Feature Analysis of Learner Adaptation across Socio-Economic Divide in South African Public High Schools

J. Visser[1*] & L. Venter[2]

ARTICLE INFO

Article details
Submitted by authors 29 Jan 2025
Accepted for publication 26 Jun 2025
Available online 29 Aug 2025

Contact details
* Corresponding author
jemmaannvisser@gmail.com

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

2 Department of Logistics, Stellenbosch University, Stellenbosch, South Africa

ORCID® identifiers
J. Visser
https://orcid.org/0009-0003-5207-1978

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

DOI
http://dx.doi.org//10.7166/36-2-3187

ABSTRACT

Persistent socio-economic disparities in the South African education system hinder learner progression. This study examines the factors shaping learner adaptation across socio-economic divides in public high schools, focusing on the contrasting experiences of learners from lower-income backgrounds in different school environments. Using data from the 2022 General Household Survey, the study applies factor analysis, multiple regression analysis, structural equation modelling, and machine-learning techniques to identify key determinants of learner progression, such as family structure, supported retention, welfare income, and household conditions. The findings show the significant role of school meals, age-appropriate grade placement, and consistent attendance in academic success. These insights point to the need for targeted, context-sensitive interventions to address the socio-economic barriers to educational achievement in post-apartheid South Africa.

OPSOMMING

Aanhoudende sosio-ekonomiese ongelykhede in die Suid-Afrikaanse onderwysstelsel verhinder steeds akademiese vordering. Hierdie studie ondersoek die faktore wat leerderaanpassing oor sosio-ekonomiese skeidslyne heen bepaal, met besondere aandag aan die uiteenlopende ervarings van leerders uit lae-inkomste agtergronde binne verskillende skoolomgewings. Die analise steun op data uit die 2022-weergawe van die Algemene Huishoudelike Opname en gebruik metodologiese benaderings soos faktoranalise, meerveranderlikeregressie-analise, strukturele vergelykingsmodellering en masjienleer om die kernbepalers van onderrigvordering te identifiseer, waaronder gesinstruktuur, ondersteunende behoud, welsyninkomste en huishoudelike omstandighede. Die bevindings toon die beduidende rol van skoolmaaltye, ouderdomsgepaste graadplasing en bestendige bywoning in akademiese welslae. Hierdie insigte wys op die noodsaak vir geteikende, konteksgevoelige ingrypings om die sosio-ekonomiese struikelblokke tot onderwysuitkomste in post-apartheid Suid-Afrika aan te spreek.

1. INTRODUCTION

Despite achieving democracy in 1994, South Africa continues to grapple with stark inequality and pervasive corruption. Even with its industrial, agricultural, and mineral wealth, the country faces a severe socio-economic crisis, including a 31.9% unemployment rate in 2023 [1], with 34.2% of youth aged 15 to 24 years disengaged from education or work [2]. Education is managed by the Department of Basic Education and the Department of Higher Education and Training. Public education spans four phases, culminating in the National Senior Certificate, and schools are grouped by socio-economic status into quintiles. Curricular frameworks aim to address inequalities and to foster social mobility.

The South African education system faces declining academic performance, persistent disparities, and significant problems with literacy and numeracy. The COVID-19 pandemic exacerbated disparities, though grade progression improved because of relaxed policies [3]. Between 2019 and 2021, systemic test scores fell in Grades 3, 6, and 9, with learners estimated to be from 40% to 70% of a year behind in language proficiency [4]. In 2019 fewer than half of the Grade 9 learners reached basic mathematics or science benchmarks [5]. With just 6% of learners proficient in mathematics or science, urgent reforms are essential to secure equitable education and future economic stability.

South African education exhibits a bimodal distribution, with learners segregated into two distinct performance groups [6]. One group, from higher socio-economic status (SES) or well-resourced schools, achieves significantly higher academic outcomes, while the other from lower SES contexts struggles. Progress in International Reading Literacy Study 2021 revealed that learners tested in English or Afrikaans performed significantly better than those tested in African languages, with the majority of African language-speaking learners failing to reach the low international benchmark in reading [7]. This divide reveals two separate education systems, necessitating targeted interventions to bridge disparities and to ensure equitable educational opportunities for all learners.

2. LITERATURE REVIEW

The South African education system faces considerable problems, leading to efforts to explore and analyse its complexities using advanced simulation modelling techniques, such as system dynamics and agent-based modelling. The primary aim of these analyses is to address the deterioration of basic education and the widening disparities in academic opportunities across the country.

Slamang [8] investigated the factors affecting academic performance in Western Cape public high schools using a system dynamics approach in order to model key determinants. The study identified three intervention areas as home-based, classroom-based, and poverty-related strategies. Classroom interventions, focusing on teacher effectiveness, school resources, and learning environments, were most impactful, significantly improving secondary education outcomes. Key influences included school resources, class size, teacher effectiveness, economic status, family health, learner motivation, and retention rates, particularly in Grades 9 to 12, informed by General Household Survey (GHS) data for 2015 to 2019.

Becker [9] analysed the factors influencing learner progression from Grade 8 to Grade 12 in South Africa, using the 2019 GHS data. Nine key determinants were identified: including supported retention, school resources, family structure, household wealth, food insecurity, access to utilities, environmental problems, household extravagances, and educational outcomes. Stable family structures and higher household wealth were strongly linked to improved academic outcomes and retention. The study emphasised the critical role of socio-economic factors in academic success, advocating targeted interventions to address educational disparities.

Van der Heever, Becker, Venter and Bekker [10] combined machine learning and agent-based modelling to simulate learner progression in public high schools, highlighting the critical factors influencing educational outcomes. An XGBoost model, trained on identified features, was embedded in an agent-based framework to simulate progression from Grade 8 through to Grade 12. Residing in a dwelling with a formal postal address was identified as the most significant predictor, suggesting its association with stable and affluent living conditions. Parental education, particularly whether parents or guardians had matriculated, was found to play a pivotal role in predictive accuracy, revealing the importance of household context in understanding learner progression.

