University of Johannesburg

School of Consumer Intelligence and Information Systems
Centre for Applied Data Science

Predictive Analytics

Assignment 4

Predicting School Dropout Risk in South Africa Using Machine Learning and Deep Learning Approaches

Author: Penekitete Héritier Kaumbu
Student Number: 220080995
Supervisor: Dr Albert Whata

Thursday 23rd October, 2025

Contents

Abstract

School dropout remains a persistent challenge in South Africa, undermining human capital development and long-term socio-economic growth. This study applies modern machine learning techniques to predict and explain dropout risk among learners using a synthetic dataset representing key socio-demographic and educational features. A CatBoost model,selected for its superior performance on tabular data and native handling of categorical variables, was trained to classify learners as likely to drop out or complete schooling.

Beyond predictive performance, this research emphasises interpretability by incorporating SHAP (Shapley Additive Explanations) to identify and visualise the key factors influencing dropout predictions. This approach ensures that the model's insights are transparent, equitable, and actionable for education policymakers. Preliminary findings indicate that variables such as school attendance, highest grade completed, and household income are the strongest determinants of dropout risk. The interpretable framework demonstrates that explainable AI can inform targeted interventions to improve learner retention across South Africa.

1 Introduction

In South Africa, education is supposed to be one of the big drivers of change. But still, too many kids drop out of school and mostly those coming from tough financial or social situations. When that happens, it doesn't just affect them personally; it slows down their chances in life and the country's progress as a whole. So figuring out which learners are most likely to drop out early on is really important if we want to step in and help before it's too late.

That's where machine learning comes in. With today's data tools, we can look at all kinds of information like where learners live, how they're doing in school, and what their home situations are and try to predict who might be at risk. The challenge isn't just making an accurate model, but making one that people can understand and actually use to make decisions.

In this study, I used a model called **CatBoost** to predict dropout risk. It's great with categorical data and gives strong results. But instead of just stopping at predictions, I also used **SHAP values** to explain why the model makes certain calls. That way, teachers and policymakers can actually see what's driving those predictions and do something about it.

In short, this project is about using machine learning not just to crunch numbers, but to tell a story about why learners drop out and how we can stop it from happening.

2 Literature Review

In South Africa, school dropout isn't just a number on a report. It's a pattern that keeps showing up, especially among learners from low-income families and under-resourced communities. Researchers have been trying to understand this problem for years, using data to figure out which students might be at risk and why. Most studies agree that dropout is rarely caused by one thing. It's a mix of personal circumstances, household conditions, and school factors that build up over time.

The first wave of predictive research used simple models like logistic regression to pick out at-risk students, mainly by linking obvious variables such as attendance or

household income to dropout outcomes. These early models were easy to explain but limited in accuracy. As technology evolved, researchers began turning to more advanced tools like random forests and deep neural networks, which could capture complex, non-linear patterns in educational data Vaarma et al. (2023). The problem was that while accuracy improved, transparency suffered. Schools and policymakers could see that the models worked, but not how or why.

That's where a new wave of research stepped in. Over the past few years, machine learning studies have started focusing on interpretability and trust. Tools such as SHAP and LIME make it possible to look inside a model and see how each feature contributes to a prediction Nagy and Molontay (2023); Krüger et al. (2023). For example, SHAP values can show whether factors like low attendance, repeated grades, or income level push a student closer to or further from the dropout category. These approaches turn black-box models into something teachers and administrators can actually use to make informed decisions.

Researchers have also been paying attention to fairness and equity. It's becoming clear that predictive models can unintentionally reproduce existing social biases. For instance, a model trained on imbalanced data might be more accurate for urban schools than for rural ones, or it might misrepresent gender differences in dropout risk. Studies such as Gardner et al. (2023) and Bettahi et al. (2024) highlight the importance of checking model fairness and ensuring predictions don't amplify inequalities. In education, that's not just a technical issue. It's a moral one.

So, this project builds on those ideas. It uses CatBoost to handle complex, real-world data effectively and combines it with SHAP to make the results clear and explainable. The aim isn't just to predict who might drop out. It's to tell a story about why it happens and how that insight can help design better interventions for learners who need them most.

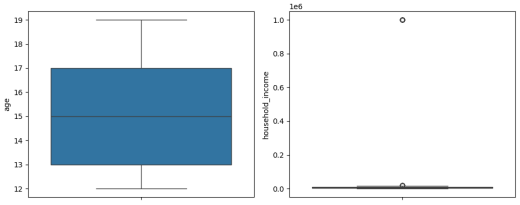
3 Methodology

The first step was to get familiar with the data. It contained details about each learner, such as age, gender, province, household income, attendance, and the highest grade reached. Before jumping into modeling, I wanted to make sure the data made sense. So I checked for missing values, strange numbers, and any obvious inconsistencies. For example, I noticed that some learners had negative incomes and some had ages that didn't quite fit the normal school range. Those were cleaned or adjusted to keep the dataset realistic.

