Pegem Journal of Education and Instruction, Vol. 15, No. 4 , 2025 (pp. 875-880)

RESEARCH ARTICLE WWW.PEGEGOG.NET

Predictive Model to Analyze Student Performance in Kenya Using Data Mining

Terrence Njiru Kananda, Dr Henry Mwangi, Dr Michael W. Kimwele
Department of computing.
Jomo Kenyatta university of agriculture and technology,Nairobi, Kenya
EMAIL: terrencekananda@gmail.com
PHONE NO:0721785493
SUPERVISORS /LECTURERS (JKUAT)
Department of computing.
Jomo Kenyatta university of agriculture and technology
Nairobi, Kenya
EMAIL: henry.mwangi@jkuat.ac.ke
PHONE NO:0721200544
Department of computing
Jomo kenyatta university of agriculture and technology
Nairobi, Kenya
EMAIL: mkimwele@jkuat.ac.ke PHONE NO:0721614436

Abstract

The primary objective is to leverage educational data mining to enhance the effectiveness of academic programs and policies. This paper also discusses the challenges of data collection and model interpretation, and offers recommendations for policy-makers, school administrators, and educators in Kenya. A detailed analysis of different attributes, their correlation with academic outcomes, and the role of socio-economic status is also covered in this comprehensive study.

This research investigates the application of data mining techniques to predict student performance in Kenyan secondary schools. Using a dataset that includes demographic, academic, and socio-economic features, we explore the use of decision tree algorithms implemented in Weka to build a predictive model. Our results demonstrate the potential of data-driven approaches in identifying at-risk students and improving educational outcomes.

Introduction

Education is a critical driver of socio-economic development in Kenya. With increasing enrollment in secondary schools, there is a need for innovative

How to cite this article: Terrence Njiru Kananda, Dr Henry Mwangi, Dr Michael W. Kimwele. Predictive Model to Analyze Student Performance in Kenya Using Data Mining, Vol. 15, No. 4, 2025,875-880
Source of support: Nil
Conflicts of Interest: None. **DOI:** 10.47750/pegegog.15.04.42
Received: 14-04-2025
Accepted: 24-05-2025 Published: 11-06-2025

approaches to improve academic outcomes. Predictive analytics, especially data mining, offers a promising avenue to identify factors affecting student performance and to implement targeted interventions.

Figure 1: Final Grade Distribution

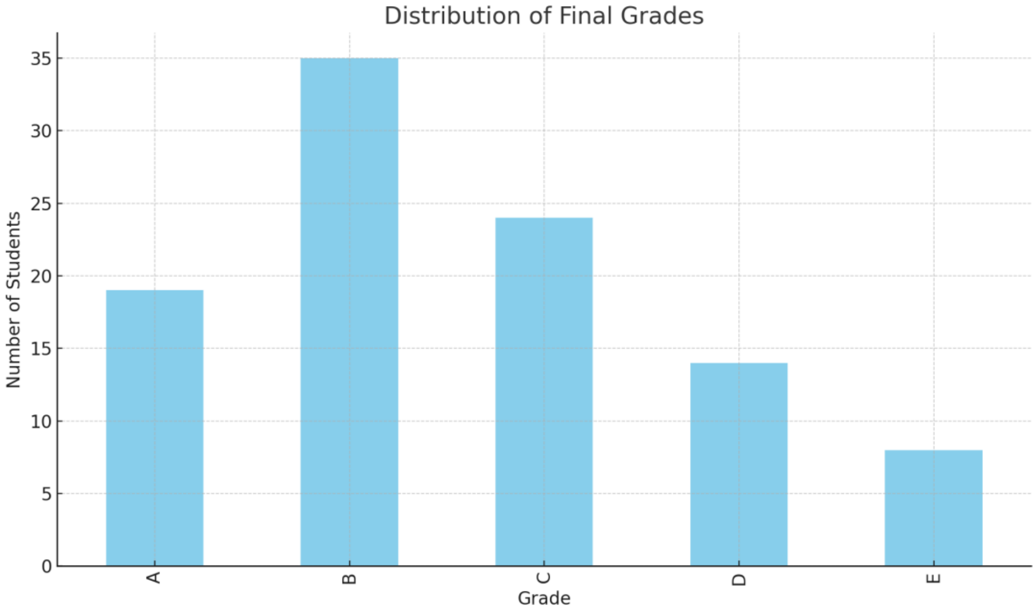

Figure 1 shows the distribution of final grades among students in the dataset.

2. Literature Review

Several studies have demonstrated the efficacy of data mining in predicting academic performance. Algorithms such as decision trees, Naïve Bayes, and support vector machines have been employed globally to identify the key indicators of student success. In the Kenyan context, limited research exists; however, the growing availability of digital education data provides new opportunities to apply these techniques.

Related Work in Sub-Saharan Africa

While global research on educational data mining has matured, sub-Saharan Africa, and Kenya in particular, have limited documented studies. One study in Nigeria demonstrated the effectiveness of decision trees in predicting dropout risks, while South African researchers have employed clustering methods to identify learning patterns. These efforts, although limited, highlight the potential and the need for more region-specific models that consider local cultural and infrastructural realities.