While these studies reveal the drivers of South Africa's failing education system, neither Becker [9] nor Van der Heever *et al.* [10] accounted for the influence of socio-economic divides among learners. Spaull [6] emphasised the absence of an "average learner", prompting investigation about how socio-economic disparities might influence the factors and relationships that characterise South African learners.

Learners and schools are grouped into socio-economic quintiles, creating predominantly homogenous systems based on residential proximity and zoning. Exceptions occur when lower-quintile learners access higher-quintile schools through bursaries or parental investments, introducing diversity into systems designed to serve an assumed average learner. In this paper we use statistical modelling to explore how heterogeneous learner groups progress in a unified education system. Recognising inherent systemic bimodality, we examine the problems faced by lower SES learners when grouped with higher SES peers. The findings highlight the inadequacy of simplistic solutions, and advocate for inclusive, equity-focused strategies to address South Africa's complex educational disparities. This study aims to discover which factors differentiate public high school learners from lower socio-economic communities in two distinct school systems in South Africa, how are these factors interrelated, and in what way these factors contribute to predicting learner progression.

3. METHODOLOGY

This study follows a four-step methodology, as shown in Figure 1. First, the 2022 GHS data was divided into two separate datasets. Next, factor analysis (FA) was conducted to reveal the latent factors in the data. Following this, multiple regression analysis (MRA) and structural equation modelling (SEM) were used to map the relationships between these factors. In the final step, machine-learning (ML) techniques were applied to evaluate the relative importance of these factors in predicting learner progression.

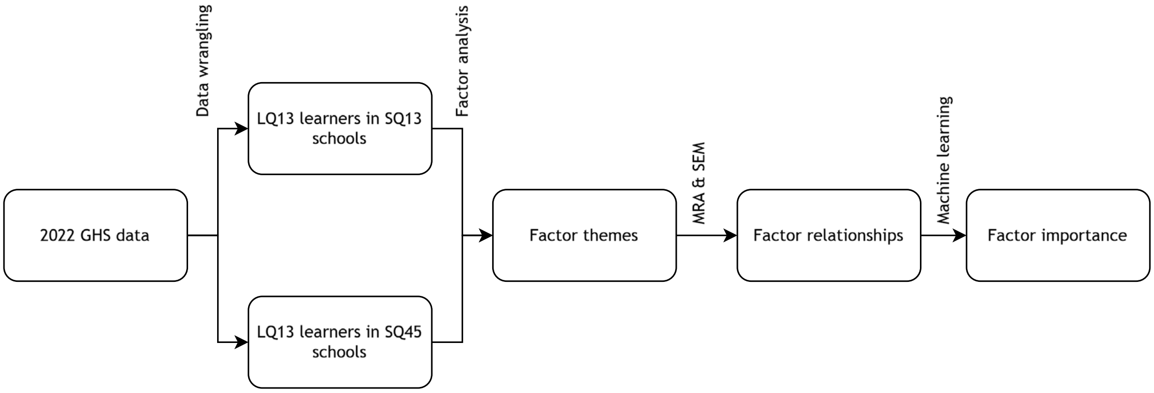


Figure 1: The four-step methodology for this study

The 2022 GHS collected data through face-to-face interviews on a range of household and individual characteristics in South Africa. Given that 2022 is considered the first normalised year following the systemic disruptions of the COVID-19 pandemic, it was deemed the most appropriate data for this study, following on from the analysis of the 2019 GHS dataset by Van den Heever *et al.* [10]. The survey recorded household-level attributes such as home ownership, access to water, sanitation, transportation, and agricultural production, alongside individual characteristics such as demographics, education, income, health, and access to social services.

For this analysis, the data was categorised according to SES. South Africans are grouped into quintiles based on their SES, with the lower three quintiles representing lower socio-economic groups. Similarly, schools are categorised into quintiles, with Quintiles 1 to 3 being no-fee schools, and Quintiles 4 and 5 consisting of wealthier institutions. This study investigated how learners from lower socio-economic households (LQ13) progress in both lower-status (SQ13) and higher-status (SQ45) schools, so that the dataset was divided accordingly into two distinct subsets. The dataset exhibited a hierarchical structure, as multiple learners resided in a single household. To address this, one learner from each household was selected to represent their household. This was appropriate, as Becker [9] demonstrated significantly low intraclass correlation coefficients between individual and household records for public high school learners in the 2019 GHS dataset.

The original GHS dataset contained 302 features. FA was used to reduce this complexity for further analysis. The use of FA reduced the complexity of the subsets by identifying latent patterns among variables [11]. Initially, the raw data was encoded into binary indicator variables through one-hot encoding. This eased the transformation of categorical variables, particularly to handle non-numeric data in the survey responses. This ensured consistency where the data comprised discrete, non-continuous variables.

To assess the suitability of the data for FA, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (MSA) was applied. The KMO MSA for each indicator was compared. Indicators with values below 0.5 were excluded from the FA; those with values between 0.5 and 0.8 were considered with caution; and those with values above 0.8 were deemed ideal for FA.

To determine the appropriate number of factors to extract, either the eigenvalue rule or the scree plot could be used. According to the eigenvalue rule, factors with eigenvalues significantly greater than 1 are retained. Alternatively, the scree plot is used to identify the “elbow point” where the curve levels off, indicating the appropriate number of factors to retain.

Once the data had been suitably encoded and the appropriate number of factors determined, FA was performed. The process began with exploratory FA (EFA), which aims to uncover underlying data structures without predefined hypotheses. Next, confirmatory FA (CFA) was conducted to test specific hypotheses. FA is an iterative process; thus indicators were added and removed until a well-fitting model was achieved. This iterative approach ensured that the final model accurately represented the data and aligned with the research objectives.