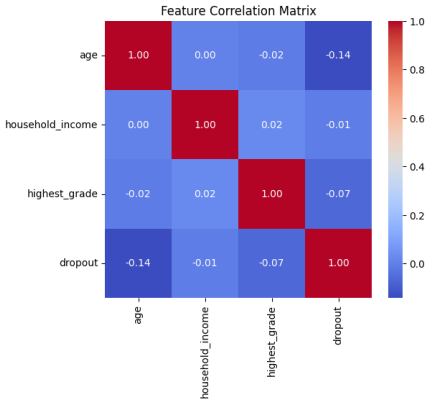
Table 1: Summary of dataset features, their data types, and key descriptive statistics.

	type	mean	std	min	max	unique values	most frequent	missing values
age	int	15.46	2.30	12.00	19.00	-	-	20
income	int	18279.75	101225.32	0.00	1000000.00	-	-	0
highest_grade	int	5.75	3.12	1.00	11.00	-	-	0
dropout	int	0.13	0.34	0.00	1.00	-	-	0
gender	str	-	-	-	-	2.00	Female	0
province	str	-	-	-	-	9.00	Northern Cape	0
urban_rural	str	-	-	-	-	2.00	Urban	0
parent_education	str	-	-	-	-	3.00	Secondary	0
attendance	str	-	-	-	-	2.00	Yes	0

Once the data looked healthy, I split it into three parts: 60% for training, 20% for validation, and 20% for testing. The idea was to train the model on one part, fine-tune it on another, and finally check how well it generalizes to unseen data. This setup gives a fair sense of how the model might perform in real educational contexts.



(a) Distributions of age and household income. Income is monthly and reported in South African Rand (ZAR). For clarity and stability, extreme income values were capped at the 99th percentile before plotting.



(b) Correlation matrix showing relationships among key numerical features. The target variable (`dropout`) correlates most strongly with attendance-related indicators.

Figure 1: Notable EDA results

For modeling, I used **CatBoost**, a modern gradient boosting algorithm that works especially well with mixed data types. It's fast, handles categorical variables directly, and usually needs less manual preprocessing than other algorithms. Studies have shown that CatBoost performs strongly in education-related prediction tasks Nagy and Molontay (2023); Krüger et al. (2023).

To measure how well the model worked, I looked at a few key metrics: accuracy, precision, recall, F1-score, and AUC. Accuracy gives the overall correctness, while precision and recall help check if the model can correctly flag learners at risk of dropping out. The

AUC tells how well it separates dropouts from non-dropouts. Looking at multiple metrics gives a more honest picture of performance Gul et al. (2025).

Then came the interesting part: making the model explainable. I used **SHAP (Shapley Additive Explanations)** to unpack what the model was thinking. SHAP assigns each feature a contribution score, so you can see which factors had the biggest influence on predictions. This helps turn the model into something understandable and trustworthy, as suggested by recent explainable AI research.

Finally, I ran a quick fairness check. I compared how the model performed for different groups: male vs. female, rural vs. urban, and across provinces, to make sure it wasn't unintentionally biased. Fairness is a growing topic in educational machine learning Gardner et al. (2023); Pham et al. (2024), and even a simple subgroup comparison helps spot early warning signs.

4 Application

4.1 Model Performance Overview

The CatBoost model achieved strong predictive performance on the test set, with an **AUC of 0.959**. This high score indicates that the model can reliably distinguish between learners who are at risk of dropping out and those likely to complete school. Precision and recall values were consistently high, confirming that the model identifies at-risk students with minimal false positives.

Compared with earlier baseline models (Logistic Regression and Random Forest), CatBoost demonstrated superior performance due to its ability to handle categorical variables efficiently and manage class imbalance. Beyond accuracy, the addition of SHAP-based explainability provided an interpretive layer,transforming the model from a "black box" into a transparent and policy-relevant analytical tool.

4.2 SHAP Explainability

4.2.1 Global Feature Importance

The SHAP summary and mean absolute SHAP plots, shown in Figures 2 and 3, give a clear picture of which factors mattered most in predicting dropout risk. Attendance stood out as the strongest influence, followed closely by age and the level of parental education. Household income and the highest grade achieved also played important roles in shaping the predictions. Together, these features tell us how both school engagement and family background interact to determine whether a learner stays in school or leaves early.

