USING MACHINE LEARNING AND AGENT-BASED SIMULATION TO PREDICT LEARNER PROGRESS FOR THE SOUTH AFRICAN HIGH SCHOOL EDUCATION SYSTEM

M. van den Heever[1*], E. Becker[2], L. Venter[2] & J. F. Bekker[1]

ARTICLE INFO

Article details
Presented at the 34th annual conference of the Southern African Institute for Industrial Engineering, held from 14 to 16 October 2024 in Vanderbijlpark, South Africa

Available online 29 Nov 2024

Contact details
* Corresponding author
maymarievdheever@gmail.com

Author affiliations
1 Department of Industrial Engineering, Stellenbosch University, South Africa

2 Department of Logistics, Stellenbosch University, South Africa

ORCID® identifiers
M. van den Heever
https://orcid.org/0000-0002-4156-7199

E. Becker
https://orcid.org/0009-0008-3200-2613

L. Venter
https://orcid.org/0000-0002-1529-9784

J. F. Bekker
https://orcid.org/0000-0001-6802-0129

DOI
http://dx.doi.org//10.7166/35-3-3080

ABSTRACT

The South African high school education system faces numerous challenges, including high dropout rates and unequal educational outcomes, which call for innovative methods to analyse and address these problems. This study uses an integrated approach that merges machine learning and agent-based modelling to simulate learner progression in public high schools, illuminating the critical factors that influence educational outcomes. Using data from the 2019 General Household Survey in South Africa, factor analysis is first conducted to identify and quantify the principal characteristics that define learners. These identified features then train an XGBoost machine-learning model, which is integrated with an agent-based framework to simulate learner progression from Grades 8 to Grade 12. Validating the model against the learner unit record information and tracking system dataset results in a root square error of 2.94%, which is indicative of the model's ability to predict learner progression. As a result, the simulation model functions as a strategic platform for evaluating and refining educational interventions.

OPSOMMING

Die Suid-Afrikaanse hoërskoolonderwysstelsel staar talle uitdagings in die gesig, insluitend hoë uitvalsyfers en ongelyke onderwysuitkomste, wat innoverende metodes vereis om hierdie probleme te ontleed en aan te spreek. Hierdie studie gebruik 'n geïntegreerde benadering wat masjienleer en agent-gebaseerde modellering saamsmelt om leerdervordering in openbare hoërskole te simuleer, wat die kritieke faktore wat opvoedkundige uitkomste beïnvloed, belig. Deur gebruik te maak van data van die 2019 Algemene Huishoudelike Opname in Suid-Afrika, word faktorontleding eers gedoen om die hoofkenmerke wat leerders definieer te identifiseer en te kwantifiseer. Hierdie geïdentifiseerde kenmerke word dan gebruik om 'n XGBoost-masjienleermodel op te lei, wat geïntegreer is met 'n agentgebaseerde raamwerk om leerdervordering van graad 8 tot graad 12 te simuleer. Validering van die model teen die leerdereenheidrekordinligting en opsporingstelseldatastel lei tot 'n vierkantwortelfout van 2.95%, wat 'n aanduiding is van die model se vermoë om leerdervordering te voorspel. Gevolglik funksioneer die simulasiemodel as 'n strategiese platform vir die evaluering en verfyning van opvoedkundige intervensies.

1. INTRODUCTION

Education stands as a cornerstone of societal development, fuelling economic growth and fostering innovation [1]. In pursuing this, it is crucial to understand the underlying factors that influence educational outcomes in order to enable targeted interventions and strategic improvements. In South Africa, the legacy of Nelson Mandela looms large, highlighting his belief that education is the most powerful tool for societal transformation. Governed by the Department of Basic Education, the public education system is designed to be inclusive, with no-fee schooling provided to learners in the lowest three quintiles of the five-quantile classification of schools.

South Africa allocated the second highest proportion of its total government expenditures on education among 44 countries in 2019, as noted by [2, 3]. Despite this substantial investment, the system continues to face problems such as overcrowded classrooms and insufficient resources [4, 5]. The deficiency in the standard of education is evident in both national and international assessments, which reveal that numerous learners fail to acquire the fundamental literacy and numeracy skills appropriate for their grade levels [6, 7, 8]. The consequences of these educational shortcomings are significant, contributing to high youth unemployment rates and restricting their opportunities for social and economic advancement [3, 9].

This study uses machine learning and agent-based modelling, grounded in industrial engineering principles, to simulate the educational progression of public high school learners in South Africa. The objectives include identifying the key factors influencing learner outcomes, using advanced simulation techniques, and validating the model's effectiveness in guiding the selection of interventions. If successful, this model could serve as a valuable analytical tool for strategising educational improvements, given the challenges and opportunities in the South African educational system.

To simulate learners' progression through the public high school system, this study uses a multifaceted approach that combines factor analysis, machine learning, and agent-based modelling. The input data is sourced from the 2019 General Household Survey, a comprehensive national survey that collects information from thousands of households across all nine provinces. The rationale for selecting the 2019 data is twofold: the 2023 data lacks the comprehensive factors that were present in previous years, and the data from 2020 to 2022 is influenced by COVID-19-related anomalies. Therefore, the 2019 data was chosen to provide a more stable and representative analysis.

