques in Rwanda. *ICCK Transactions on Educational Data Mining*, 2(1), 38–51.

of learners in our classrooms. The non-match between learner diversity and instructional practices is resulting in learning gaps, absenteeism and unequal learner achievement [1].

Personalized learning is one of the innovations to help address the problems facing our classrooms today. Personalized learning refers to learning approaches that use technology to tailor learning experiences in terms of content, speed, and course of study to the individual needs, abilities, and pace of each learner. Advances in the fields of artificial intelligence (AI) and machine learning (ML) as well as learning analytics are enabling the design of scalable learning technologies that can realize the promises of personalized learning [2, 15]. Learning systems incorporating learner modeling technologies that capture the learning behavior of the individual learner as it occurs can be used to support data-driven instructional decision-making.

Personalized learning has made tremendous progress over the last decade and personal learning systems have been widely implemented and rigorously validated in a variety of learning environments. However, most of these implementations are based on the assumption that there is sufficient data available for the analysis required for just-in-time pedagogical decisions. Such data is typically derived from clickstream data, learning management system interaction data or even real-time behavioral data from ITS. Unfortunately, many of the national education systems in Sub-Saharan Africa (SSA) including Rwanda lack the required data infrastructure for such fine-grained data collection. As such, many of the personalization approaches that have been designed for high data availability settings are not applicable to resource-poor settings.

The education sector in Rwanda has been preparing itself to effectively implement technology for data governance. So much data is being collected at the national level by Ministry of Education and at sector level by Ministry of Education. In schools, learners' education administration data is being collected on a daily basis. It remains underutilised and could be leveraged at the national level to support the Personalized Learning ambition.

Personalized learning is increasingly considered as a way to improve learning outcomes for students. It is however unclear how meaningful personalization can be achieved in the context of: - Off-grid, digital resource and data poor environments - Where teachers do not have enough time and competency to utilise personal learning shells for meaningful interaction with learning content. This research investigates the conditions required for personalized learning to be delivered to students using administrative education data and machine learning methods. Rather than relying on interaction data that is often unavailable, our research focuses on developing and applying an effective machine learning method to enable the effective personalization of learning experiences for all students within the context of a national MIS. Specifically the method developed should be scalable, such that it can handle large numbers of students and teachers, the outputs should be interpretable and easy to understand by teachers, and the approach should be compatible with existing, national-level government administered MISes.

Based on the above, the objective of this work is to design and validate a data-driven machine learning framework for the personalized learning system in the context of Rwanda as a case study. This paper focuses on the design and validation of the proposed framework. Therefore, the contribution of this paper is towards the body of work on AI-based education systems with consideration of the state of education systems in developing countries.

In this paper we present three main contributions. First, we introduce a structural model which explains how administrative data about students can be transformed into the so-called learner profiles and learning predictions models using machine learning techniques. Second, we validate and analyze the adequacy of our approach for the administrative needs of a national educational system. Third, we provide useful information for educational policy makers and system developers, with respect to how to implement the necessary technology for "learning in a controlled environment" using only the administrative data that already exist in current educational systems.

2 Literature Review

2.1 Personalized Learning and Learner Diversity

Personalized learning is rooted in the constructivist and learner-centered philosophy of education. The constructivist and learner-centered approach to education assumes that the learner is active and not passive, constructing knowledge from personal experiences and learning environment. Modern views of personalized learning extend these core tenets to incorporate the use of data and technology to deliver personalized learning experiences through dynamic and student-centric models of instruction to reach many learners. Personalized learning has been linked to increased learner engagement, motivation, and achievement. Hattie and Timperley [1] identify personalizing the learning to the needs of the learner as a factor which has a significant relationship with learning.

Personalized learning is still a concept that needs to be developed. Common to all Personalized Learning systems is the construction of adequate learner models in order to provide real personalization. Such learner models contain information about the psychological state of the learner, his learning history, his learning environment etc. - in general all types of learner data. The quality of the learner data used by a learning system has a direct influence on the degree of personalization provided by the system.

2.2 Machine Learning in Education

Machine learning is an area of study and research that was introduced to the educational research community to address learning activities and to analyze learning data such as learner modeling, performance prediction and providing feedback to learners. The majority of studies are based on the use of supervised learning algorithms like regression and classification. In the context of educational systems, these models can be used to forecast student performance, identify at risk students and allow teachers to intervene appropriately [2, 8]. Thus, recently, unsupervised learning approaches such as clustering have also been used to model learner behavior. Clustering is a method of partitioning the data into classes or groups such that similar instances are grouped together and diverse instances are separated.

In 2019 and 2020, a number of systematic reviews were published examining the use of machine learning within the context of education [4, 15]. The reviews highlighted that there are an increasing number of applications of machine learning in education, but also highlighted a number of challenges, such as the lack of interpretability of models; the lack of validation for non-discriminatory learning; and that there is a lack of studies that investigate the robustness of findings. These are generic challenges for the use of machine learning in all domains, but within the educational domain, one of the key challenges that must be addressed is that of interpretability. The reason for this is that teachers, head teachers and policymakers will need to be fully aware of the implications of any learning analytics system before they will be in a position to make decisions that are associated with the outcomes of the system.

