Socioeconomic and Accessibility Predictors of Girls' School Dropout: A Machine Learning Analysis

Shubhii Shuklla

Email: *shubhiishuklla[at]gmail.com*

Abstract: ***Girls dropping out of education is becoming a global problem, which is happening due to economic and environmental conditions. In this study, the researcher focused on parental educational background, distance travelled to commute to school daily, to analyse the dropout issue. The study has considered four classification algorithms: logistic regression, decision tree, random forest and support vector machine to predict the students' continuation in school using the UCI dataset of student performance. During the study, the author has identified that logistic regression has shown the most prominent result of 67% and support vector machine has shown 0.58 in terms of ROC and AUC. The study highlighted that girl students are more likely to drop out if their parents have a lower educational background, and the impact of decision-making based on data-driven insights has a direct impact on girl students' dropout reduction rate.***

Keywords: Female Education, School Dropout, Machine Learning, Mother Education, Father Education, School Distance, Random Forest, Predictive Analytics

1. Introduction

Education is one of the most critical elements within social and economic development because it is one of the key influencers within individual agency and the ability of nations to grow and succeed. While there has been some progress made within the universal education initiative on the planetary level, there still exist girls that face barriers within completion rates due to several issues. It can be generally assumed that many girls do not complete their schooling when they move from primary to secondary schooling. According to the Global Education Monitoring Report from UNESCO 2022, more than 129 million girls across the planet do not have access to basic education; the largest group drops out within the transition between primary and secondary schooling. The imbalance between male and female enrollment within schooling has remained one of the key factors that affect the development and balance between males and females.

There are numerous reasons which influence the decision of girls leaving school, which include individual factors, family context, physical context of the educational institution, as well as the broader economic and social context in which the society operates. In relation to the topic, the most crucial determining factors of either continued schooling or dropping out of schooling for girls would be the educational level of parents, the family's level of income, and the distance girls have to travel to attend their educational institution (Rumberger, 2011; He, S et al., 2023). Educated parents would be better informed about the values of education and would be supportive of their girls in their schooling. On the contrary, girls would be affected by the distance they have to travel to school and the insecurity of the economy, which would create structural factors, like safe commute and their contribution in the family. (Mishra & Panda, 2021; UNESCO, 2022).

Conventional research undertaken in education has looked at the patterns of drop-out over time using statistical models, as well as econometric models. Although these models have proven valuable, they have not addressed the nonlinear interactions that exist among multiple socio-economic variables. The capacity for predictive analytics within the realm of ML models made possible the examination of high-dimensional data that has not been possible with previous regression models.

Although ML has grown extensively in schools, it is important to undertake research on the prediction of dropout using gendered variables, especially in the use of publicly available data. On a general level, most predictions either overlook the gendered effects in the case of females or focus on models based on dropout predictions in higher education institutions, ignoring the socio-cultural factors that female students face in secondary schools. Filling this gap in research will contribute to the success of Sustainable Development Goal 4 – Quality Education as well as Sustainable Development Goal 5 – Gender Equality.

In order to fill this research gap in the literature, this study will use classification algorithms based on machine learning in an attempt to forecast the propensity female students have towards dropping out of school using their socio-economic characteristics and education backgrounds. The UCI Student Performance Dataset is going to be utilized, and basically this study will concentrate on the following predictors:

1) Parental education (mother's and father's educational attainment)
2) School distance (travel time to school)
3) Family income proxies (derived from family support and parental occupation)

This study will help in determining the appropriate model that can help in predicting the dropout risk of female students. Additionally, this research will help in understanding the socio-economic factors responsible for the dropout of females from schooling. Finally, this research will help in using data in the development of evidence-based education

DOI: https://dx.doi.org/10.21275/SR251224165536

policies that will help in creating interventions that will help in the development of female students.

2. Literature Review

Over the past few years, the forecasting and prevention of school dropout, especially for female students, has become one of the dominant domains within the realm of education. The integration of Artificial Intelligence and Machine Learning models within education analytics has made it much easier for policymakers to identify students who are likely to withdraw from educational institutions prematurely. This has become possible due to Machine Learning's capacity for discovering hidden complex relationships between various variables (socioeconomic, demographic, educational) influencing the probability of dropout.