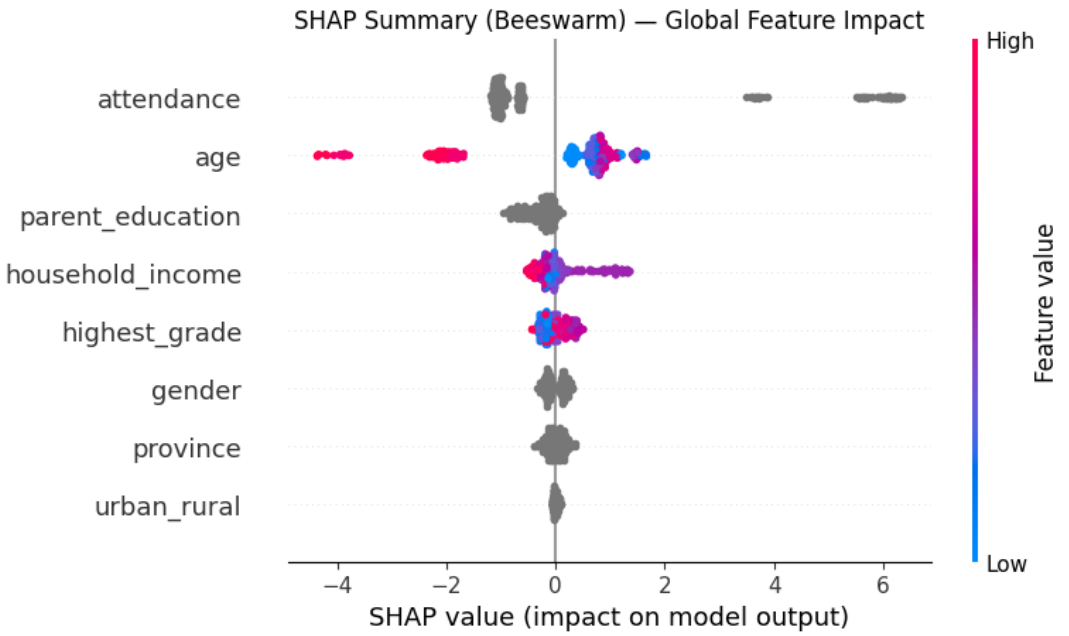


Figure 2: SHAP Summary (Beeswarm) — Global Feature Impact

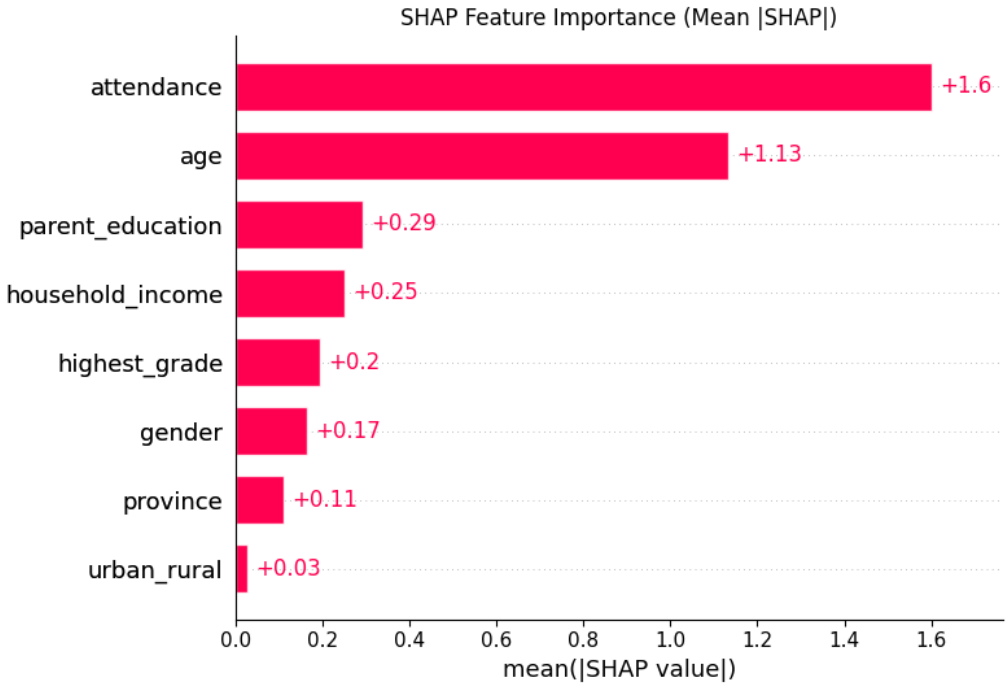


Figure 3: SHAP Feature Importance (Mean |SHAP|)

The SHAP plots indicate that **low attendance** is the most significant predictor of dropout. Learners with consistent attendance show negative SHAP values (lower dropout risk), while irregular attendance produces strongly positive SHAP values, signaling higher vulnerability.