With this data, factor analysis is initially applied to identify key variables that define a learner. These identified variables serve as input features for training a XGBoost machine learning model. This model is then integrated into the AnyLogic simulation environment, enhancing the agent-based model's ability to predict and simulate the realistic behaviours of learners. Finally, the agent-based model uses these predictions to simulate the educational progression of each learner—whether they advance to the next grade, repeat the same grade, or drop out of school. The simulation incorporates the rule stipulated by national education policy, which is that a learner may only repeat a grade once in any of the four educational phases: Foundation, Intermediate, Senior, and Further Education and Training [10].

The paper is structured as follows: the literature pertaining to the field of study is presented next, followed by a description of the data preparation process. The methodology is presented in Section 4, discussing the combination of factor analysis, machine learning, and agent-based modelling. Model validation is described in Section 5, followed by some results that are presented in Section 6. Finally, the conclusions are presented, and future work is discussed.

2. LITERATURE REVIEW

This section reviews the literature that is relevant to the simulation of educational systems. An overview of both related studies and various simulation techniques is presented, providing a context for the current study, and demonstrating the relevance of the techniques that are adopted.

2.1. Related work

Educational systems are complex, and simple cause-and-effect approaches often fall short when predicting outcomes. Table 1 illustrates past research that aimed to simulate these complexities, detailing the techniques they used, the results achieved, and the contexts in which these studies were conducted.

Table 1: Related work

Theme	Purpose	Method	Findings
Student progression in universities [11]	Analyse the influence of diversity distributions on graduation in university enrolment.	System dynamics Quantitative simulation	Found no significant improvement in graduation diversity by increasing first-year enrolments of under-represented groups. Requires changes in pass rates among these groups to have a meaningful influence.
Agent-based modelling in education [12]	Examine educational outcomes by simulating individual behaviours in primary schools.	Agent-based modelling	Model helps in understanding the complex dynamics of educational interactions and their impact on learning outcomes, suggesting that personalised interventions could improve educational achievement.
Teacher training and educational outcomes [13]	Assess the influence of different teacher training programmes on educational outcomes.	System dynamics Qualitative data analysis	Highlights the critical role of teacher training quality and methodology on educational outcomes, recommending enhancements to training programmes to improve teacher effectiveness.
Systems perspective of basic education [14]	Explore dynamics in basic education systems to identify leverage points for policy intervention.	System dynamics Policy analysis	System dynamics approach revealed key factors influencing educational quality and outcomes, emphasising the importance of systemic changes over isolated interventions.
Overview of available educational datasets [15]	Provide a comprehensive overview of the key datasets available on education in South Africa, and their uses.	Data analysis Policy impact analysis	Summarises important datasets, offering insights into their applications for policy-making and educational research. Indicates gaps in data collection, and suggests areas for improvement.

2.2. Synthesis and gap identification

From the studies that have been reviewed, several key insights emerge. Educational systems involve complex interactions and require sophisticated modelling techniques such as system dynamics and agent-based modelling to understand and predict outcomes. Teacher training quality significantly affects educational outcomes, underscoring the need for improved training programmes. System dynamics can uncover critical leverage points in educational systems, which are crucial for effective policy interventions. Existing datasets provide valuable insights, but also highlight significant gaps, especially in tracking long-term educational progress and outcomes.

Despite the valuable contributions of these studies, a critical gap remains in understanding the progression of learners in South Africa, particularly at the high school level, where dropout rates are most significant [5]. Existing research does not sufficiently address the multi-faceted reasons behind high school failure rates, nor does it provide comprehensive strategies for improving retention and pass rates. Several factors contribute to high school dropouts, including family responsibilities, economic pressures, and a lack of quality education, which compel learners to leave school prematurely [16]. Given the troubling trend of about 50% of learners being lost from the system before reaching Grade 12, with the majority exiting during Grades 10 and 11 [17], this study focuses on high school progression, pinpointing this as a critical area of concern in the South African educational landscape.

2.3. Simulation techniques

This section describes various simulation techniques that could be used to study educational progression. Table 2 explains methods such as discrete event simulation, system dynamics, Monte Carlo simulations, agent-based modelling, and predictive analytics with machine learning models from previous studies.

Table 2: Simulation techniques

Technique	Description	References
Discrete event simulation	Models the system as a sequence of discrete events. Each event occurs at a specific time and changes the state of the system.	[18] [19]
System dynamics	Uses stocks, flows, feedback loops, and time delays to model the overall behaviour and dynamics of an education system, emphasising long-term trends and impacts of policies.	[20] [21] [22]
Monte Carlo simulation	Relies on repeated random sampling to predict the probability of different outcomes; particularly effective in assessing the impacts of varied educational interventions under uncertainty.	[23] [24] [25]
Agent-based modelling	Each individual entity, or 'agent', is modelled with distinct characteristics. Agents interact within a system, leading to emergent behaviours, and allowing for the detailed exploration and analysis of complex systems.	[26] [27] [28]
Machine-learning models	Uses historical data and machine-learning techniques to predict educational outcomes. Techniques such as regression, decision trees, and neural networks analyse various inputs, including socio-economic data.	[29] [30] [31]

The decision to use both machine learning and agent-based modelling in this study stems from their complementary strengths in simulating complex systems such as educational progression. Machine learning provides powerful tools for recognising patterns in large datasets, enabling the prediction of outcomes based on historical data [32]. This capability is particularly useful for understanding the factors that influence learner success and for identifying potential interventions [32]. Techniques such as regression, decision trees, and neural networks can analyse inputs ranging from individual learner characteristics to broader socio-economic factors, thus providing insights that are critical for strategic planning in education [29, 31].