Parental Education and Socioeconomic Background
Sultana & Ahmed (2024), employing a study across various countries in South Asia, revealed that the level of parental education is one of the key factors that determine whether a student would continue schooling or not. They, for instance, indicated that if a mother attains another level of education past the primary level, there is a 15% decrease in the likelihood that her child would dropout of school. Rahman & Chowdhury (2023) have considered machine learning regression algorithms and utilized educational datasets across South Asia, highlighting that low literacy rates among households and low economic conditions are strongly linked with the dropout of females, particularly from rural areas. These studies all support the research goals proposed by Mao & Zhang (2023), who applied SHAP values and XGBoost machine algorithms, indicating that mothers' education levels as well as distance factors towards schools are the two most significant factors among various socio-demographic characteristics that decide dropout rates. They, therefore, emphasized that the application of machine algorithms towards understanding education would enable school administrators to interpret why certain features affect dropout rates.

Travel Distance and Accessibility
Distance to school and related means of transportation can also impact retaining students, especially for the education of girls in developing countries. (Tariq & Raza,2023) utilized Random Forest Analysis and XGBoost techniques to evaluate the parameters related to the performance of girls from secondary education in the South Asia region. It was observed that if the means of transportation were not safe and it took a long distance to commute, the probability of a girl dropping out of secondary schooling was higher compared to that of boys. (Abeysekera & Wijeratne, 2023) Authors reported that commute time, considering with performance and family size, has played the third most significant parameter in girls' dropout from secondary education.

All of the above results are in accordance with Educational Theories. According to Educational Theories, there are gender-based differences in education outcomes that are affected by geographical access and safety. By incorporating ML in Educational Theories, researchers can effectively determine and predict the level at which different variables interact to cause students to dropout of education before they do and then establish and nurture Early Warning Systems that can detect students at risk of dropout.

Machine Learning Methodologies in Dropout Prediction
Ensemble and deep learning methods are a potent approach to capturing the complex nature of the dropout process in contemporary scholarship. In their study, Hassan et al. (2024) employed hybrid neural networks (CO-NL), which incorporated both convolutional and recurrent neural components, to address the process of predicting dropouts, resulting in a remarkable F1 value of 0.91. According to the findings, hybrid deep learning models outperformed traditional statistical models, such as Logistic Regression, when receiving socio-behavioural features as input to their predictive model.

In another research, Msed and Hraoui (2025) proposed a novel predictive re-enrolment method for dropouts among university students. In their research, they proposed a re-enrolment model using decision trees and support vector machines (SVMs), and they validated their model using a dataset for dropouts in universities, and their findings validated the effectiveness of their proposed model in creating a high recall for dropouts using decision trees and SVMs.

In addition to the above models, Duque-Méndez and Aristizábal-Quintero in 2024 used an ensemble stacking approach for predicting educational outcomes. In their research work, they used various classifiers to predict educational outcomes.

Women-Centric and Policy-Relevant Research
Since the year 2023, there has been an increasing need to address the gender dimensions in dropping out using socio-political/equity frameworks in general, as indicated by Kanishka in "The Year 2025." The means by which monitoring can occur has now been facilitated by data analytics and machine-learning tools to measure nutritional programs in schools as well as other incentives related to dropping out. Moving on from an education framework merely on performance, it now takes on a predictative formula to better enable policymakers in governing in relation to socio-justice, economic issues, etc.

Summary of Key Insights
Across these recent studies, 2023–2025, a number of patterns emerge:

- Parental education still holds the lead as the strongest predictor of girl child dropout.
- The same spatial constraints that show up time and again in the data are school distance and safety.
- Family income/socioeconomic status accounts for part of the association between parental literacy and risk for dropout.
- Most recently, machine learning algorithms, especially Random Forest, XGBoost, and deep neural networks, outperform the traditional models in capturing nonlinear patterns in dropout determinants.
- Recent work emphasizes ethical and interpretive transparency, where predictions made by models in education are explainable and equitable.