Following the identification of factor themes, it was essential to examine how they were interrelated. This was accomplished using MRA and SEM. MRA investigates the relationship between a quantitative dependent variable and multiple predictors, accommodating both linear and non-linear associations. Significance was determined through hypothesis testing, with *p*-values below 0.05 considered indicative of statistical significance.

SEM explores more complex relationships among variables, extending beyond traditional regression by incorporating latent variables and evaluating both direct and indirect effects. SEM assesses model fit using indices such as the comparative fit index (CFI), standardised root mean square residual (SRMR), and root mean square error of approximation (RMSEA), while the strength of relationships is evaluated through standardised path coefficients and *t*-values [11].

Finally, machine learning was applied to predict learner progression, focusing on the key factors influencing grade advancement. The model included features such as province, grade, age-grade delta, gender, home language, and indicator variables derived from FA. The target variable — learner progression in 2021 — was determined retrospectively, based on a 2022 GHS survey question that recorded whether the learner was repeating their current grade.

The target variable exhibited a significant class imbalance, with about 85% of learners being classified as having progressed, compared with a smaller proportion representing repeaters or dropouts. To manage this imbalance, the synthetic minority over-sampling technique (SMOTE) was used to generate synthetic samples for the minority class, enhancing the model’s ability to predict failures and improving its accuracy.

Van der Heever *et al.* [10] demonstrated the efficacy of gradient boosting — specifically XGBoost — in predicting learner progression in the 2019 GHS dataset. Therefore, gradient boosting was also used in this study as the appropriate machine-learning technique. A grid search determined the optimal hyperparameters, and the model’s feature importance was assessed to evaluate each variable’s contribution to prediction accuracy. An iterative process was used to identify the minimum number of features needed to maintain accuracy, guiding identification of the relevance of the features that were most critical in determining learner progression.

4. RESULTS

The first analysis focused on LQ13 learners attending SQ13 schools, using a sample of 2,839 learners, comprising 52% male and 48% female. Home languages varied, with 28% speaking isiZulu, 17% Sepedi, 15% isiXhosa, and 11% Sesotho.

FA revealed factors describing learners’ education, health, and family structure with an overall MSA of 0.62. Three person-level factors (Person_F1 to Person_F3) were identified, as presented in Table 1, where each indicator’s KMO MSA score exceeded 0.5. Factor loadings, scaled for clarity, showed clear clustering. CFA confirmed excellent fit (CFI = 0.92, SRMR = 0.03, RMSEA = 0.07).

Table 1: KMO MSA and the scaled rotated factor pattern for person-level characteristics for LQ13 learners in SQ13 schools

Indicator	*KMO* *MSA*	*Person_F1* *Family structure*	*Person_F2* *Learner health*	*Person_F3* *Supported retention*
Maternal participation	0.60	**85***	-1	1
Co-residence with parents	0.67	**85***	1	1
Maternal vitality status	0.60	**63***	-4	1
Paternal participation	0.65	**59***	2	-2
Good hygiene	0.62	0	**80***	6
Good memory	0.65	-1	**76***	2
Good communication	0.68	0	**73***	0
Retention	0.56	2	9	**84***
School feeding	0.57	4	5	**80***
Walk to school	0.70	-4	-3	**60***

Person_F1 reflects family structure and parental involvement; *Person_F2* encompasses institutional support, including access to transport, school meals, and retention rates; and *Person_F3* represents physical and cognitive capabilities, covering communication skills, self-care activities such as dressing and hygiene, and cognitive functions such as memory and concentration.

FA was again applied to household indicators describing social security, economic activities, welfare, and hunger, with an overall KMO MSA of 0.72 confirming suitability. Table 2 presents the KMO MSA values and factor loadings for the five factors (*Hhold_F1* to *Hhold_F5*) that were identified. The CFA results showed excellent model fit (CFI = 0.97, SRMR = 0.05, RMSEA = 0.05).

Table 2: KMO MSA and the scaled rotated factor pattern for household characteristics for LQ13 learners in SQ13 schools

Indicator	*KMO* *MSA*	*Hhold_F1* *Food security*	*Hhold_F2* *Welfare income*	*Hhold_F3* *Sanitation*	*Hhold_F4* *Energy*	*Hhold_F5* *Environment*
Food available	0.84	**93***	-3	2	2	9
Food sufficient	0.84	**92***	-3	2	2	9
Food has variety	0.85	**88***	-3	1	2	8
Household not hungry	0.85	**87***	-2	1	4	9
General toilet available	0.85	-2	-6	**93***	-3	-1
Managed waste collection	0.67	0	-4	**90***	-4	3
Inhouse water	0.83	6	-8	**82***	9	4
Main electricity meter	0.60	6	-1	3	**91***	2
Energy access	0.66	1	-1	8	**83***	-2
Paid electricity	0.72	1	2	-9	**78***	7
Grant received	0.54	1	**95***	0	-2	-2
Childcare grant received	0.54	1	**94***	3	-4	-3
Unearned income earned	0.90	-9	**55***	-18	4	3
No air pollution	0.90	0	0	-2	9	**81***
No water pollution	0.70	14	2	8	0	**80***
No littering	0.78	13	-2	1	1	**74***

Hhold_F1 represents food security, encompassing availability, access, and utilisation; *Hhold_F2* pertains to welfare income derived from state support; *Hhold_F3* relates to sanitation, including water and waste management; *Hhold_F4* indicates energy access and type; and *Hhold_F5* addresses environmental conditions such as pollution and litter.