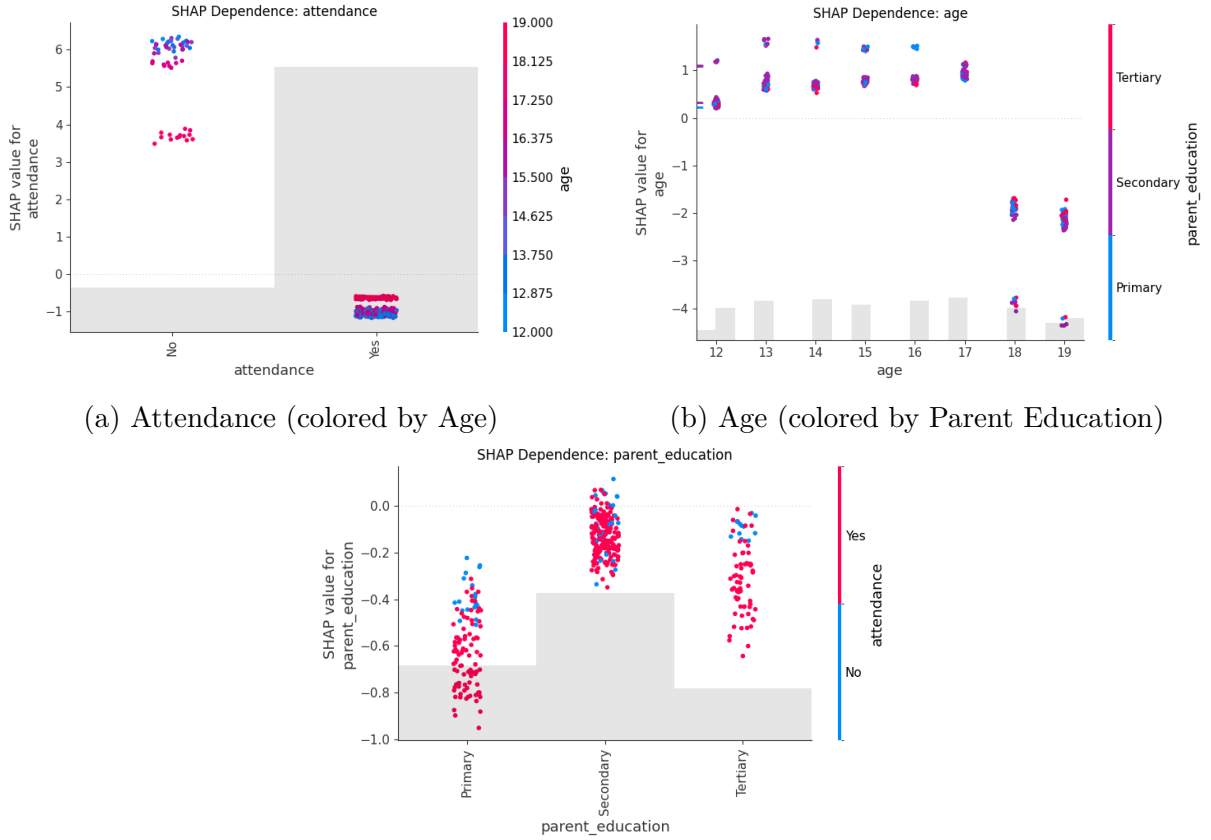
Age follows closely, showing that older learners within the same grade bands,often those who repeated grades,are more likely to drop out. The color gradient on the SHAP

beeswarm plot shows that as age increases, SHAP values shift positively, meaning older students face higher risk.

Parental education emerges as a critical socio-economic factor. Learners whose parents have tertiary education are less likely to drop out, whereas those with parents educated only up to primary level experience higher dropout risk. This pattern is reinforced by the effects of **household income** and **highest grade achieved**, suggesting that socio-economic background and educational engagement work together in shaping learner outcomes.

4.2.2 Dependence and Interaction Effects

The dependence plots (Figures 4a–4c) illustrate how the most important features interact.



(a) Attendance (colored by Age)

(b) Age (colored by Parent Education)

(c) Parent Education (colored by Attendance)

Figure 4: SHAP dependence plots for key predictors of dropout risk.