Agent-based modelling allows for the simulation of individual behaviours and interactions in an educational setting. Each learner or 'agent' can be modelled with unique characteristics and decision-making processes, which interact with the system and with other agents [26, 27]. This method is invaluable for examining how individual differences and interactions could lead to emergent behaviours in the educational system [32]. It helps to understand the micro-level dynamics that contribute to the overall educational outcomes, such as the impact of peer influence, teacher-learner interactions, and personalised educational pathways [28].

Combining machine learning with agent-based modelling allows for an integration of macro-level insights from machine-learning predictions with micro-level understanding from agent-based scenarios. This synergy is supported by recent studies that emphasise the value of combining predictive analytics with simulations to understand complex systems better and to plan effective interventions [33].

3. DATA PREPARATION

This section details the process of preparing data for the machine learning and agent-based simulation model. It starts with a description of the General Household Survey, covering its scope, methodology, and relevance to the study. It then explains the application of factor analysis to the General Household Survey data to identify key features that characterise a learner. Finally, it outlines the process of defining the target variable by tracking learner progression between 2019 and 2020.

3.1. General Household Survey

The General Household Survey is an annual survey conducted by Statistics South Africa. Initiated in 2002, the General Household Survey aims to provide a nuanced picture of household living conditions, aiding the government and researchers in tracking progress towards development goals, identifying emerging issues, and tailoring policies to meet the population's needs better. To achieve this, the survey uses a stratified random sampling method, ensuring that it represents the national population. About 30,000 households are selected from all nine provinces, capturing diverse backgrounds to reflect a broad spectrum of socioeconomic statuses, age groups, and geographic locations.

The General Household Survey covers a wide range of topics, including education, health, housing, access to services, employment, and income. The questions are designed to capture both objective data, such as the number of rooms in a dwelling or the highest level of education attained, and subjective perceptions, such as satisfaction with health services or perceived neighbourhood safety. In the realm of education, the General Household Survey collects crucial data on school attendance, educational attainment, and access to educational resources. It also examines barriers to education, such as financial constraints, distance to schools, and household responsibilities that may impede regular school attendance.

3.2. Factor analysis

Factor analysis is a statistical technique that is used to simplify complex datasets by identifying underlying patterns and reducing them to a smaller number of factors. The 2019 General Household Survey, with its 356 questions, posed significant difficulties for trend analysis owing to the sheer volume of data. Factor analysis was used to reduce the dataset's complexity, eliminating unnecessary variables and enhancing the clarity of insights. This technique, commonly applied in fields such as psychology, economics, medicine, and the social sciences, helps to elucidate key constructs and themes in large datasets. For instance, in a psychological study, factor analysis might distil multiple observed variables into a single factor such as 'extraversion', encompassing traits such as 'enjoys social gatherings' and 'feels energised by crowds'.

The General Household Survey is organised into six sections that cover topics such as education, welfare, household specifics, health, social security, and housing conditions. Factor analysis was applied to each section individually, ensuring that the most relevant variables were retained. As detailed in Table 3, this process identified 15 factors encompassing a total of 84 variables, which represented the most critical information derived from the survey questions. By reducing the dataset's dimensionality by 76.4%, factor analysis simplified the data while retaining the essential information.

Each factor in Table 3 represents a distinct theme or underlying construct in the survey data. For example, in the education section, various questions related to school facilities, teacher availability, and access to learning materials were analysed. This analysis identified a key factor, 'edu1_resources', which encompasses the overall quality of educational resources. This factor is defined by variables such as the condition of school facilities and the availability of educational materials, offering a measure of the school environment and resources available to learners.

By focusing on the 84 most pertinent variables, this approach enhances the efficiency of subsequent analyses, and ensures that the factors contribute meaningfully to the interpretation of the data. These variables were used to train the machine learning model and subsequently to define each agent in the simulation model, enabling the intervention capabilities discussed in Section 4.3.

3.3. Target variable

The 2019 General Household Survey dataset contains 68,789 entries, including household heads, spouses, children, and other household members. To focus specifically on learners in order to be relevant to this study, the dataset was filtered to include only those attending public high schools. Learners aged 13 to 18, as well as those attending high school beyond these ages, were selected. This filtering process resulted in a total of 5,962 entries, representing the target learner population.

To establish the target variable for the machine-learning model, a search was conducted to identify learners from the 2019 General Household Survey who also participated in the 2020 survey. This longitudinal tracking identified 1,246 learners with unique identifiers in both datasets. These learners provided the necessary

data to define the target variable—whether they passed or failed their grade. For instance, if a learner was in Grade 9 in 2019 and advanced to Grade 10 in 2020, it was recorded that the learner had passed. It should be noted that, while the progression policy of the Department of Basic Education influenced learner progression, the data does not differentiate between passing as a result of academic success or because of the policy [34]. The longitudinal search in the survey indicated only whether they had passed, regardless of the reason.