The MRA examined the relationships among eight composite factors for LQ13 learners in SQ13 schools, with the correlation significance at $p < 0.05$ (*) and strong significance at $p < 0.001$ (**). These results are presented in Table 3. The application of SEM to these factors revealed intricate interrelationships, highlighting their mutual influence. The SEM results, presented in Table 4, yielded excellent fit indices (CFI = 0.99, SRMR = 0.0015, RMSEA = 0.006), confirming a well-fitting model and providing valuable insights into system dynamics.

Table 3: Correlation matrix for the eight composite factors for LQ13 learners in SQ13 schools. Rows represent predictors, columns represent outcomes. ** indicates $p < 0.001$, * indicates $p < 0.05$.

	Family structure	***Learner health***	***Supported retention***	***Food security***	***Welfare income***	***Sanitation***	***Energy***	***Environment***
Family structure	1.00	-0.00	0.01	0.00	0.02	0.09**	-0.08**	0.02
Learner health	-0.03	1.00	0.02	0.06	0.03	-0.14	-0.03	0.10
Supported retention	0.04	0.28**	1.00	-0.02	0.28**	-0.08	0.02	-0.02
Food security	0.00	0.01	-0.02	1.00	-0.03	0.02	0.08*	0.22**
Welfare income	0.02	0.00	0.05**	-0.03	1.00	-0.07**	-0.04	-0.01
Sanitation	0.06**	-0.02	-0.02	0.02	-0.07**	1.00	-0.02	0.05*
Energy	-0.01**	-0.02	0.02	0.03*	-0.04	-0.01	1.00	0.04**
Environment	0.02	0.02	-0.01	0.18**	-0.01	0.05*	0.12**	1.00

Table 4: Standardised effects of predictors on outcomes, including the estimated path coefficients (*β*), standard errors (SE), and *t*-values for each relationship of the eight composite factors for LQ13 learners in SQ13 schools

Predictor	*Outcome*	*β*	*SE*	*t-value*
Environment	Food security	0.21	0.02	11.73
Welfare income	Supported retention	0.13	0.02	6.83
Environment	Energy	0.09	0.02	4.69
Supported retention	Learner health	0.08	0.02	4.26
Family structure	Sanitation	0.08	0.02	4.13
Sanitation	Welfare income	-0.08	0.02	-4.05
Sanitation	Environment	0.05	0.02	5.56
Family structure	Energy	-0.05	0.02	-2.94

Figure 2 shows the relational mapping of predictors and outcomes, represented as a network diagram of composite factors for LQ13 learners in SQ13 schools. The path coefficients on the arcs illustrate the strength of these relationships. While causation cannot be inferred, the presence of one factor is associated with the likelihood of a corresponding outcome. The diagram presented in Figure 2 is static and emphasises associations rather than causal links.

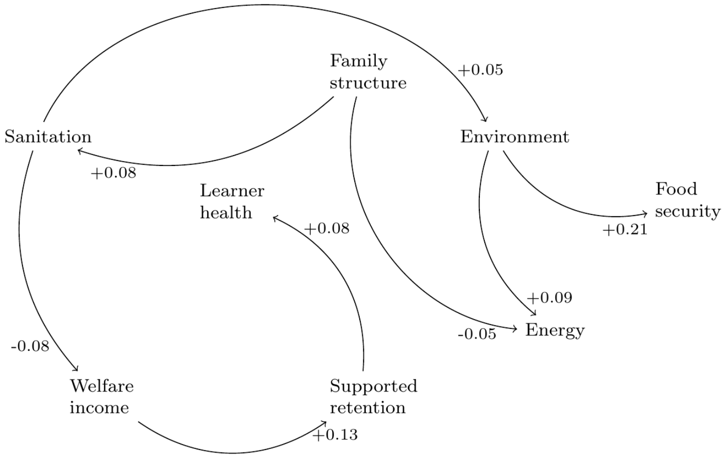


Figure 2: Relational mapping of the predictors and outcomes included in the estimated path coefficients as a network diagram of the eight composite factors for LQ13 learners in SQ13 schools

The feature importance analysis for LQ13 learners in SQ13 schools used a dataset consisting of 68 features. The parameters selected from the grid search are presented in Table 5. The XGBoost model's performance was assessed using precision, recall, and F1-score. The model achieved a precision of 0.88, recall of 0.95, and an F1-score of 0.92 for predicting learners who progressed, demonstrating its efficacy in predicting learner progression, as presented in Table 6.

Table 5: Key gradient boosting classifier parameters for LQ13 learners in SQ13 schools

Parameter	*Value*
learning_rate	0.3
*n*_estimators	200
Max_depth	4
Min_samples_split	5
Min_samples_leaf	2
Loss	Log_loss
Subsample	1.0
Validation_freaction	0.1

Table 6: Classification metrics by class for LQ13 learners in SQ13 schools

Class	*Precision*	*Recall*	*F1-score*
Repeat	0.49	0.26	0.34
Progress	0.88	0.95	0.92

The confusion matrix (in Figure 3) evaluated the model's performance, categorising the results into true positives, true negatives, false positives, and false negatives. The model achieved an overall accuracy of 85.00%, correctly identifying 95.25% of the progression cases but only 26.19% of the repeat cases. These results highlight strong recall and precision for progression, but indicate a need for improvement for repetition predictions.

True label	Fail	26.19%	73.81%
	Pass	4.75%	95.25%
		Fail	Pass

Predicted label

Figure 3: Confusion matrix of the machine learning model for LQ13 learners in SQ13 schools

Figure 4 illustrates the feature importance derived from the machine-learning model, highlighting the key predictors of learner success. Learner retention is the most significant feature, emphasising the importance of regular school attendance for grade progression. Age, reflecting the grade-age gap, also plays a key role, while the child support grant feature shows the importance of financial support in academic success for learners in lower-quintile schools.