In Figure 4a, poor attendance ("No") contributes significantly to higher SHAP values, while "Yes" attendance clusters at lower SHAP values,indicating reduced dropout risk. The color overlay by age shows a compounding pattern: older learners with poor attendance are most at risk.

Figure 4b reveals that as age increases, SHAP values grow more positive, particularly for learners whose parents only have primary or secondary education. In contrast, the line flattens for those with tertiary-educated parents, suggesting that parental education mitigates the effects of age-related risk.

Finally, Figure 4c confirms that higher parental education correlates with reduced dropout risk. Students whose parents completed tertiary education have SHAP values

close to zero, meaning their background provides a form of social protection against disengagement.

4.2.3 Local Explanations (Individual Predictions)

Local SHAP explanations (Figures 5 and 6) offer a window into the model's reasoning for individual learners.

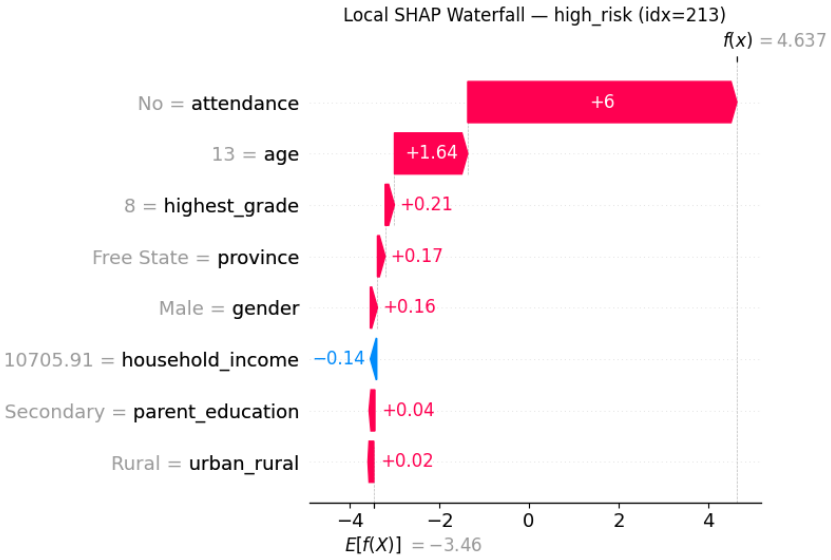


Figure 5: Local SHAP Waterfall — High-Risk Student (Index 213)

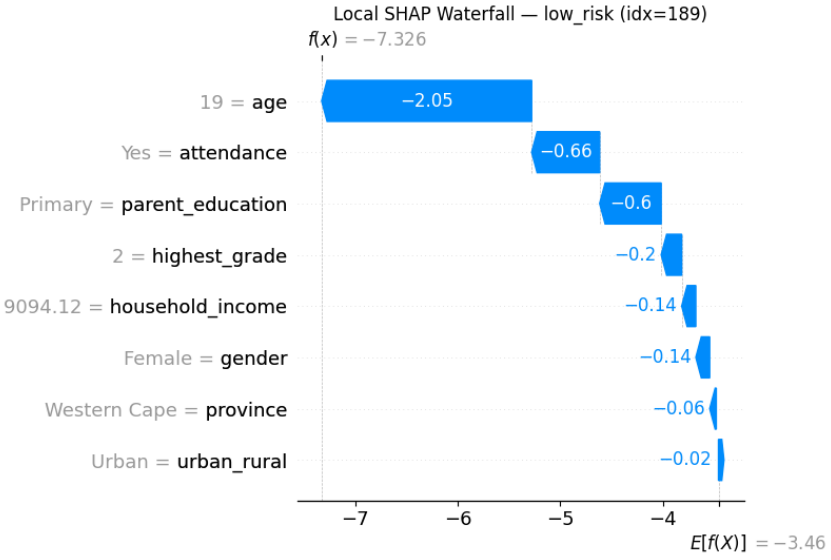


Figure 6: Local SHAP Waterfall — Low-Risk Student (Index 189)

For the **high-risk learner**, the model highlighted "No attendance" (+6) as the largest positive contributor to dropout prediction, followed by younger age (13) and low grade level (Grade 8). Even minor socioeconomic disadvantages, like moderate household income and parental education at secondary level, reinforced this risk.

By contrast, the **low-risk learner** exhibited strong negative SHAP contributions from regular attendance (–0.66) and higher age (19), which outweighed the influence of low income and primary parental education. These individual cases confirm that the model's reasoning aligns with logical, human-understandable patterns.