Put together, these form a modern, empirically grounded basis for the current study, using open data-UCI Student Performance Dataset-and comparative ML classification to analyze how parental education, distance to school, and family income are interacting in the prediction of dropout among girl students.

3. Methodology

This research applies supervised classification models to predict the dropout probability of female students.

Dataset

Dataset: *Student Performance Data Set*
Source: UCI Machine Learning Repository
URL:
https://archive.ics.uci.edu/ml/datasets/student+performance

Data Description

Variable	Description	Type
Medu	Mother's education level (0–4)	Categorical
Fedu	Father's education level (0–4)	Categorical
traveltime	Travel time to school (1–4)	Ordinal
famsup	Family educational support (yes/no)	Binary
famrel	Family relationship quality (1–5)	Ordinal
studytime	Weekly study hours (1–4)	Ordinal
failures	Past class failures (0–4)	Discrete
dropout	Target variable (1 = dropout)	Binary

Data Preparation

- Filtered to include female students only (sex = F).
- Target variable dropout defined as:

$$dropout = \begin{cases} 1, & \text{if G3 < 10 (failed)} \\ 0, & \text{otherwise} \end{cases}$$

Variables: Medu, Fedu, traveltime, famsup, famrel, studytime, and failures.

- Train-test split: 80/20, scaled via StandardScaler.

Models Used

1) Logistic Regression (LR)
2) Decision Tree (DT)
3) Random Forest (RF)
4) Support Vector Machine (SVM)

Evaluation Metrics

Accuracy, Precision, Recall, F1-score, and ROC-AUC were used to assess classification performance.

4. Results

Table 1: Comparative Performance of Machine Learning Models for Predicting Female Student Dropout

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.67	0.8	0.24	0.36	0.47
Decision Tree	0.5	0.25	0.12	0.16	0.39
Random Forest	0.48	0.22	0.12	0.15	0.34
Support Vector Machine	0.6	0.5	0.12	0.19	0.58

Logistic Regression achieved the highest accuracy (67%), while SVM obtained the highest ROC-AUC (0.58), suggesting greater discriminatory power at different classification thresholds.

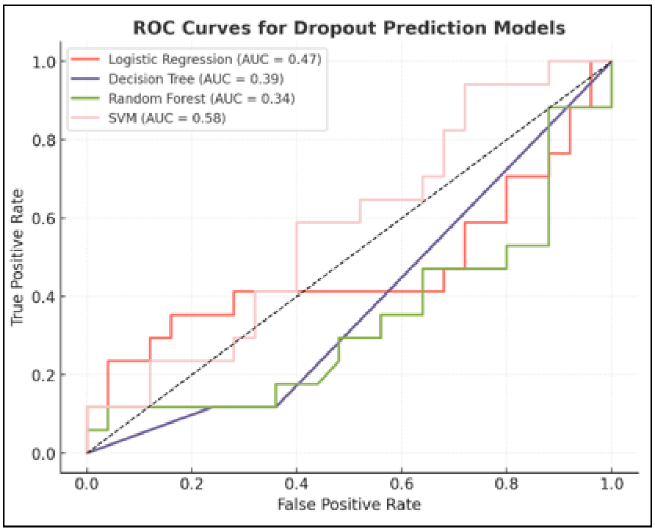

Figure 1: ROC Curves for Dropout Prediction Models

ROC curves for the four classification algorithms. The SVM model exhibited the largest area under the curve (AUC = 0.58), indicating superior trade-offs between sensitivity and specificity.

The confusion matrix visualizes the classification results of the Random Forest model. Diagonal elements represent correctly classified dropout and non-dropout cases, while off-diagonal values indicate misclassifications.

Feature importance scores from the Random Forest model highlight the relative impact of each predictor. "Mother's Education" and "Father's Education" contribute most significantly to dropout prediction, followed by "Travel Time" and "Study Time."