Table 3: Factors derived from the factor analysis

Section	Factor	Abbreviation	Explanation
Household Specific	Traditional household structure	hhc1_nuclear	Variables related to marital status, gender of the household head, and the father's presence
	Maternal presence in the household	hhc2_maternal	Mother alive and present
	Families with grandparents and low education levels	hhc3_grand	Grandparents apart of household and education attained
Education	School environment and resources	edu1_resources	Condition of school facilities, availability of teachers, and access to learning materials
	Learners' academic proficiency and cognitive development	edu2_proficiency	Variables reflecting academic abilities and cognitive growth
	Cost of attending school	edu3_cost	School feeding programs, transportation, and tuition fees
Health & functioning	School absence	hlt1_injury	Variables related to injury resulting in school absences
	Sensory disabilities	hlt2_disability	Variables associated with eyesight difficulties and lack of medical aid coverage
	Chronic illnesses and health status	hlt3_health	Chronic illnesses, pregnancy and overall health status
Social & economic activities	Household income and expenses	soc1_income	Monthly income, expenses, and reliance on wages
	Social grants and relief assistance	soc2_grant	Variables indicating social grants or relief assistance
House conditions	Dwelling conditions and infrastructure	hsg1_dwelling	Physical state of the dwelling, sanitation, electricity, and access to drinking water
	Environmental cleanliness and waste management	hsg2_enviro	Surroundings, absence of pollution, and regular waste removal services
Welfare & hunger	Food security and hunger	hhw1_foodsecurity	Adequate food supply and security
	Access to amenities and assets	hhw2_assets	Private healthcare, ownership of vehicles and household assets, and employment of domestic workers

4. METHODOLOGY

This section details the integrated approach that combines factor analysis, machine learning, and agent-based modelling. The machine learning model predicts whether a learner will pass or fail, while the agent-based model simulates the progression of learners through high school. As illustrated in Figure 1, after conducting factor analysis on the dataset, the XGBoost machine-learning model was trained (Section 4.1). Subsequently, an application programming interface was developed (Section 4.2) to pass the machine-learning model's predictions to the agent-based model (Section 4.3).

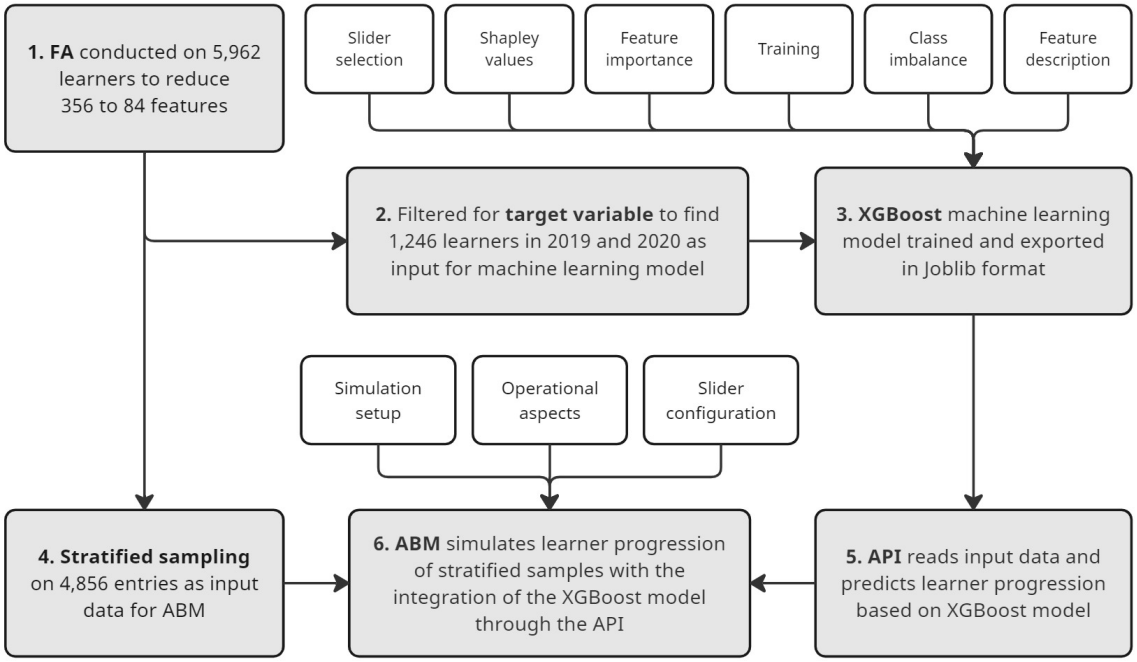

Figure 1: Methodology flow diagram

4.1. Machine-learning model

The machine-learning component of this study aimed to predict whether a learner would pass or fail a grade. The model was built using 1,246 entries from the General Household Survey of 2019 and 2020, with features including the learner's province, grade, over-age status, and 81 variables derived from factor analysis. The target variable was whether the learner had passed in 2019.

However, a difficulty emerged from the imbalance in the target variable. Most learners in the dataset had passed in 2019, with a significantly smaller proportion failing. This imbalance could have skewed the model's predictions. To mitigate this issue, the synthetic minority over-sampling technique (SMOTE) was applied. SMOTE generates synthetic samples for the minority class (those who failed), which helps to balance the dataset and enhances the model's ability to predict the minority class, ultimately improving its overall accuracy.