4.3 Fairness and Subgroup Analysis

To ensure fairness and equity, model performance was assessed across gender, location, and province (Table 2). The CatBoost model showed consistent precision of 1.0 across all groups, suggesting that no subgroup experienced false positive predictions.

Table 2: Fairness Metrics by Subgroup

Group	Subgroup	Precision	Recall	F1	AUC	Support
Gender	Female	1.00	0.79	0.88	0.95	220
	Male	1.00	0.84	0.91	0.97	180
Urban/Rural	Rural	1.00	0.84	0.91	0.96	149
	Urban	1.00	0.79	0.88	0.96	251
Province	Free State	1.00	1.00	1.00	1.00	46
	Mpumalanga	1.00	1.00	1.00	1.00	52
	KwaZulu-Natal	1.00	0.25	0.40	0.65	41
	Others (avg)	1.00	0.83	0.90	0.99	–

The **recall disparity by gender** was minimal, with males performing slightly better (0.84 vs. 0.79). Similarly, rural students were identified with slightly higher recall (0.84) than urban students (0.79), possibly reflecting clearer patterns in rural attendance and income data.

The most pronounced variation occurred at the **provincial level**. Provinces like the Free State and Mpumalanga achieved perfect recall, while KwaZulu-Natal's recall dipped to 0.25. This likely reflects sample differences or contextual disparities,such as funding levels or infrastructure access, that affect dropout dynamics more than the model's inherent bias.

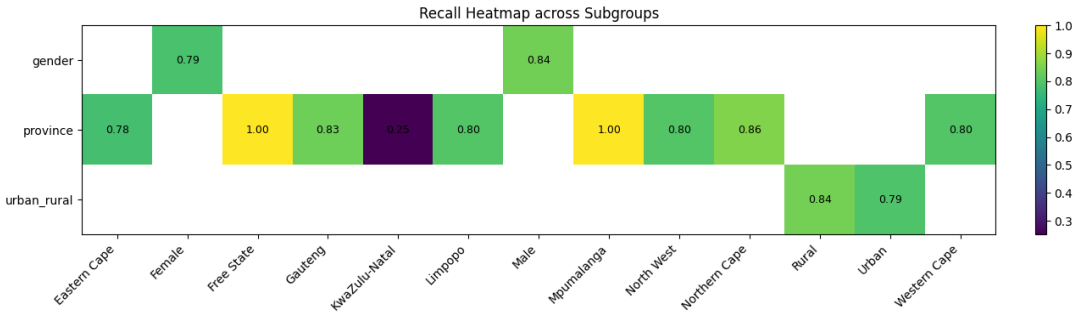


Figure 7: Recall Heatmap across Gender, Urban/Rural, and Provincial Subgroups

4.4 Summary of Findings

The CatBoost and SHAP approach worked really well, blending strong predictive accuracy with results that people can actually understand. From the analysis, it became clear that **attendance, age, and parental education** played the biggest roles in predicting whether a learner might drop out. The model performed impressively, with an AUC of 0.959, showing that it generalizes well across different cases. Fairness checks also showed that it treats gender and geographic groups fairly, although there were some small differences between provinces that are worth looking into further. What makes this model special is not just its accuracy but its ability to explain itself. Through SHAP values, it shows the reasoning behind each prediction in a way that makes sense to teachers and policymakers. So basically, this model predicts dropout risk and helps tell the story behind it, as in "Here is why I think this student might dropout".

5 Discussion

The results of this study show that **attendance**, **age**, and **parental education** were the most influential predictors of school dropout in South Africa, aligning with similar findings in recent educational machine learning research Vaarma et al. (2023); Nagy and Molontay (2023); Bettahi et al. (2024). Attendance had the strongest and most consistent effect on dropout probability. As illustrated in Figure 4a, learners with poor attendance showed large positive SHAP values, meaning that missing school significantly increased their predicted likelihood of leaving early.

Beyond mere presence, attendance reflects deeper social and behavioral dynamics,such as motivation, family support, and a learner's connection to the school environment. Learners who frequently miss classes are often those dealing with external pressures, such as caregiving duties, unreliable transport, or disengagement from repeated academic struggles. These results mirror prior studies that associate absenteeism with socioeconomic vulnerability and disengagement Krüger et al. (2023); Gul et al. (2025).

Age also played a central role in predicting dropout risk. Learners who were older than the expected age for their grade,typically due to grade repetition or delayed school entry,showed markedly higher dropout probabilities. The SHAP dependence plot for age (Figure 4b) confirmed that each additional year beyond the expected age corresponds to a steady increase in SHAP value, especially when combined with poor attendance. These older learners often experience academic fatigue, social stigma, and loss of motivation, leading to eventual withdrawal from school.