2.3 Framework-Based Approaches to Personalized Learning

While the individual models are the main focus of most research, there are numerous studies exploring the frameworks based on integration of data, models and activities/decision processes. Frameworks give a conceptual and operational clarity by explaining data flows within activities and decision processes in models and by explaining how results from analyses are used for learning [7].

Most current frameworks depend on an abundance of learning analytics data as well as on the availability of modern digital learning platforms, which are often not available in many contexts. This is largely due to the current educational system inequality: either there is unequal access to technology, or there is traditional classroom-based teaching.

2.4 Personalized Learning in Developing-Country Contexts

There is limited research on personalized learning in developing countries. While there are a few studies on personalized learning in developing countries, most are based on experiments that are restricted to a small number of schools or classrooms and therefore do not amount to a systemic approach to personalized learning. UNESCO [9], UNICEF [13] and Pedro et al. [14] policy reports on the use of AI in education for developing countries highlight the need for context-specific AI-based innovations that can utilize the existing data resources and address the ethical, governance and capacity challenges.

While less granular than digital learning trace data, administrative education data can be used as an input data source for this type of solution. When combined with appropriate machine learning techniques, the data can be used to classify students, to track student performance and to determine when interventions are required in a large scale learning analytics system.

A number of recent studies and policy reports published in the last year confirm the growing relevance of the use of AI in education in low-income and developing countries. UNESCO published a report on Artificial Intelligence in Education for 2030, which highlights the opportunities and challenges that AI presents in relation to education. While AI has the potential to revolutionise and transform education, Holmes et al. [10] argues that it is crucial to bear in mind the context of each country and therefore tackle the infrastructure and capacity challenges that must be addressed, in order to avoid exacerbating existing inequalities. In 2024 and 2025, UNICEF [11, 12] published reports entitled Applying AI in Education: A Scan of Current and Emerging Use Cases in Developing Countries and Applying AI in Education: Equity and inclusion considerations. The reports note that any educational innovations that rely on the use of AI must be designed to be accessible and do not increase the existing inequalities faced by the poorest and most disadvantaged children.

According to a study published this month in the journal Science, more AI projects are being implemented in low resource settings in order to develop "human-centered, sustainable, and teacher-friendly" tools, rather than attempting to develop fully automated solutions. Recent studies on the use of AI in teaching in developing countries have found that by complementing the work of teachers with technology, it is possible to improve the design and organization of the material to be taught in the classroom, and thus ease the workload of teachers, while ensuring that the content is relevant to the context of the country. This is particularly relevant in countries where access to technology for e-learning is poor, and there is limited access to real-time learning data.

Although the aforementioned researches attempted to meet the need for large scale, system level studies that use administrative data for personalizing learning, there are still many open research questions that need to be explored. Furthermore, although AI researches have shown many promising results in the field of education in developing countries, these achievements are generally not scaled into national education systems. Therefore, the need for large scale, interpretable and policy aligned approaches as suggested in this paper is still valid.

Using artificial intelligence (AI) to create personalized learning experiences for students is claimed to have many benefits, but a closer look at the ethics of learner profiling reveals many concerns, particularly in developing countries. Current AI profiling is based on past performance and administrative data, which is likely to carry over the current inequalities that have been caused by the broader social structures in society. These inequalities may be perpetuated and exacerbated through unequal access to learning resources and activities [9].

Another factor in the debate is algorithmic bias. A biased learning app is one whose instructional algorithms are designed with some categories of students in mind and not with others in mind. So, if the app is designed for all students who are capable of using a tablet, it may not be helpful to students who cannot use a tablet. Bias can arise from a number of places. Data used for learning apps can be biased and so the students that the apps are designed to serve may not be those who are different from the source data. For example, if a learning app is trained using a data set of students that are a random sample of all students in a country, then students who are very different from the average student in that country may be poorly served if the algorithm is not able to recognize or accommodate their differences. It is also possible for bias to arise from unbalanced data sets. An example of this is a literacy app that is designed to assist students who are having difficulty reading and writing. Such an app is by definition biased towards students with literacy difficulties and will be of little use to students who are already literate. This is a particular problem in developing countries where there is often a paucity of high quality learner data. Researchers have also recently identified the potential for apps to propagate existing forms of discrimination, making these discrimination's more entrenched over time.

Another concern revolves around data privacy, transparency and accountability. For example, creating learner profiles can lead to learners being 'labelled' in some way, which can stigmatise those who are in some way different – for example, those with learning difficulties [16].

One of the most pressing challenges in developing countries is the lack of AI governance frameworks, and insufficient institutional capacity to track the effects of new algorithms on society. Our research calls for the development of people-oriented, transparent and ethical AI technologies, methods and practices that can support the development of curriculum, safeguard the rights of children in the digital age and improve the quality of learning outcomes for all.

2.5 Research Gap

We identified a literature gap: While there is a large body of research on the effects of machine learning-based personalized learning systems in high-data contexts, there is much less work on the frameworks required to successfully leverage administrative education data to personalize learning at a national level – particularly in low-data contexts such as many developing countries in Sub-Saharan Africa. There is also limited research into the challenges, considerations and mechanisms involved in designing large-scale, personalizable learning systems that remain interpretable and actionable to policy-makers.