Various machine learning algorithms were considered for this study, including random forest, linear regression, neural networks, and support vector machines. Random forest was considered for its robustness and ability to handle large datasets with high dimensionality, and neural networks for their capability to model complex non-linear relationships. However, XGBoost was chosen for its ability to handle imbalanced datasets. XGBoost's gradient boosting framework optimises accuracy through iterative refinement, making it particularly suitable for the intricacies of General Household Survey data when subtle patterns need to be captured.

Feature importance in machine learning quantifies each feature's contribution to the model's prediction accuracy, as illustrated in Figure 3, which presents the top 20 contributing features and their importance scores. A high feature importance score does not necessarily correspond to higher passing probabilities. Instead, it reflects how much a feature informs the model about the learner's context, providing critical information that helps the model accurately to distinguish between learners who are likely to pass and those who are likely to fail.

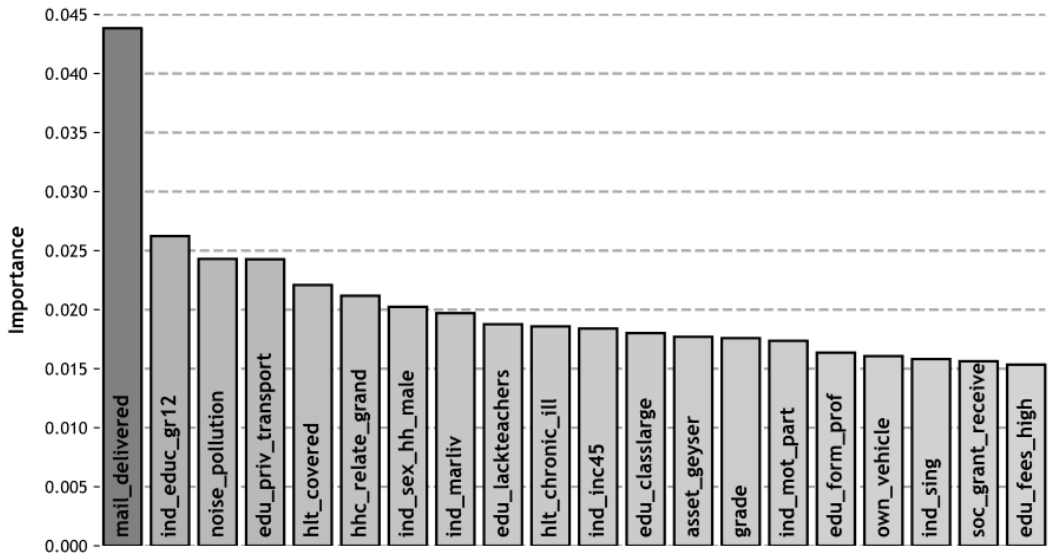

Figure 2: Feature importance in machine-learning model

The feature 'mail_delivered' is the most significant contributor to the model's predictions, suggesting that access to regular mail delivery, which may indicate a stable and affluent living environment, is crucial in understanding the learner's context. Next to it, 'ind_educ_gr12' pertains to the educational level of the learner's parents or guardians, and specifically whether they have matriculated, emphasising the importance of parental education levels in predictive accuracy.

Features such as 'edu_priv_transport', 'hlt_covered', 'ind_inc45', 'asset_geyser', 'own_vehicle', and 'soc_grant_receive' suggest that financial security factors are significant in the model's predictions. School-related features such as 'edu_lackteachers', 'edu_classlarge', 'grade', 'edu_form_prof', and 'edu_fees_high' also play a significant role, highlighting the influence of academic resources and formal education settings. In addition, features such as 'hhc_relate_grand', 'ind_sex_hh_male', 'ind_marliv', 'ind_mot_part', and 'ind_sing' all relate to the household structure of the learner.

4.2. Application programming interface

Integrating the machine-learning model's predictions into the agent-based model posed several difficulties. The complexity of the agent-based model environment, coupled with the computational demands of the machine-learning model, made direct integration difficult.

To overcome these issues, an application programming interface was developed to facilitate the interaction between the simulation model and the machine-learning model. This approach enabled effective communication between the two components without requiring direct integration, thereby ensuring smoother operation and increased flexibility. The application programming interface was built using the Flask web framework, which is well-suited for creating lightweight web services. The application programming interface accepts input data in JSON format, processes this data, and uses the pre-trained XGBoost model to make predictions. The predictions are then returned to the simulation model in real time.

4.3. Agent-based simulation model

The agent-based model simulates the progression of learners through high school, aiming to capture the dynamics and complexities of the educational system. Stratified sampling was performed on 4,856 learners to create a representative dataset of 250 entries, with 50 learners per grade, serving as input for the simulation. Each learner is modelled as an agent with 84 distinct parameters, including variables from factor analysis, province, grade, and over-age status.

As can be seen in Figure 4, learners enter the model at any of the five grades, depending on their grade in 2019. At each grade level, the agent-based model sends the learners' variables to the application programming interface, which uses the pre-trained machine learning model to predict whether each learner will progress or not. These predictions are then fed back into the agent-based model, which uses them to simulate the learners' progression through high school. If the model passes a learner, the agent moves to the next grade. Learners who pass Grade 12 are marked as 'matriculated'. The integration of the machine learning model via the application interface enables the agent-based model to make data-driven predictions on a learner-to-learner basis.