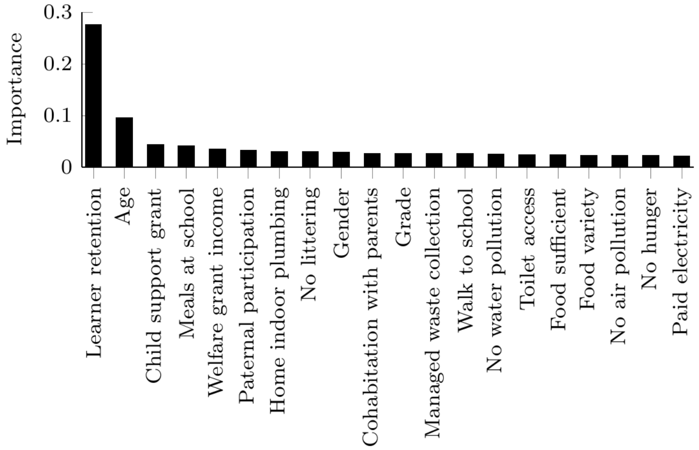


Figure 4: Feature importance for LQ13 learners in SQ13 schools

Figure 5 presents the results from an iterative approach, showing that model accuracy remains around 80% until the 16th feature is removed. This suggests that, with 80% confidence, a learner's progression in an LQ13 context in an SQ13 school can be reasonably predicted using only 16 features.

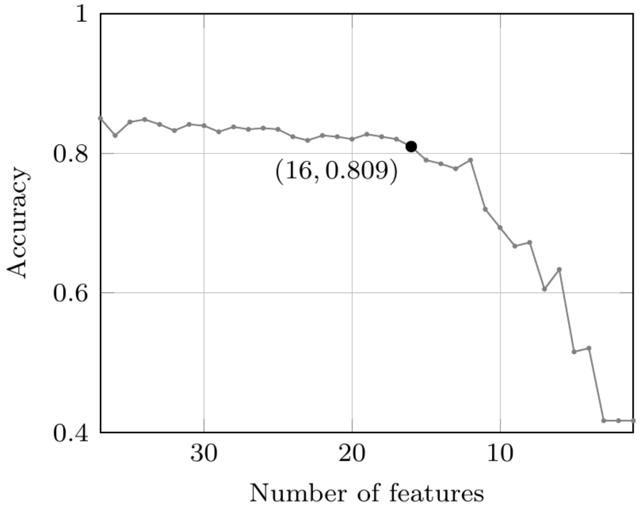


Figure 5: Change in learner progression prediction accuracy against number of features for LQ13 learners in SQ13 schools

The second analysis focused on LQ13 learners attending SQ45 schools, using a sample of 1,716 learners, comprising 53% male and 47% female. Home languages varied, with 25% speaking isiZulu, 18% isiXhosa, 12% Afrikaans, and 11% Setswana.

FA revealed the factors describing learners' education, health, and family structure, with an overall MSA of 0.59. Three learner-level factors were identified, as detailed in Table 7, where each indicator's MSA value exceeded 0.5. Factor loadings, scaled for clarity, demonstrated clear clustering. CFA confirmed excellent model fit (CFI = 0.99, SRMR = 0.03, RMSEA = 0.03).

Table 7: KMO MSA and the scaled rotated factor pattern for person-level characteristics for LQ13 learners in SQ45 schools

Indicator	*KMO MSA*	*Person_F1 Family structure*	*Person_F2 Learner health*	*Person_F3 Supported retention*
Co-residence with parents	0.57	**89***	1	-2
Maternal participation	0.60	**81***	1	4
Paternal participation	0.65	**72***	2	-11
Good hygiene	0.62	**4**	**88***	3
Good walking	0.54	3	**87***	0
Good communication	0.79	-1	**52***	-4
School feeding	0.57	-10	**2**	**80***
Walk to school	0.59	-13	0	**78***
Retention	0.68	11	-4	**57***

The three factors that emerged (i.e., family structure, supported retention, and learner health) were consistent with SQ13 schools. The analysis extended to household characteristics, including social security, economic activities, welfare, and hunger. EFA, justified by an MSA of 0.71, identified five factors where indicators had MSA scores above 0.5, as detailed in Table 8. CFA confirmed an excellent fit (CFI = 0.97, SRMR = 0.05, RMSEA = 0.05), validating the model.

Table 8: KMO MSA and the scaled rotated factor pattern for household characteristics for LQ13 learners in SQ45 schools

Indicator	*KMO MSA*	*Hhold_F1 Food security*	*Hhold_F2 Welfare income*	*Hhold_F3 Sanitation*	*Hhold_F4 Energy*	*Hhold_F5 Environment*
Food available	0.76	**93***	1	7	-7	6
Food sufficient	0.76	**93***	1	10	-10	6
Household not hungry	0.82	**92***	1	8	-4	3
General toilet available	0.71	7	3	**91***	-9	0
Managed waste collection	0.73	6	2	**89***	-11	1
Inhouse water	0.84	9	13	**82***	-13	7
Main electricity meter	0.59	2	1	4	**87***	7
Energy access	0.67	1	-3	6	**77***	0
Paid electricity	0.68	-1	0	6	**77***	4
Grant receiver	0.57	-4	**95***	-4	0	-1
Childcare grant receiver	0.93	-2	**95***	-4	0	-2
Unearned income receiver	0.76	-15	**60***	-29	-3	-5
No air pollution	0.72	4	4	2	5	**81***
No noise pollution	0.66	-4	2	0	0	**77***
No littering	0.58	14	4	6	-8	**74***

FA on learners in low and higher socio-economic schools revealed similar factors across strata, indicating that the identified school characteristics reflect broader environmental influences rather than being driven solely by socio-economic background, capturing both learner experiences and overall school functioning.