While there has been an increasing number of studies exploring the integration of Artificial Intelligence (AI) and education, empirical studies on the implementation and impact of using AI in education have been mainly carried out in the developed world and based on artificial or simulated datasets. So, the potential of AI in education for developing countries such as Rwanda remains largely unexplored. In this study, we developed and empirically evaluated a data-driven machine learning framework for personalizing learning experiences using administrative education data sourced from Rwanda. Since our study focuses on the development of a framework rather than implementing a learning platform, it is both conceptual and practical in nature.

3 Methodology

3.1 Research Design

Personalization of learning experience is one of the strategic implementations to achieve the success of the learning process. However, currently the learning experience personalization process has not been maximized yet. This study applies Design Science Research (DSR) approach to develop machine learning-based framework for personalizing learning experience. The DSR approach is chosen for this study because the methodology is relevant to research studies that design and evaluate artifacts such as frameworks, models and systems to solve particular problems in practice [5]. In this study, the artifact of interest is a data-driven learning experience personalization framework that can produce learner profiles and learning performance prediction.

Our focus is on the design, development and validation of the concept model with educational data. This is appropriate for our goal of defining the concept of personalising learning within the constraints of a national education system structure and data.

3.2 Data Source and Context

The framework was developed and validated on administrative education data provided by the Rwandan Ministry of Education. Administrative education data is data used for educational management and reporting. Such data is typically collected over time and provides information on the characteristics of students enrolled in various education activities including different categories of basic education. For this study, several annual cohorts of administrative data for basic education were used.

Dataset created for Fall 2023 Data Science competition provided 31,513 records from 41 schools in 6 different districts for two academic years 2023/2024 and 2024/2025. Some variables in the dataset are district id, school id, gender and grade level (mixing both gender and grade level) and academic year and annual marks for students.

We have a range of structured variables for learner demographic attributes, school information (with school ID, grade level and subject combination), as well as information about the year the data was collected. In addition, there is a set of aggregated learning outcome measures. The EMIS data we have is different from learning platform data in that it is more like other sources of data that are typically available at a national level, and it provides a representative sample of the national data. This is an important benchmark for developing scalable personal learning solutions for the Rwandan context as well as for other developing countries.

For building the learning models, the data was preprocessed by removing missing values from the selected attributes. For the attributes having categorical data, encoding was done. For the attributes having numerical data, normalization was done. The attributes required for learner profiling and performance prediction were selected.

The feature set for this case study was derived from the administrative dataset variables. For the learner profiling (clustering) task, the most relevant features were the following: - Aggregate features that describe the learner's performance at a more global level, like the average score. - Subject level performance description when available (e.g. average score in Mathematics). - Grade level. - Contextual features like school id, and academic year.

The features for the academic performance prediction task for supervised learning were selected from a