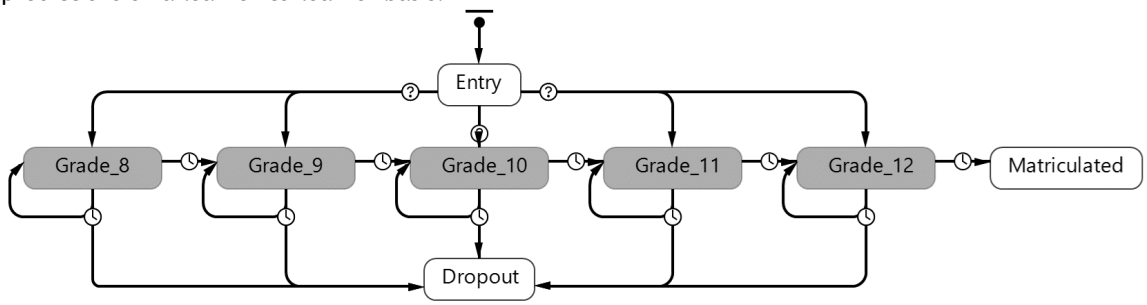

Figure 3: Agent-based simulation setup

5. VALIDATION

Validation is essential to ensure the accuracy and reliability of the model used in this study, providing clear insights into the factors that influence educational outcomes. This section assesses the performance of both the XGBoost machine learning model (Section 5.1) and the agent-based simulation model (Section 5.2), highlighting their strengths and identifying areas for improvement. Furthermore, the agent-based model's results are compared with historical data from the learner unit record information and tracking system (LURITS) to evaluate its accuracy.

5.1. Machine learning validation

The performance of the XGBoost machine-learning model was evaluated using precision, recall, and the F1-score. As shown in Table 4, the model achieved a precision of 0.84 for predicting learners who passed. Precision is the ratio of true positive predictions to all predicted positives. Recall, which measures the ability to capture all actual positives, was 0.81 for passing learners. The F1-score, which balances precision and recall, was 0.82 for 'pass' predictions, demonstrating the model's effectiveness in predicting learner progression.

Table 4: Performance metrics

Prediction	Precision	Recall	F1-score
Fail	0.39	0.44	0.41
Pass	0.84	0.81	0.82

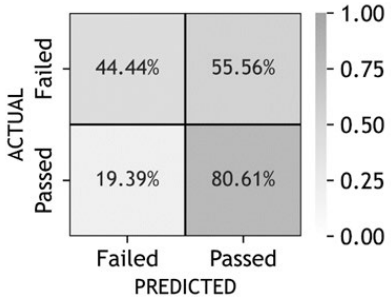

Figure 4: Confusion matrix

A confusion matrix, as can be seen in Figure 5, is a performance measurement tool used for classification models. Despite its name, it provides clarity by displaying the model's predictions against the actual outcomes. The matrix includes four components: true positives (correctly predicted positive cases), true negatives (correctly predicted negative cases), false positives (incorrectly predicted positive cases), and false negatives (incorrectly predicted negative cases). This matrix helps to identify where the model performs well and where it needs improvement, particularly in distinguishing between classes. Figure 5 clearly shows that the model is more accurate in predicting learners who have passed than those who have failed.

The final results showed an overall prediction accuracy of 72.80%. As depicted in Figure 5, the model correctly identified 44.44% of the 'Fail' class and 80.61% of the 'Pass' class. However, it misclassified 55.56% of the negative cases as positive and 19.39% of the positive cases as negative. These results demonstrate the model's strong recall and precision for the 'Pass' class, but also highlight the need for improvement in handling the minority 'Fail' class.

5.2. Agent-based simulation validation

To validate the agent-based model, a comparison was made between the predicted simulation results and historical data from the LURITS dataset – an essential tool developed by the South African Department of Education to manage detailed data on individual learners from Grade R to Grade 12. Each learner is assigned a unique tracking number throughout their schooling, facilitating consistent record-keeping even if they move between schools or provinces. LURITS relies on data from computerised school administration systems, such as the South African School Administration and Management System (SA-SAMS), which is provided to schools free of charge. This system ensures fairly accurate enrolment numbers, and tracks learner movements, supporting the Department of Education in identifying and addressing issues such as school dropouts.

The validation process began with data preparation, in which the LURITS dataset was cleaned and filtered to remove any duplicates and entries with missing identifiers. This ensured that the dataset used for validation was accurate and representative of the actual learner population. Learner progression percentages from the 2019 LURITS data were compared with the simulation results, as shown in Figure 7, where 'Pass' percentages are indicated by the hatched pattern and 'Fail' percentages in white.