MRA, when applied to the data for SQ45 schools, examined the relationships among eight composite factors, with the correlation coefficients detailed in Table 9. Thereafter, SEM revealed the interrelationships among the composite factors; the results are detailed in Table 10, and depicted as a relational mapping diagram in Figure 6.

Table 9: Correlation matrix for the eight composite factors for LQ13 learners in SQ45 schools. Rows represent predictors, columns represent outcomes. ** indicates $p < 0.001$, * indicates $p < 0.05$

	Family structure	*Learner health*	*Supported retention*	*Food security*	*Welfare income*	*Sanitation*	*Energy*	*Environment*
Family structure	1.00	0.34	-0.05	-0.01	-0.10**	0.15**	0.00	-0.02
Learner health	0.34	1.00	0.08	-0.11	-0.14	0.06	-0.04	0.32**
Supported retention	-0.05	0.08	1.00	-0.10**	0.38**	-0.19**	0.01	0.01
Food security	-0.01	-0.11	-0.10**	1.00	-0.08**	0.16**	-0.00	0.14**
Welfare income	-0.10**	-0.14	0.38**	-0.08**	1.00	-0.14**	0.00	-0.01
Sanitation	0.15**	0.06	-0.19**	0.16**	-0.14**	1.00	0.28**	0.06
Energy	0.00	-0.04	0.01	-0.00	0.00	0.28**	1.00	0.05**
Environment	-0.02	0.32**	0.01	0.14**	-0.01	0.06	0.05**	1.00

Table 10: Standardised effects of predictors on outcomes, including the estimated path coefficients (*β*), standard errors (SE), and *t*-values for each relationship of the eight composite factors for LQ13 learners in SQ45 schools

Predictor	*Outcome*	*β*	*SE*	*t-value*
Welfare income	Supported retention	0.33	0.02	15.54
Sanitation	Welfare income	-0.21	0.02	-9.45
Family structure	Sanitation	0.19	0.02	8.34
Sanitation	Supported retention	-0.16	0.02	-7.00
Sanitation	Food security	0.15	0.02	6.38
Family structure	Welfare income	-0.13	0.02	-5.45
Sanitation	Energy	0.13	0.02	5.30
Environment	Food security	0.12	0.02	4.97
Welfare income	Food security	-0.11	0.02	-4.51
Food security	Supported retention	-0.09	0.02	-4.09
Energy	Environment	0.10	0.02	4.02

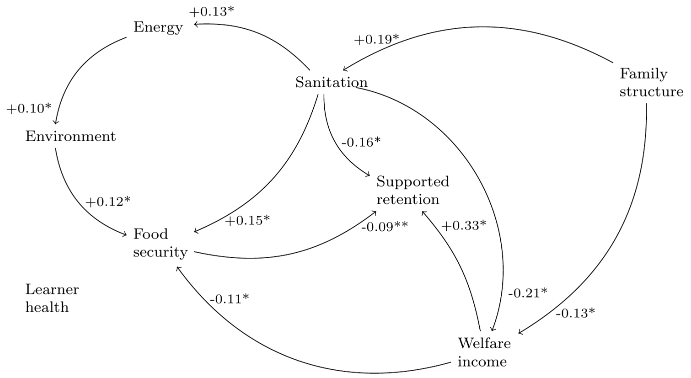


Figure 6: Relational mapping of the predictors and outcomes included in the estimated path coefficients as a network diagram of the eight composite factors for LQ13 learners in SQ45 schools

Feature importance analysis for across 62 features used XGBoost, with the grid search parameters detailed in Table 11. Model performance achieved a precision of 0.93, recall of 0.94, and an F1-score of 0.94 for predicting progressing learners, as shown in Table 12, indicating strong predictive capability.

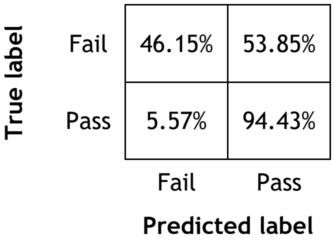
Table 11: Key gradient boosting classifier parameters for LQ13 learners in SQ45 schools

Parameter	*Value*
learning_rate	0.3
*n*_estimators	300
Max_depth	5
Min_samples_split	2
Min_samples_leaf	1
Loss	Log_loss
Subsample	1.0
Validation_freaction	0.1

Table 12: Classification metrics by class for LQ13 learners in SQ45 schools

Class	*Precision*	*Recall*	*F1-score*
Progress	0.51	0.41	0.49
Repeat	0.93	0.94	0.94

The model achieved 89.00% accuracy, accurately classifying 94.43% of progression and 46.15% of repetition instances. While precision and recall were strong for the progress class, 53.85% of repeat cases were misclassified, again indicating a need to improve minority class prediction. The results are presented in the confusion matrix in Figure 7.



True label	Fail (Predicted label)	Pass (Predicted label)
Fail	46.15%	53.85%
Pass	5.57%	94.43%

Figure 7: Confusion matrix of the machine learning model for LQ13 learners in SQ45 schools

Figure 8 shows the ranked feature importance from the machine learning results. Learner retention is the most influential factor, revealing the logical need for a learner to be retained for them to progress. Meals at school support academic success in higher-quintile schools, while age reveals the importance of grade-age alignment for effective learning.