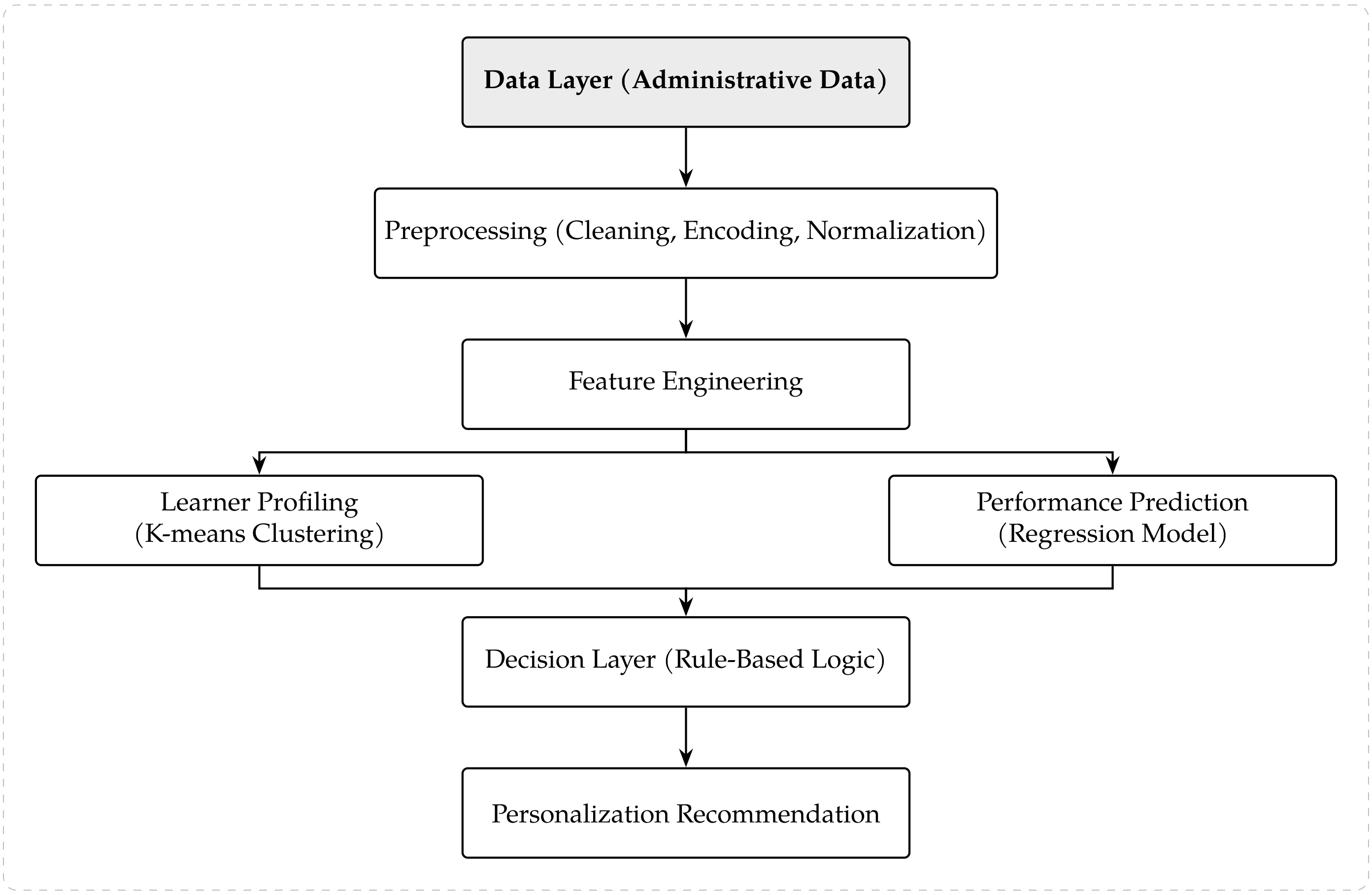

Figure 1. Personalized learning framework architecture.

combination of learner demographic information (including gender where available), academic performance metrics, grade progression metrics and school level metrics. The target variable is the aggregated academic performance score for a given year.

All the categorical variables were engineered with appropriate techniques and all the numerical variables were normalized. The features were selected based on three criteria - educational gain of the selected features, amount of data available in those features and suitability of the features with respect to the clustering and regression algorithms.

3.3 Framework Architecture

As shown in Figure 1, the proposed personalized learning framework is structured into four interconnected layers: data layer, machine learning layer, decision layer, and personalization output layer.

The data layer serves as the foundation of the framework and consists of structured administrative education data. This layer ensures that the framework remains compatible with existing education information systems without requiring additional data collection mechanisms.

The machine learning layer is responsible for transforming raw data into meaningful analytical outputs. It incorporates two core components: learner profiling through unsupervised learning and academic performance prediction through supervised learning. These components operate on engineered feature sets derived from the administrative data.

The decision layer integrates outputs from the machine learning models and applies rule-based logic to translate analytical results into actionable insights. This layer bridges the gap between technical model outputs and educational decision-making.

Finally, the personalization output layer generates learner-specific recommendations, such as targeted academic support, differentiated instructional pacing, or enrichment opportunities. These outputs are designed to support teachers, school leaders, and system administrators in making informed decisions rather than replacing human judgment.

3.4 Learner Profiling Using Unsupervised Learning

Learner profiling was constructed by applying unsupervised clustering to classify learners into similar groups based on their profiles of attributes and performance. Clustering is one of the most widely used methods in data mining for uncovering hidden patterns in a data set especially when there is no prior knowledge or understanding of the data set. It is also applicable in exploratory learner modeling where a large amount of detailed learning behavior data may not be available.

A scalable clustering algorithm was applied to the dataset. The input features for the clustering algorithm were derived from the aggregated learning outcomes per student (per activity) and contextual features. As a result, relevant learner profiles were derived, capturing distinct performance patterns and required levels of support for each cluster.

In this work, K-means clustering algorithm is employed to build the learner profile. K-means clustering algorithm has the advantage of less computational effort, it can handle high dimensional data and can deal with large structured data efficiently. K-means clustering algorithm divides the data into K clusters based on the within cluster variance, so that the homogeneity of clusters are maintained. So K-means algorithm is very useful in the classification of homogeneous learners based on their academic performance and attributes of learning context.

We have employed the K-means method. This choice has been based on three criteria. First, from a computational perspective, an algorithm that can efficiently manage large datasets such as those provided by an administration, should be sought. Second, the simplicity of the obtained results is an advantage: with K-means, not only the assigned clusters but also their centroids can be obtained, and the centroids can be easily translated into the learner profiles that are easily interpretable by the educational community.

For the first phase we examined the possibility of applying alternative clustering methods such as hierarchical clustering and density-based clustering (DBSCAN). Due to the high complexity of hierarchical clustering for large datasets, it was not applied in the first phase of this research. As for the DBSCAN method, it needs more parameter settings than K-means and the results are more sensitive to parameter settings. Another important aspect is that K-means is more efficient in clustering objects with similar variances in data, which is the case for our structured educational data. Hence, K-means is selected as clustering method.

In this step we investigated the quality of the clustering performed by analyzing the values of some internal validation indices. We had already seen the Davies-Bouldin Index which evaluates the relationship between the cluster compactness and their separation. In fact, low values of the index indicate clusters that are highly separated and compact. The cluster interpretability was considered to be fundamental to obtain a meaningful set of profiles that can be really used by the education community. To determine the optimal number of clusters K, we considered some of the common internal validation indices taking into account the need for interpretability of the clusters and at the same time ensuring the best possible performance of the model.

3.5 Academic Performance Prediction Using Supervised Learning

This framework extends the learner profiling activity in iCub by means of a supervised learning block for academic performance prediction. There are two main reasons that motivated the academic performance prediction activity. Firstly, to give early warnings to those learners who are facing poor academic performance and in this way allow the educators to act in a timely fashion. Secondly, to aid the differentiation at pedagogic planning level.