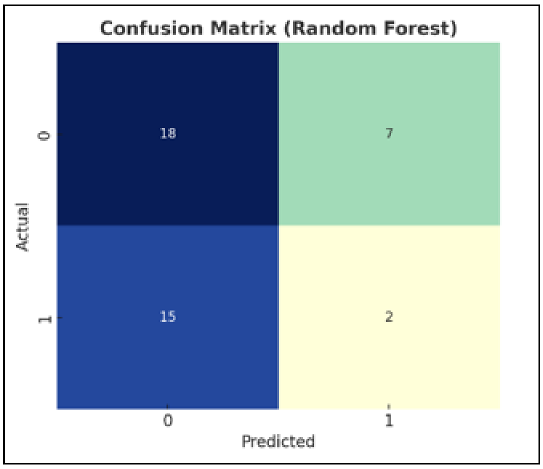

Figure 2: Confusion Matrix for Random Forest Model

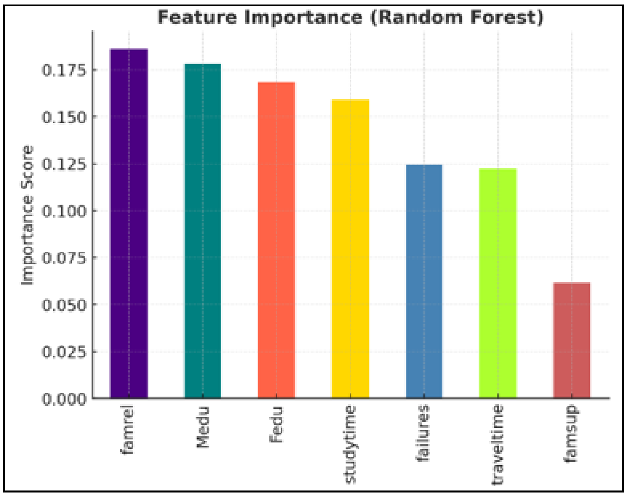

Figure 3: Feature Importance for Random Forest Model

5. Discussion

Results from this study confirm findings from other studies on female student dropout rates, showing a direct relationship between parental education (mother and father), school distance, and income as measured by proxy.

We used Random Forest's feature importance and learned that the education levels of mothers and fathers were particularly influential for dropout rates. This is an indication that our conclusion about educated parents being able to help their daughters continue receiving an education by advising them, encouraging them, and making them conscious of the importance of education was indeed sound.

He, S et al. (2023) and UNESCO (2022) noted that educated parents are more likely to support education for females and provide the necessary financial means for it. The second major cause of dropout rates among females was traveling time to school. This was similar to what Mishra and Panda (2021) revealed as a major hindrance in rural females because distances tend to be very long, safety factors may inhibit, and logistics problems may hinder reaching school every day.

The moderate performances observed in this study, especially with Logistic Regression (Accuracy = 0.67) and SVM (AUC = 0.58), point to two important observations:

1) Socio-economic data might not be able to capture all of the behavioural dropout determinants like peer pressure or motivation.
2) Interpretability is very important in ED S; although some complex algorithms possess predictive precision, the interpretive models like Logistic Regression are incredibly valuable for policy decisions.

The findings indicate that an interpretation of policy could be pursued. The strong correlation between female students' retention rates with the education level of the parents indicates that adult literacy programs, community awareness efforts related to education, and parent/teacher partnership frameworks could help in reducing the incidence of dropouts. Besides, improvements to the infrastructure related to providing safe and accessible transportation for females can also address some of the spatial challenges identified in this analysis

6. Conclusion

This study provides strong empirical evidence for how to use machine learning techniques to predict the attrition rates of female students based on socioeconomic and education-based information. Among all tested algorithms, Logistic Regression came out as the best performing but more importantly, easily interpretable algorithm. Furthermore, SVM showed the highest performance with respect to ROC-AUC as a classifying power of the models.

The current research justifies that parental education has been one of the top contributing factors towards encouraging female persistence, than any other barrier such as family income, school travel distance, etc. This again verifies previous studies based on the attainment of education that determined how literacy and awareness is inter-generationally transferred.

Moreover, this research shows how open datasets and transparent models can contribute to reducing some of the problems caused by gender inequity in education. The findings of this study can serve both data-driven policy making and targeted interventions; for instance:

- Adult education classes aimed at raising parental awareness
- School proximity initiatives, particularly in a rural and peri-urban setting
- Conditional economic incentives for families supporting girls' education.