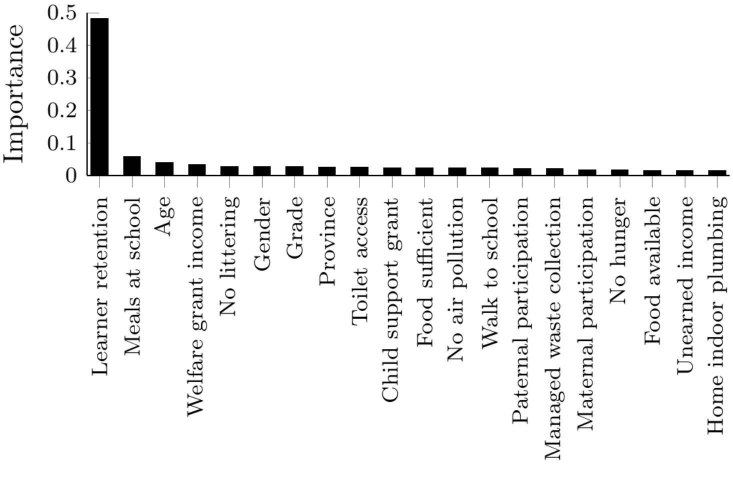


Figure 8: Feature importance for LQ13 learners in SQ45 schools

Figure 9 shows that model accuracy initially remains consistently high at around 90%, but declines gradually. Even after removing the seventh feature, accuracy exceeds 80%, thus highlighting the model's robustness against feature elimination in SQ45 schools.

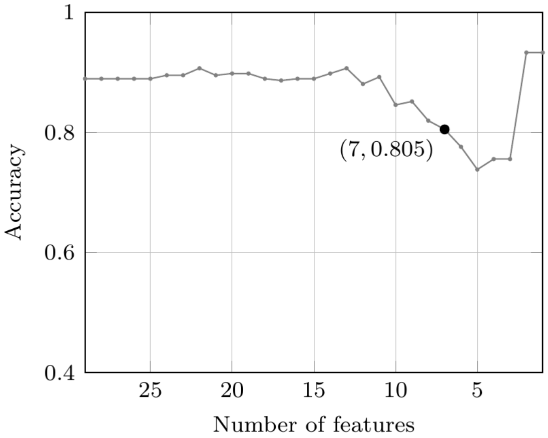


Figure 9: Change in learner progression prediction accuracy against number of features for LQ13 learners in SQ45 schools

5. DISCUSSION

For lower SES learners in SQ13 schools, the key findings reveal several interrelated factors that influence their well-being and education. Households in clean environments are more likely to be food- and energy-secure, while learners receiving welfare income tend to attend schools with support programmes. Traditional family structures are associated with better access to sanitation, and learners in schools with support programmes are likely to be healthier. These relationships highlight how environmental, economic, and familial factors collectively shape the educational and health outcomes of learners in SQ13 schools.

Lower SES learners in SQ45 schools who receive welfare income are more likely to attend schools with support programmes. In contrast, households with good sanitation infrastructure are less likely to receive welfare income or have learners attending schools that offer support programmes. In addition, traditional family structures are more likely to have good sanitation infrastructure.

Lower socio-economic learners in SQ13 and SQ45 schools may be compared. A composite factor score, calculated using binary responses weighted by relative factor loadings, provided deeper insights into their characteristics. Averaging the composite scores for all LQ13 learners in their schools enabled an overall assessment of each group. This approach reveals similarities and differences between the two learner groups. In this comparison, the resultant goodness scores closer to 1 reflect empowered learners, while scores near 0 indicate those who are worse off, as summarised in Table 13.

Table 13: Resultant goodness scores for the factor themes for LQ13 learners in SQ13 and SQ45 schools respectively

Factor	*SQ13*	*SQ45*
Family structure	0.58	0.57
Learner health	0.99	0.99
Support retention	0.92	0.53
Food security	0.80	0.84
Welfare income	0.61	0.38
Sanitation	0.36	0.65
Energy	0.94	0.95
Environment	0.76	0.82

An analysis of the average factor scores reveals notable differences between LQ13 learners in SQ13 and SQ45 schools, especially regarding supported retention, welfare income, and sanitation. LQ13 learners in SQ13 schools benefit from higher levels of supported retention and more frequent welfare income, probably because of feeding programmes that contribute to retention. Welfare income variation suggests that LQ13 learners in SQ13 schools may be more likely to apply for government grants, while those in SQ45 schools might rely on scholarships, reducing their need for welfare. In addition, LQ13 learners in SQ45 schools have access to better sanitation, probably because they reside near wealthier areas.

When examining feature importance, regular attendance and age-appropriate placement emerge as the most critical factors for academic success among LQ13 learners in SQ13 schools. This finding reflects the widespread issue of grade repetition and over-age learners in South African schools, often linked to delayed school entry, repeated grades, or interrupted schooling because of socio-economic problems. The slightly lower prioritisation of school meals in this context may indicate the presence of community-based support systems that address food insecurity, providing alternative nutritional support for learners and mitigating some of the effects of poverty on academic outcomes.

For LQ13 learners attending SQ45 schools, regular attendance and age-appropriate placement remain essential for academic success. However, school meals assume greater importance, suggesting that these learners experience higher levels of food insecurity than their more affluent peers. This reveals the critical role of feeding schemes in these schools, which are often not provided owing to the assumed affluence of the learners enrolled in these schools.

The findings show the enduring significance of social support systems in education. Targeted interventions are vital to promoting regular attendance and ensuring age-grade alignment, particularly through early intervention strategies that address delayed school entry and grade repetition. School feeding schemes are equally important in all educational contexts, contributing significantly to learners' well-being and academic performance. Financial grants remain essential for enabling access to education for learners who face economic hardships, highlighting the need for sustained investment in such support mechanisms to address persistent inequalities.

While Van der Heever et al. [10] recognised financial factors as crucial for improving education, their disregard of a learner's SES context camouflaged the key factors for lower socio-economic learners. For example, the age delta and the critical role of school meals, as discovered in this study, are essential for understanding the difficulties faced by these learners. This supports Spaull's observation of the socio-economic divide in South Africa, and shows that the factors that influence the so-called "average learner" differ significantly between lower and higher socio-economic groups [6]. Consequently, there is a clear need for SES-specific educational interventions to address these disparities effectively.