Continuous outcomes such as the scores obtained at the academic achievement tests were estimated using regression models. Linear and penalized models were chosen for their interpretability, robustness and because they represent the gold standard for a large class of structured data such as administrative data and because they are a simple way to avoid over-explaining the data, especially in the presence of correlated predictors.

Here is the process of training a model. Our data is split into training set and test set. We are tuning the model parameters.

Data is divided into training and testing datasets using an 80/20 split where 80% of the data is used for training the model and 20% for testing. It's a random split so both datasets retain the characteristics of the full data set.

To get a robust and reliable evaluation of the model, k-fold cross-validation (k = 5) was applied during the model development. The training data was split into 5 folds. In each iteration of the cross-validation process, the model was trained on 4 folds and evaluated on the 5th fold. The final model performance was then evaluated on the test set.

Cross-validation with a hold-out test set is used to avoid over-fitting the training data and to verify that the performance metrics calculated are reliable and indicative of how well a model will perform on a new set of data.

Performance was assessed using standard regression metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). These metrics provide complementary perspectives on prediction accuracy and model explanatory power.

4 Findings

4.1 Framework Implementation Overview

This research proposes a new framework to support Personalized Learning by integrating Learner Profiling and Performance Prediction into one framework. Using real world administrative data from a university, the data is transformed into a feature space that is then streamed into a data processing pipeline. The pipeline has two main components – an unsupervised Learner Profiling component and a supervised Performance Prediction component.

Table 1. Summary of dataset characteristics.

Attribute	Description
Number of Learners	31,513
Number of Schools	41
Number of Academic Years	2 (2023/2024 and 2024/2025)
Key Features Used	Academic scores, grade level, school
Data Type	Structured administrative data

As summarized in Table 1, the dataset comprises 31,513 learners across 41 schools over two academic years. This Implementation Activity has demonstrated the re-use of administrative education data for learning personalisation, demonstrating the potential to derive data from sources which do not necessarily require learner participation. In this instance, the results of the models derived from the data are presented in formats which allow re-use of the data in the models and also potentially re-use of the derived results in other applications.

4.2 Learner Profiling Results

4.2.1 Cluster Formation and Validation

This learner profiling component performed clustering on the learner population by means of unsupervised learning methods and obtained four clear-cut clusters. The Davies–Bouldin Index which is a popular validity index for cluster evaluation has a value of about 1.88. It calculates the similarity between each cluster and evaluates the separateness and the cohesiveness of the clusters. Therefore it means that the clustering algorithm is able to divide the learners into meaningful clusters.

Table 2. Learner clustering results and characteristics.

Cluster	Learner Profile	Avg. Score Range	Key Characteristics
C1	Remedial	0 – 40	Low performance, high support required
C2	Developing	40 – 60	Moderate performance, improving learners
C3	Proficient	60 – 80	Meets expected standards
C4	Advanced	80 – 100	High performance, enrichment needed

Table 2: The 4 learner profiles were obtained by clustering the data at the different achievement levels. From Table 2 we can see that each cluster is clearly separated by the score ranges and learning behaviors which means that we have successfully clustered the data for each learner profile.

In order to measure the stability of the resulting clusterings, the K-means algorithm was run many times with different sets of initial locations and all the resulting clusterings were very close to the one derived by using the quality index with the centroid-based approach. One can easily derive that all meaningful runs have resulted in the same cluster structure, which consists of about 4 clusters. There is little variation among the centroids of the clusters, and very small variation in the cluster assignments. Therefore, the 4 clusters are believed to be the true underlying structure of the data.

Cross-validation of subsamples Further cross-validation was performed by reapplying the clustering algorithm to subsamples of the data. This was achieved by dividing the full data set into subsamples (subsets) at various levels (i.e. the number of subsamples varied). Irrespective of the number of subsamples applied the cluster patterns were confirmed to be stable regarding the performance classification and the corresponding learner profiles derived from the cluster analysis.

4.2.2 *Interpretation of Learner Profiles*

Each cluster was examined in terms of its central tendency and dispersion of academic performance. Based on these characteristics, the clusters were interpreted as representing four learner profiles:

- **Remedial learners**, characterized by consistently low academic performance and a need for intensive instructional support.
- **Developing learners**, demonstrating moderate performance with signs of progression but requiring targeted reinforcement.
- **Proficient learners**, meeting expected academic standards and benefiting from structured consolidation activities.
- **Advanced learners**, exhibiting high academic performance and suitability for enrichment and accelerated learning opportunities.

The interpretability of these profiles is a key outcome of the framework. Rather than producing opaque numerical groupings, the learner clusters align with categories commonly used in educational practice, enhancing their usability by teachers and school administrators.

4.3 Academic Performance Prediction Results

4.3.1 *Model Training and Evaluation*

This section of the assignment is to do with the prediction of the annual educational achievement of learners based on a set of administrative variables. It is for the supervised learning part of the assignment. The performance of the model on the test set was evaluated. The average Mean Absolute Error (MAE) for the test set was approximately 9 marks and the average Root Mean Squared Error (RMSE) was approximately 27.55 marks There was a moderate margin of error on average.

Table 3. Table of performance metrics for our prediction model. The low R^2 value suggests that our model explains a very small proportion of the variation in the data. Although this is concerning, we note that our model was designed to predict individual student learning outcomes based on highly aggregated administrative data, so it is likely to have only limited precision in this respect, even if the MAE and RMSE give some suggestion of reasonably good accuracy at the level of individual data points.