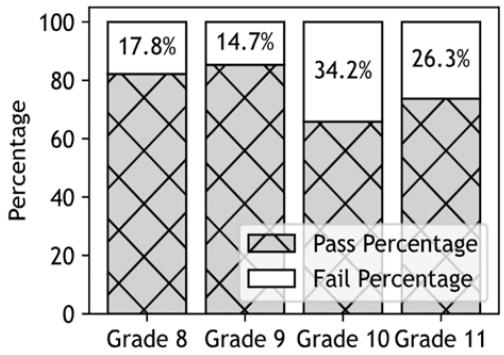

Figure 5: Predicted progression for 2019

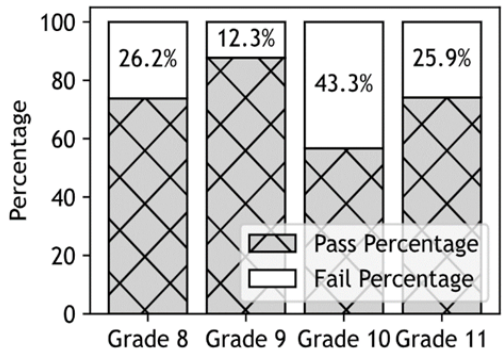

Figure 6: Actual LURITS progression

The comparison reveals that the simulation model closely replicates real-world learner progression, particularly in Grades 9 and 11, where simulated 'Pass' and 'Fail' percentages align closely with actual data. However, discrepancies are noted in Grades 8 and 10, indicating areas for further improvement.

To quantify the model's accuracy, the root mean square error (RMSE) is calculated. RMSE is a standard metric that measures the average magnitude of error between predicted and observed values, with a lower RMSE indicating better performance. In this study, the RMSE for the pass percentage is 2.94%, meaning that the simulation's predictions deviate by an average of 2.94 percentage points from the LURITS data. This low RMSE value suggests that the model performs well in predicting pass rates, with only a small margin of error, which could be further reduced with additional tuning, though it already demonstrates a strong alignment with actual learner outcomes.

It is important to note that results for Grade 12 are not included in this comparison. The LURITS system only tracks learners from Grade 0 to Grade 12, and once Grade 12 learners matriculate, they are no longer included in the dataset for the following year. Consequently, any Grade 12 learners who graduated in 2019 would not appear in the 2020 dataset, making it difficult to compare the pass rate in the simulation model with the actual data. Therefore, this validation of the agent-based model using LURITS data is limited to Grades 8 to 11.

6. SCENARIO TESTING AND RESULTS

This section explores three scenarios that are tested using the agent-based simulation model in order to understand their impact on learner progression in the South African high school education system. The scenarios illustrate the usability of the simulation model, and are not meant to be comprehensive.

6.1. Scenario 1: Increase in higher-cost education by 10%

The first scenario examines the effect of a 30% increase in the presence of the 'edu_high_fees' variable, which reflects the financial cost of attending a school. As indicated by the results in Table 5, schools with higher fees tend to perform better. This outcome is consistent with the reality that more affluent South African schools, which typically charge higher fees, often produce better passing rates. This finding highlights the potential benefits of investing in educational resources, which could improve learner outcomes in schools that have traditionally been underfunded.

6.2. Scenario 2: Decrease in basic educational proficiency by 30%

The second scenario explores the impact of a 30% reduction in the 'edu_form_prof' variable, which measures learners' basic proficiency in essential skills such as reading, writing, and comprehension. Table 5 demonstrates that a decline in basic proficiency hinders passing rates. Learners who struggle with foundational skills are less likely to progress and are more likely to fail. This scenario underscores the importance of ensuring that all learners attain a minimum level of basic proficiency to support their academic success.

6.3. Scenario 3: Increase in teachers and decrease in classroom size by 15%

The third scenario involves simultaneously increasing the number of teachers and decreasing classroom size by 15% to assess the combined effect on passing rates. As shown in Table 5, this dual approach yields positive results, with improvements observed across various grades. The rationale behind this outcome is straightforward: smaller classroom sizes enable more individualised attention, and an increased number of teachers enhances the overall quality of education. Together these factors contribute to better learner performance and higher pass rates.

Table 5: Passing rate percentage improvement on scenario testing

Scenario	Grade 8	Grade 9	Grade 10	Grade 11
1	7.49	17.45	5.50	15.66
2	-8.82	-0.95	-3.37	-19.33
3	26.47	24.03	12.67	17.95

6.4. Scenario results

As detailed in Table 5, the scenario of increasing higher-cost education variables by 10% shows moderate improvements across Grades 8, 9, 10, and 11, indicating that financial investment in schools could lead to better educational outcomes. In contrast, the scenario of decreasing basic proficiency by 30% results in higher failure rates, particularly in Grades 8 and 11, highlighting the detrimental impact of low proficiency on learner progression. The third scenario, involving a 15% increase in teachers and a decrease in classroom size, demonstrates substantial improvements in all grades, especially Grades 8 and 9, underscoring the positive effect of reducing class sizes and increasing teacher availability on educational progression.

7. CONCLUSIONS AND FUTURE WORK

Using machine learning and agent-based modelling, this study illuminates critical aspects of learner progression in the South African high school education system. By incorporating factor analysis to distil key features and by using the XGBoost model for predictive analytics, the simulation captures the dynamics of learner progression. The model includes variables pertaining to educational resources, household income, health status, housing conditions, social grants, and food security, alongside additional variables such as grade, province, and whether the learner is over-aged for their grade.

The agent-based model's ability to simulate individual behaviours provides a granular understanding of the factors that influence educational outcomes. Validation against the LURITS dataset confirms the model's accuracy, particularly in Grades 9 and 11. The RMSE for the pass percentage is 2.94%, indicating that the simulation's predictions deviate by an average of 2.94% from the LURITS data. This low RMSE demonstrates the model's strong alignment with actual learner outcomes, although room remains for refinement through additional tuning.