6. CONCLUSION

This study has explored the socio-economic disparities in the South African public high school system, focusing on learners from lower socio-economic backgrounds and their progression in both low and high socio-economic schools. Using a combination of statistical techniques such as FA, MRA, SEM, and ML, the findings have highlighted the multidimensional nature of learner adaptation across these socio-economic divides.

For learners in lower quintile schools, the results demonstrate the critical role of traditional family structures, welfare support, and school feeding programmes in fostering academic success. These factors contribute to better health, food security, and retention rates, showing the interconnectedness of economic, environmental, and social dimensions in shaping educational outcomes. In higher quintile schools, access to improved sanitation and environmental conditions has emerged as significant, although disparities persisted in welfare support and school retention strategies for learners from lower socio-economic households.

The study reaffirms the importance of targeted, context-specific interventions to address the unique difficulties faced by learners from disadvantaged backgrounds. Regular attendance, age-appropriate placement, and access to school feeding schemes were consistently highlighted as key drivers of academic progression. These findings align with the reality of educational bimodality, and reveal systemic gaps that require nuanced solutions that are tailored to the socio-economic realities of learners.

Comparing learners in the two school contexts, the results illustrate the need for equitable resource allocation and the sustained implementation of support mechanisms such as financial grants and feeding programmes. While the presence of these interventions in lower quintile schools has a measurable impact, the absence of or reduced access to similar programmes for disadvantaged learners in higher quintile schools highlights systemic assumptions about affluence and self-sufficiency that fail to accommodate their needs.

This study advances the understanding of how socio-economic factors influence learner outcomes in South Africa's dual education system. The integration of ML techniques, particularly XGBoost, into predictive modelling has provided robust insights into the factors that are most critical to learner progression. Features such as school retention and access to welfare income have been identified as key predictors, emphasising the need for comprehensive and scalable policy interventions.

Future research should build on these findings by examining the longitudinal effects of interventions that aim to reduce the socio-economic barriers in education. Expanding the scope to include qualitative insights from learners and educators could offer deeper contextual understanding, while comparative studies in rural and urban settings could further inform policy adjustments. In addition, refining predictive models to address class imbalances in ML outcomes would enhance their applicability to policy formulation.

REFERENCES

[1] **South African Government**, "Statistics South Africa on quarterly labour force survey quarter three 2023." [Online]. Available: https://www.gov.za/news/media-statements/statistics-south-africa-quarterly-labour-force-survey-quarter-three-2023-14 [Accessed: Nov. 28, 2024].

[2] **G. Mudiriza and A. de Lannoy**, "Profile of young people not in employment, education or training (NEET) aged 15-24 years in South Africa: An annual update." [Online]. Available: https://www.stateofthenation.gov.za/assets/pyei/resources/neet-youth-2023_saldru-wp.pdf [Accessed: Nov. 28, 2024].

[3] **S. van der Berg and B. Böhmer**, "COVID-19 and the crisis in South Africa's schools: Assessing the impact of school closures on the learning loss," in *Learning losses and recovery: Evidence from the COVID-19 pandemic*, Springer, 2023, pp. 153-169. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-69284-0_11 [Accessed: Nov. 28, 2024].

[4] **S. van der Berg, U. Hoadley, J. Galant, C. van Wyk, and B. Böhmer**, "Learning losses from COVID-19 in the Western Cape: Evidence from systemic tests," *Research on Socio-Economic Policy (ReSEP)*, Stellenbosch University, Stellenbosch, South Africa, Feb. 2022. [Online]. Available: https://ssrn.com/abstract=4212977 [Accessed: Nov. 28, 2024].

[5] **V. Reddy, L. Winnaar, F. Arends, A. Juan, J. Harvey, S. Hannan, and K. Isdale**, "The South African TIMSS 2019 Grade 9 results*: Building achievement and bridging achievement gaps*, 2022." [Online]. Available: https://www.timss-sa.org/wp-content/uploads/2022/02/FINAL-HSRC-TIMSS-Gr9-report_ELECTRONIC.pdf [Accessed: Nov. 28, 2024].

[6] **N. Spaull**, "Education in SA: A tale of two systems," *Politicsweb*, Aug. 31, 2012. [Online]. Available: https://www.politicsweb.co.za/news-and-analysis/education-in-sa-a-tale-of-two-systems [Accessed: Nov. 28, 2024].

[7] **K. Roux, S. van Staden, & M. Tshele**, "PIRLS 2021: South African highlights report," Centre for Evaluation and Assessment, University of Pretoria, 2023. [Online]. Available: https://www.up.ac.za/media/shared/164/ZP_Files/2023/piirls-2021_highlights-report.zp235559.pdf [Accessed: Nov. 28, 2024].

[8] **S. Slamang**, "A systems perspective on public high school management within the Western Cape," Honours research project, Department of Logistics, Stellenbosch University, Stellenbosch, South Africa, 2020. Available: https://scholar.sun.ac.za/server/api/core/bitstreams/706cb647-fc9a-42d3-b47b-7eeed74987c8/content [Accessed: Nov. 28, 2024].

[9] **E. Becker and L. Venter**, " Statistiese kartering van faktorverwantskappe vir leerdervordering in SA openbare hoërskole," *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie*, vol. 44, no. 1, pp. 64-74, 2025.

[10] **M. van der Heever, E. Becker, L. Venter, and J. F. Bekker**, "Using machine learning and agent-based simulation to predict learner progress for the South African high school education system," *South African Journal of Industrial Engineering*, vol. 34, no. 3, pp. 15–27, 2024.

[11] **N. O'Rourke and L. Hatcher**, *A step-by-step approach to using SAS for factor analysis and structural equation modelling*, 2nd ed. Cary, NC: SAS Institute, 2013.