Table 3. Performance evaluation metrics of prediction model.

Metric	Value	Interpretation
Mean Absolute Error (MAE)	9.08	Moderate average prediction error
Root Mean Squared Error	27.55	Higher penalty on large prediction errors
R-squared (R^2)	0.02	Low explained variance

The stability of the model was assessed by investigating the variability of performance metrics across the cross-validation splits. The Mean Absolute Error (MAE) has a mean value of 9.08 $\pm$ 1.12. The Root Mean Squared Error (RMSE) has a mean value of 27.55 $\pm$ 3.45.

Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were calculated for each of the 1000 stratified splits of the dataset and 95% confidence intervals were calculated for these metrics. The mean absolute error is calculated to be in the interval [7.90, 10.26] and the root mean square error is calculated to be in the interval [24.10, 31.00].

4.3.2 *Interpretation of Predictive Outputs*

The model was not a particularly good fit to the data but the model predictions were still very useful. The predicted performance values were able to provide a target for students who were not performing well and at risk of not meeting the performance criteria. In conjunction with the student profiles obtained from the clustering procedure, the model predictions were used to guide teaching strategies for each of the clusters of students.

A class of students who were assigned to the Remedial or Developing performance prediction categories with lower predicted performance levels were selected for a remedial intervention. Conversely, students assigned to the Advanced performance prediction category with the higher predicted performance levels were selected for an enrichment intervention. This example shows that different data sources can be used to support different types of intervention in a learning analytics environment.

4.4 Integrated Personalization Outputs

For this iteration, the stage that produces Personalized Learning Recommendations received translated analytical output and used a rule-based decision-making process to generate learning recommendations for each learner. The decision-making process in this stage connected learner characteristics and predicted performance with specific actions such as remediation, mastery activities, on-grade level pace activities, or enrichment activities.

Table 4. Mapping of learner profiles to personalized interventions.

Learner Profile	Predicted Performance	Recommended Action
Remedial	Low	Intensive remedial support
Developing	Moderate	Targeted reinforcement
Proficient	Average	Standard progression
Advanced	High	Enrichment and acceleration

Table 4 illustrates how learner profiles and predicted performance levels are translated into actionable educational interventions. This mapping demonstrates the practical applicability of the framework in supporting data-driven instructional decision-making.

4.5 Summary of Key Findings

The findings indicate that the developed framework successfully operationalizes personalized learning using administrative education data. Key outcomes include:

- the identification of interpretable learner profiles through unsupervised clustering,
- the generation of actionable performance predictions despite data constraints, and
- the integration of analytical outputs into coherent personalization recommendations.

Together, these findings demonstrate the feasibility of implementing a machine learning–based personalized learning framework at the system level within the Rwandan education context. The results presented in Tables 2, 3 and 4 collectively demonstrate the effectiveness of the framework in generating interpretable learner profiles, predictive insights, and actionable personalization strategies.

5 Discussion

5.1 Interpreting Learner Profiling Outcomes

From the learner profiling results it can be seen that the unsupervised machine learning is capable of producing meaningful learner profiles from the administrative education data. The four profiles that are obtained viz. Remedial, Developing, Proficient and Advanced Learners are similar to the performance groups that are traditionally created by educators to plan and deliver instruction [3].

Whereas most of the literature on the topic focuses on the low level details of how a student interacts with a learning resource, our profiling component focuses on higher level information, such as the aggregate knowledge of the students at different levels of granularity, as well as contextual information. The Davies-Bouldin Index achieved in our experiments demonstrates that it is possible to effectively distinguish students with a relatively small number of features, which is in agreement with other works that suggest that there is no need to monitor or track students at a very fine level of detail [8].

5.2 Significance of Performance Prediction Results

The resulting model barely improved on the null model in terms of fit (low R^2). Although it is common to associate low R^2 values with poor models, the R^2 is a metric more suitable to other fields of study than education and, in particular, not to the use of administrative data in education. The input variables for any model of student learning are obviously known such as the student's previous achievements. However, there are also many unknown factors that can affect the learning of students such as their socio-economic status, their level of engagement with the material, the pedagogical practices employed by the educators, the condition of the infrastructure of the school and so on.

Our model is not intended to be used for accurate score estimation but rather to provide early indication of the learning direction. The MAE and RMSE values suggest that the model provides a reasonable estimate of the learner performance that can be used for early warning of at-risk learners. This is in line with other learning analytics studies that have argued that decision support is more important than striving for high levels of predictive accuracy [6].

5.3 Value of Combining Profiling and Prediction

One of the main contributions of the proposed framework is the integration of learner profiling and performance prediction into a unique system. Although clustering and prediction models are individually meaningful, their integration through profiling and prediction models provides more valuable information for the identification of learner needs. Profiling expresses the relative position of the learner with respect to the cluster, while prediction expresses the future academic performance of the learner with respect to the threshold values.

Complementary to the clustering and regression classification approaches in this study, more advanced personalisation methods can be developed. Therefore, two students that have the same regression classification results may be assigned to different clusters and consequently receive different teaching interventions. This multi-layered personalisation is in line with the recommendations in the literature, for instance by Klašnja-Milićević et al. [7] who propose multi-dimensional learner models based on learning framework theory.