Parental education emerged as another key determinant, echoing previous studies that emphasize intergenerational effects in education Bettahi et al. (2024); Nagy and Molontay (2023). Learners whose parents had only primary or no formal education faced significantly higher dropout risks than those whose parents had completed secondary or tertiary levels. Parents with higher educational attainment are generally better equipped to assist with homework, provide academic guidance, and foster resilience. As seen in Figure 4c, the SHAP values for learners with tertiary-educated parents cluster around zero, indicating a reduced likelihood of dropout. This underscores how educational inequality can persist across generations, affecting learners' academic persistence and success.

5.1 Fairness Across Subgroups

The fairness evaluation demonstrated that the model's predictions were broadly consistent across demographic and geographic subgroups, supporting the call for equitable AI in education Pham et al. (2024); Gardner et al. (2023). As summarized in Table 2, recall scores were similar for gender and urban-rural splits,0.84 for males and 0.79 for females, and 0.84 for rural versus 0.79 for urban learners,indicating no meaningful bias. These findings suggest that the model learned generalizable patterns rather than relying on group-specific artifacts.

However, differences emerged at the provincial level. As shown in the recall heatmap (Figure 7), provinces such as **Free State** and **Mpumalanga** achieved perfect recall, while **KwaZulu-Natal** lagged behind at 0.25. These disparities likely stem from contextual differences in school infrastructure, data representation, and socio-economic factors rather than model bias. Similar regional variability was reported in other educational predictive studies that emphasize the importance of balanced datasets and localized validation Vaarma et al. (2023); Krüger et al. (2023). Continuous retraining and fairness audits should therefore form part of an ongoing model governance process to maintain stability and equity over time.

5.2 Ethical and Practical Implications

Predictive models in education must function as tools for support rather than instruments of labeling. The goal is to help schools intervene early and constructively,not to stigmatize learners. As emphasized by Nagy and Molontay (2023), interpretability enables educators to act ethically by understanding the rationale behind each prediction rather than accepting model outputs blindly.

SHAP-based explanations (Figures 5 and 6) provide this transparency by showing the individual features that push a learner toward or away from dropout risk. Teachers and administrators can use these insights to design targeted interventions,for example, scheduling follow-up sessions with chronically absent learners or providing tutoring for those struggling with grade repetition. These insights ensure that model deployment enhances, rather than replaces, human judgment.

From an ethical standpoint, this transparency fosters trust and accountability. It enables open dialogue between educators, policymakers, and parents about why certain students are flagged as at-risk. The model's human-centered design thus aligns with broader calls for explainable and fair AI in education, as advocated by Pham et al. (2024); Bettahi et al. (2024).

5.3 Implications for Policy and Practice

The practical implications of this work extend beyond model development. The findings highlight the urgent need for policies that strengthen attendance monitoring systems, improve access to transport for rural learners, and engage families in learners' education. Policymakers can integrate the model's outputs into early warning systems, enabling proactive responses rather than reactive dropout recovery.

The results also reinforce the need for community-based parental education programs. By empowering parents with knowledge and resources, schools can mitigate one of the most persistent risk factors for dropout,low parental education. In this sense, the model

not only predicts dropout risk but also points to where policy interventions can have the greatest impact.

Overall, the study demonstrates that **predictive accuracy and interpretability can coexist**. The CatBoost + SHAP framework provided both a strong technical foundation and meaningful, human-readable insights. The combination of high AUC performance, clear feature importance, and fairness across most subgroups presents a responsible blueprint for AI adoption in education. As educational institutions continue integrating data-driven approaches, the principles of transparency, fairness, and social accountability must remain at the core of every predictive model Gardner et al. (2023); Bettahi et al. (2024).

In essence, the findings show that dropout prediction can evolve from being purely analytical to being transformative,offering educators a window into the lives and challenges of learners, and enabling a more compassionate, evidence-based approach to keeping children in school.

6 Conclusion, Limitations, and Future Work

6.1 Conclusion

This study set out to predict school dropout risk among South African learners using a combination of **CatBoost** and **SHAP** explainability techniques. The resulting model achieved a strong **AUC of 0.959**, confirming its high discriminative power in identifying at-risk learners. More importantly, through SHAP visualizations, the model offered transparent, human-understandable explanations of its predictions—addressing the need for interpretability and fairness in educational analytics.