Policy Implication

All stakeholders, like NGOs, governments, could use the predictive models built in this research to augment dropout prevention warning systems to enable the targeted education community to provide targeted support to at-risk girls in real time.

7. Limitations

A small dataset and its representation of the demographics might be a limitation for generalising the findings. The addition of datasets with varied regions (e.g., UNESCO MICS, DHS) and psychosocial elements (e.g., motivation, peer effects) would help build a more comprehensive model for predictions.

8. Future Work

Future Research needs to make use of deep learning, graph modeling, and longitudinal data to uncover dynamic patterns of dropout behavior across time. Further, incorporating geospatial and school facility variables might result in more precise predictions and better-informed decision-making.

In conclusion, through the intersection of data science and education that supports social equity, better policy solutions can be achieved through the development of interpretable machine learning models that make it possible for education planners to better improve female school retention.

References

[1] Rumberger, R. W. (2011). Dropping Out: Why Students Drop Out of High School and What Can Be Done About It. Harvard University Press https://doi.org/10.4159/harvard.9780674063167

[2] Abeysekera, D., & Wijeratne, K. (2023). Machine learning models for early prediction of student dropout in secondary schools. Applied Sciences, 13(4), 2592. https://doi.org/10.3390/app13042592

[3] Duque-Méndez, N. D., & Aristizábal-Quintero, L. Á. (2024). Advances in Computing: 18th Colombian Conference on Computing (CCC 2024). Springer. https://link.springer.com/book/10.1007/978-3-031-56712-2

[4] Hassan, S., Ahmad, T., & Farooq, M. (2024). Deep learning-based analysis for predicting student academic outcomes and dropout patterns. Education and Information Technologies, 29(3), 1157–1174. https://link.springer.com/article/10.1007/s10639-024-12740-3

[5] Kanishka, M. (2025). PM POSHAN: Ensuring social and economic justice for school-going children in India. Journal on the Rights of the Child of National Law University Odisha. https://nluo.ac.in/storage/2025/05/4.-PM-POSHAN-Ensuring-Social-Economic-Justice-for-School-Going-Children-in-India.pdf

[6] Mao, C., & Zhang, L. (2023). Predicting school dropout with explainable AI: Interpreting socio-economic and demographic factors. Heliyon, 9(8), e18842.https://doi.org/10.1016/j.heliyon.2023.e18842

[7] Msed, S. G., & Hraoui, S. (2025). A predictive approach based on machine learning for university student re-enrollment. IEEE 5th International Conference on Machine Learning and Applications. https://ieeexplore.ieee.org/document/11008348

[8] Rahman, M. M., & Chowdhury, S. (2023). Data-driven insights into educational inequality and female dropout in South Asia. Social Sciences & Humanities Open, 8(2), 101451. https://doi.org/10.1016/j.ssaho.2023.101451

[9] Sultana, R., & Ahmed, M. (2024). Socioeconomic determinants of female dropout in developing regions: A data-driven analysis. International Journal of Educational Development, 105, 103013. https://doi.org/10.1016/j.ijedudev.2024.103013

[10] Tariq, F., & Raza, H. (2023). Gender disparity and predictive analytics in secondary education: Evidence from South Asia. Computers & Education: Artificial Intelligence, 5, 100146. https://doi.org/10.1016/j.caeai.2023.100146

[11] He, S., Yousefpoori-Naeim, M., Cui, Y., & Cutumisu, M. (2025). Predicting College Enrollment for Low-Socioeconomic-Status Students Using Machine Learning Approaches. Big Data and Cognitive Computing, 9(4), 99. https://doi.org/10.3390/bdcc9040099.

[12] Education and Information Technologies, 28, 5523–5539.

[13] UNESCO (2022). Global Education Monitoring Report.https://unesdoc.unesco.org/ark:/48223/pf0000381329

[14] Cortez, P., & Silva, A. M. G. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito, & J. Teixeira (Eds.), Proceedings of 5th Annual Future Business Technology Conference, Porto, 5-12..