5.4 Implications for Personalized Learning in Rwanda

Our findings have important implications for the Rwandan education sector. First, our study demonstrates that using the current administrative data infrastructure it is possible to implement a system that delivers personal learning to every student without the need for a significant investment in technology infrastructure that is commonly associated with digital learning in developing countries where the default learning environment is still the classroom.

The second innovation of the framework is the system-level personalization model, which is a complement to the widely practiced teacher-driven differentiation. This is not meant to replace the role of the teacher. Rather, the purpose is to prepare and assist the teacher with relevant data, in order to ensure that the right interventions and resources are delivered to the learner. This is in line with the UNESCO guidelines for the use of AI in education, which advocate for a human centric approach, referred to as augmentation, rather than automation.

5.5 Contribution to AI in Education Research

Our study is one of the first works that investigates context-aware AI in the context of an education sector of a developing country. This work not only contributes to the context-aware AI in education research, but it is also the first work that specifically focuses on administrative data and large-scale technology deployment in the education sector. In contrast to personalization works that have been done mostly in constrained environments with rich data, we study large-scale, interpretable and sustainable personalization solutions that can benefit developing countries.

This paper also discusses the contribution of framework-based research. This includes assessing the level of architectural clarity and practical applicability of the frameworks as well as generally commenting on the fact that these frameworks are largely defined at the level of technical models, and thus, contribute to the larger discussion of how the results of the research can be translated into policy and practice by education systems.

5.6 Limitations and Future Directions

Although the framework proposed in this paper offers a general guideline for designing advanced learning environments and shows the feasibility of using aggregated administrative data to model individual learning behavior, a number of constraints and limitations need to be mentioned. Firstly, using aggregated administrative data at the learning environment level may result in a loss of learner model granularity, and consequently a loss of predictive accuracy for the learning trajectory of the individual learner. Secondly, the framework proposed in this paper has not made use of the real-time learner learning behavior and affective cues.

Further work is needed to investigate the value of additional data sources such as formative assessment results, classroom observations and learning management system data where this is available. A longitudinal study is also needed to investigate the value of the framework in supporting interventions involving the use of learning pathways.

6 Conclusion and Recommendations

6.1 Conclusion

Title of the project: Machine learning approach for Personalized Learning (PL) in Rwanda using administrative education data Research Summary This research applied machine learning algorithms to develop a Personalized Learning (PL) framework and validated the feasibility of the approach using administrative data from the Rwandan education sector. The developed framework integrates learner profiling and performance prediction and this research outcomes should be considered as a Proof of Concept (PoC) or functional prototype.

Our research discovered that it is possible to use administrative data to support learner differentiation, early intervention and learner at risk intervention using data analytics. That being said, the framework has not been tested within an operational learning environment, and it remains to be seen if it will have a tangible on improving student outcomes.

From a more practical perspective, this framework may serve as a foundation for the further development and implementation of AI in the national education systems. It is designed to be scalable, interpretable and adaptable to future adjustments in conjunction with the education community. Further research is needed to determine the feasibility of this framework in practice, and to evaluate its validity, applicability, and ethics.

6.2 Contributions of the Study

Our research contributes to the field in three distinct ways. Firstly, this paper introduces a context-aware learning framework that is appropriate for the low data point educational scenario. Secondly, our work is the first to investigate the validity of using administrative data from large-scale education systems to build learner profiles and models for predicting learner progress at the system-level. Thirdly, our research falls into the broader body of work on the application of AI in education and it sets itself apart from previous works by discussing, in addition to performance, the challenges of balancing interpretability, scalability and policy relevance — all important factors for the adoption and sustainability of AI and learning frameworks in education.

A large number of papers in this volume are output of the ML4Ed Workshop held in Hyderabad, India. These papers are related to the theme of how the Machine Learning for Education (ML4Ed) framework can be used to help address the challenge of translating research on learning systems into practical implementations, particularly in developing countries.

6.3 Implications for Policy and Practice

This study has implications for Education Policy and Practice in Rwanda. For policy, it shows that the current infrastructure of data in the education sector can be easily repurposed for the development of adaptive learning platforms that do not require capital intensive investments. Therefore, the Ministry of Education should consider incorporating the proposed framework for the design of learning pathways into the National Education Management Information System (NEMIS).

This Framework can be used to identify learning areas which require targeted intervention at the School and Classroom level. It can help educators to: Identify learners with similar needs who will require similar levels of support Identify areas where learners require the greatest level of support The Framework does not prescribe specific actions for teachers to take. Rather it provides data which can be interpreted in a way which is relevant to the context of each individual teacher.

6.4 Recommendations

Based on the findings, the following recommendations are proposed:

1. Education System Recommendations Educational authorities should implement personal learning plans for all students in Rwanda through the machine learning personalization frameworks developed in the scope of the project. In addition, educational management system, which has been functioning in the education system, should be used to monitor students through the system and prepare support plans.

   Train teachers, educators and other stakeholders on interpreting the data and how it can be utilized in their work.

   Train teachers and educators on using data and analysis to develop curriculum.

2. While administrative data was a great starting block for our digital portfolios, we would like to include other forms of data in the future, such as formative assessments and classroom walk-throughs, in order to make the digital portfolio more student-centric.

3. Schools, education institutions, Governments and organisations are invited to implement appropriate policies and governance mechanisms to ensure the ethical development and use of technology, provide transparency, accountability, adequate pedagogical engagement for learners and promote human responsibility.