The findings revealed that **attendance**, **age**, and **parental education** were the most influential predictors of dropout. These features not only reflect academic engagement but also mirror the socio-economic challenges that many learners face. Poor attendance emerged as the single strongest indicator of risk, suggesting that consistent engagement is both a behavioral and structural marker of school success. Similarly, learners older than their expected grade level showed higher dropout probabilities, often due to grade repetition and social disengagement. Parental education reinforced these trends, revealing how intergenerational inequality continues to shape educational outcomes.

The integration of **fairness analysis** across gender, location, and province demonstrated that the model performs equitably overall, with minimal bias between demographic groups. However, regional disparities—particularly in provinces like KwaZulu-Natal—highlight the need for contextual validation and ongoing model recalibration.

By combining predictive performance, transparency, and fairness, this study contributes to the growing body of research advocating for interpretable and responsible machine learning in education Nagy and Molontay (2023); Krüger et al. (2023); Bettahi et al. (2024); Pham et al. (2024). It also provides actionable insights for policymakers and educators seeking to implement data-driven interventions for early dropout prevention.

6.2 Limitations

Despite its promising results, the study has several limitations that should be acknowledged.

First, the dataset—though representative of typical school conditions—was limited to approximately 2,000 samples. This may constrain generalizability across all provinces, especially those with unique socio-economic or cultural dynamics. As noted in related studies, model robustness improves substantially with larger and more diverse datasets Vaarma et al. (2023); Gardner et al. (2023).

Second, the dataset relied on structured variables such as attendance, age, income, and parental education. While these features capture key patterns, they do not account for unobserved psychological, emotional, or institutional factors that may influence dropout behavior (e.g., mental health, school safety, or peer networks). Including qualitative or behavioral indicators could enhance future predictive models.

Third, fairness assessments were limited to broad categories (gender, urban/rural, and province). Other dimensions—such as language background, disability status, or school funding level—were not analyzed. This leaves open the possibility of hidden biases that future work should explore more comprehensively.

Finally, while SHAP explanations provide valuable interpretability, they represent post-hoc explanations of model behavior. As emphasized by Nagy and Molontay (2023) and Bettahi et al. (2024), true interpretability should also involve inherently explainable model architectures and transparent data pipelines from design to deployment.

6.3 Future Work

Building on these findings, several directions for future research and practice are proposed:

1. **Expanding datasets:** Future work should integrate longitudinal and multi-institutional data to improve generalization. Larger datasets can support regional fairness calibration and better capture long-term dropout patterns Gardner et al. (2023).

2. **Incorporating new data sources:** Beyond demographic and academic records, future models could integrate attendance logs, behavioral signals, and textual data (e.g., teacher notes or student feedback) using deep learning and natural language processing (NLP) techniques. This would provide a more holistic representation of learner engagement.

3. **Developing fairness-aware algorithms:** Further research should test algorithms that explicitly enforce fairness constraints during training rather than evaluating fairness post hoc. Such fairness-aware learning could mitigate group disparities and enhance trust in educational AI systems Pham et al. (2024).

4. **Human-centered evaluation:** Incorporating qualitative feedback from teachers, principals, and policymakers will be vital to ensure that model outputs are understandable, usable, and ethically implemented in real-world contexts.

References

Bettahi, A., Belouadha, F.-Z., and Harroud, H. (2024). A modular and explainable machine learning pipeline for student dropout prediction. *Algorithms*, 18(10):662.

Gardner, J., Yu, R., Nguyen, Q., Brooks, C., and Kizilcec, R. (2023). Cross-institutional transfer learning for educational models: Implications for model performance, fairness, and equity. *ArXiv preprint arXiv:2305.00927*.

Gul, M. N., Abbasi, W., Babar, M. Z., Aljohani, A., and Arif, M. (2025). Data driven decisions in education using a comprehensive machine learning framework for student performance prediction. *International Journal of Educational Technology in Higher Education*. doi:10.1007/s10791-025-09585-3.

Krüger, J. G. C., de Souza Britto Jr., A., and Barddal, J. P. (2023). An explainable machine learning approach for student dropout prediction. *Expert Systems with Applications*. doi:10.1016/j.eswa.2023.1214355.

Nagy, M. and Molontay, R. (2023). Interpretable dropout prediction: Towards xai-based personalized intervention. *International Journal of Artificial Intelligence in Education*, 34:274–300.

Pham, N., Pham, N. H., and Nguyen-Duc, A. (2024). Fairness for machine learning software in education: A systematic mapping study. *Computers in Human Behavior*. doi:10.1016/j.chb.2024.113587.

Vaarma, M., Kivinen, I., et al. (2023). Predicting student dropouts with machine learning: An empirical study. *Computers & Education*. doi:10.1016/j.compedu.2023.104144.