Future work will focus on incorporating system dynamics into the simulation framework. This addition is expected to complement the current approach by providing a macro-level perspective, capturing feedback loops and long-term trends that influence the education system. The combination of system dynamics and agent-based modelling would create a more comprehensive and interactive model that would be capable of simulating the impact of policy changes and resource allocation over time.

The integration of machine learning with agent-based simulation modelling provides a valuable tool for planning educational improvements in the South African educational system. This simulation model serves as a strategic platform for evaluating and refining interventions.

REFERENCES

[1] K. V. Astakhova, "The role of education in economic and social development of the country," *International Review of Management and Marketing,* vol. 6, no. S1, pp. 53-58, 2016.

[2] C. Kayembe and D. Nel, "Challenges and opportunities for education in the Fourth Industrial Revolution," *African Journal of Public Affairs,* vol. 11, no. 3, pp. 79-94, 2019.

[3] OECD, "Education at a glance: OECD indicators," OECD Publishing, 2022. Available: https://www.oecd.org/education/education-at-a-glance/EAG2019_CN_ZAF.-pdf

[4] Department of Basic Education, "Annual Report 2022," Education Management Information System (EMIS), 2022. Available: https://static.pmg.org.za/2022_23_DBE_Annual_Report_FINAL.pdf

[5] N. Spaull, "Poverty and privilege: Primary school inequality in South Africa," *International Journal of Educational Development,* vol. 33, no. 5, pp. 436-447, 2013.

[6] The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ), "The SACMEQ IV Project in South Africa: A study of the conditions of schooling and the quality of education," Department of Education, August 2017.

[7] V. Reddy *et al.*, "The Gauteng province TIMSS 2019 grade 9 results: Building achievement and bridging achievement gaps," HSRC Press, 2022. Available: https://repository.hsrc.ac.za/handle/20.500.11910/19474

[8] N. Spaull, "Background report for the 2030 reading panel," February 2023. Available: https://groundup.org.za/media/uploads/documents/embargoed_2023_reading_panel_background_report_7_feb_2023.pdf

[9] M. Khuluvhe and W. Gwantshu, "The highest level of education attainment in South Africa," Department of Higher Education and Training, TD/TNC 144.63, 2021. Available: https://www.dhet.gov.za/Planning%20Monitoring%20and%20Evaluation%20Coordination/Fact%20Sheet_Highest%20Level%20of%20Educational%20Attainment%20in%20South%20Africa%20-%20June%202022.pdf

[10] S. Slamang, "A hybrid simulation analysis of graduation success at Stellenbosch University," Stellenbosch, 2016. Available: https://scholar.sun.ac.za/items/82e539fb-963a-4bb5-9107-f54a38dd072d

[11] C. Perrie, "Analysing the Western Cape junior primary", Stellenbosch University, 2019.

[12] J. Grobler, "A systems perspective on teacher training in South Africa," PhD dissertation, Stellenbosch University, 2020.

[13] L. Venter, "A systems perspective of basic education in South Africa," PhD thesis, Stellenbosch University. Available: https://scholar.sun.ac.za/items/91797b6f-4383-4ae8-9063-2d9590349187

[14] C. Layton, “An overview of key datasets on education in South Africa and their uses,”, 2021.
[15] J. P. E. Motala, *The state, education and equity in post-apartheid South Africa: The impact of state policies*, Routledge Revivals, 2002. Available: https://www.routledge.com/The-State-Education-and-Equity-in-Post-Apartheid-South-Africa-The-Impact-of-State-Policies/Motala-Pampallis/p/book/9781138723641
[16] A. Law, *Simulation modeling and analysis* 5th ed., McGraw Hill Education, 2014.
[17] R. G. Sargent, “Verification and validation of simulation models,” in *Proceedings of Winter Simulation Conference,* 2012.
[18] A. Ford, *Modeling the environment: An introduction to system dynamics models of environmental systems* 2nd ed., Island Press, 2010.
[19] D. H. Meadows, *Thinking in systems: A primer*. Chelsea Green Publishing, 2008.
[20] G. P. Richardson, “System dynamics,” in *Encyclopedia of operations research and management science*, S. I. Gass and M. C. Fu (eds), Boston, MA: Springer, 2013, pp. 1519–1522.
[21] D. P. Kroese, T. Taimre, and Z. I. Botev, *Handbook of Monte Carlo methods*, John Wiley & Sons, 2011.
[22] C. P. Robert and G. Casella, *Monte Carlo statistical methods* 2nd ed., Springer, 2010.
[23] D. P. Dubhashi and A. Panconesi, *Concentration of measure for the analysis of randomized algorithms*, Cambridge University Press, 2009.
[24] S. F. Railsback and V. Grimm, *Agent-based and individual-based modeling: A practical introduction* 2nd ed., Princeton University Press, 2019.
[25] U. Wilensky and W. Rand, *An introduction to agent-based modeling: Modeling natural, social, and engineered complex systems with NetLogo*, The MIT Press, 2015.
[26] J. M. Epstein, *Agent_zero: Toward neurocognitive foundations for generative social science*, Princeton University Press, 2014.
[27] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems,* O’Reilly Media, 2019.
[28] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.
[29] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning: With applications in R,* Springer, 2013.
[30] T. Talan, “Artificial intelligence in education: A bibliometric study,” *International Journal of Research in Education and Science (IJRES),* vol. 7, no. 3, pp. 822-837, 2021.