4. A small number of pilot schools or colleges could be given priority to act as trial sites to test whether the concept works, and to establish the practicality, ease of implementation and impact on teaching practices.

6.5 Directions for Future Research

Future research should extend this work by evaluating the impact of framework-driven personalization on learner outcomes over time. Longitudinal studies examining how educators use personalization insights and how such use influences instructional practices would provide valuable evidence for large-scale adoption. Additionally, comparative studies across different education systems could further validate the framework's generalizability and adaptability.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that generative AI (ChatGPT-5) was used exclusively for language editing and grammar refinement throughout the manuscript. The use of AI did not influence the study design, data analysis, interpretation of results, or conclusions. The authors take full responsibility for the originality, accuracy, and integrity of the content.

Ethical Approval and Consent to Participate

Not applicable. This study used only secondary, de-identified administrative education data provided by the Rwandan Ministry of Education. No primary data were collected from human participants, and no identifiable personal information was accessed or analyzed.

References

[1] Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. [CrossRef]

[2] Baker, R. S., Martin, T., & Rossi, L. M. (2016). Educational data mining and learning analytics. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, 379-396. [CrossRef]

[3] Bodily, R., & Verbert, K. (2017). Review of research on student-facing learning analytics dashboards and educational recommender systems. *IEEE Transactions on Learning Technologies*, *10*(4), 405-418. [CrossRef]

[4] Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, *8*, 75264–75278. [CrossRef]

[5] Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105. [CrossRef]

[6] Ifenthaler, D., & Yau, J. Y. K. (2020). Utilising learning analytics to support study success in higher education: a systematic review. *Educational technology research and development*, *68*(4), 1961-1990. [CrossRef]

[7] Klašnja-Milićević, A., Vesin, B., Ivanović, M., & Budimac, Z. (2016). E-learning personalization based on hybrid recommendation strategy and learning style identification. *Computers & Education*, *56*(3), 885–899. [CrossRef]

[8] Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12–27. [CrossRef]

[9] UNESCO. (2021). *Artificial intelligence and education: Guidance for policy-makers*. UNESCO Publishing. [CrossRef]

[10] Holmes, W., & Miao, F. (2023). *Guidance for generative AI in education and research*. Unesco Publishing. [CrossRef]

[11] UNICEF. (2024). *The State of the World's Children 2024: The future of childhood in a changing world*. UNICEF.

[12] UNICEF. (2025). *Digital education strategy 2025–2030*. UNICEF.

[13] UNICEF. (2023). *Guidance on artificial intelligence and data for children in education systems*. UNICEF. Retrieved from https://www.unicef.org/globalinsight/reports/ai-children

[14] Pedro, F., Subosa, M., Rivas, A., & Valverde, P. (2019). Artificial intelligence in education: Challenges and opportunities for sustainable development.

[15] Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators?. *International journal of educational technology in higher education*, *16*(1), 39. [CrossRef]

[16] Luckin, R. (2008). The learner centric ecology of resources: A framework for using technology to scaffold learning. *Computers & Education*, *50*(2), 449–462. [CrossRef]

Tuyisenge Ephrem is a PhD candidate in Information Technology at the Open University of Tanzania. He holds a Master of Science degree in Information Technology of Mount Kenya University (MKU) and a Bachelor's degree in Computer Science of the University of Rwanda (UR). His research focuses on machine learning applications for personalized learning, digital education systems, and data-driven decision-making in education. His research interests include artificial intelligence in education, educational data analytics, e-learning systems, and education management information systems (EMIS). He has extensive professional experience working with international organizations, including UNICEF, VVOB and national education institutions in Rwanda. (Email: tuyisenga@gmail.com)

Rogers Bhalalusesa is a Lecturer and Programmer at the Open University of Tanzania with extensive experience in information systems design, software development, and data-driven technologies. He holds a PhD in Management Information Systems from the University of Science Malaysia, an MSc in Computer Science from the University of Kent, United Kingdom, and a BSc in Computer Science from the University of Dar es Salaam, Tanzania. He is also a Certified Cisco Network Associate (CCNA).

Dr. Bhalalusesa specializes in artificial intelligence, educational data mining, and learning analytics, with a strong focus on developing intelligent and adaptive e-learning systems. He has extensive teaching experience in programming, artificial intelligence, and data science, delivering courses in Java, Python, and C at both undergraduate and postgraduate levels. His current research interests include machine learning applications in education, natural language processing, and intelligent assessment systems. (Email: rogers.balalusesa@out.ac.tz)

Juliana Kamaghe Head of Department-Mathematics and ICT, Faculty of Science, Technology and Environmental Studies. Dr. Juliana Kamaghe is a Tanzanian lecturer and researcher specializing in computing, educational technology, AI in education, and assistive technologies for inclusive learning.

She is a Lecturer and Head of the Department of Mathematics and Information and Communication Technology at The Open University of Tanzania (OUT), Dar es Salaam. She obtained her PhD in Information Communication Science and Engineering (with a focus on Electronics and Telecommunications) in 2021 from the Nelson Mandela African Institution of Science and Technology (NM-AIST). She previously earned an MSc in Electronics and Telecommunication Engineering from ISEP, France, in 2012, and a BEng (Hons) from Dar es Salaam Institute of Technology in 2007. (Email: email@email.com)