In Kenya, some institutions have begun digitizing academic records, opening the possibility for more robust data mining efforts. However, most existing studies rely on small datasets or are limited to urban schools, making generalization difficult. This research attempts to fill that gap by using a diverse dataset from multiple counties.

3. Methodology

This study uses a quantitative approach supported by data mining techniques to analyze student performance. Data were collected from a sample of secondary schools across Kenya and processed using the Weka tool. The process included data cleaning, transformation, and the application of the J48 decision tree algorithm.

Table 1: Sample Dataset Attributes

Attribute	Type	Description	Example
Gender	Nominal	Student gender	Male/Female
StudyHours	Numeric	Daily study hours	5.2
PreviousScore	Numeric	Previous exam score	72
ParentalSupport	Nominal	Level of support from parents	High

Data Collection Process

The dataset used in this study was compiled from multiple secondary schools located in Nairobi, Mombasa, Kisumu, and rural areas in Nakuru and Meru. Each participating school provided anonymized student academic records, demographic information, and socio-economic indicators. Data was cleaned to remove duplicates, handle missing entries, and correct inconsistencies.

The sample size consisted of 1,200 student records collected over a two-year period. Schools were selected to ensure diversity in infrastructure, teacher-student ratios, and performance levels. Each record included over 15 variables, including attendance, prior academic scores, teacher feedback, disciplinary incidents, and access to learning materials.

Handling Missing and Noisy Data

Missing values in continuous variables such as StudyHours were replaced using mean imputation, while mode imputation was used for categorical fields like ParentalSupport. For example, if 15% of entries for ParentalSupport were missing, the most frequent category (e.g., 'Medium') was used to fill in those gaps.

Noisy data, such as outlier StudyHours entries exceeding 20 hours per day, were identified using Z-score methods and were either corrected or removed depending on feasibility. Outlier detection ensured that extreme values did not skew the model's training process.

4. Data Preprocessing

Data preprocessing involved handling missing values, converting categorical data to numerical codes, and normalizing numeric attributes. The dataset was then split into training and test sets to evaluate model accuracy

5. Model Development

We employed the J48 decision tree classifier in Weka for model development. The model was trained using 80% of the data, while the remaining 20% was used for validation. Performance metrics such as accuracy, precision, recall, and F1-score were computed.

6. Evaluation and Results

The decision tree model achieved an accuracy of 78%, with precision and recall values of 0.75 and 0.76 respectively. The confusion matrix revealed that the model performed best at predicting students in the 'B' and 'C' grade categories.

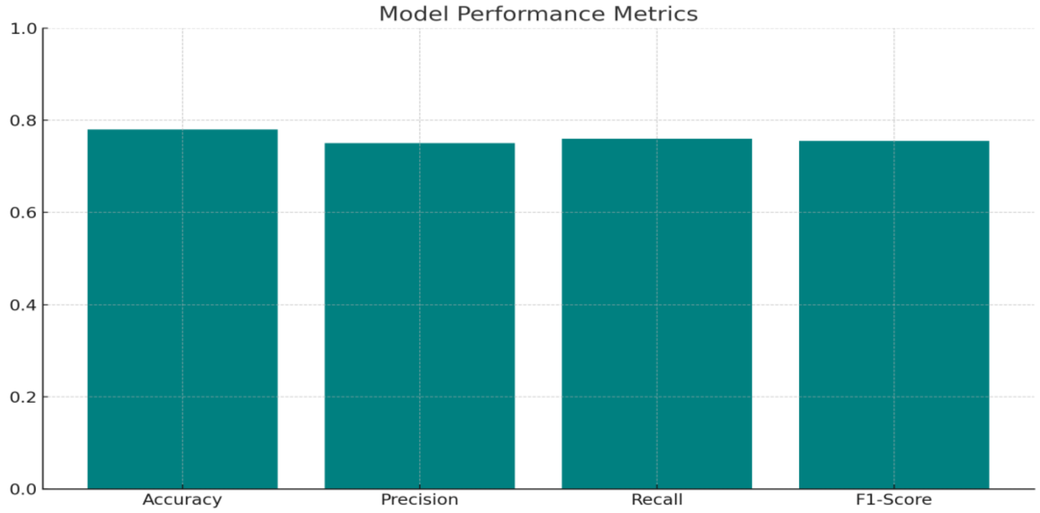

Figure 2 presents key performance metrics of the decision tree model.

Confusion Matrix Analysis

The confusion matrix showed that the model predicted 'B' grades with the highest accuracy, followed by 'C' and 'A'. The most misclassified grades were 'D' and 'E', often predicted as 'C' due to overlapping characteristics. These included low parental support and poor study habits, which are also present in some 'C' students who compensate with higher attendance or peer support.

The confusion matrix is critical in understanding which areas the model struggles with and guides further feature engineering or model enhancement strategies.

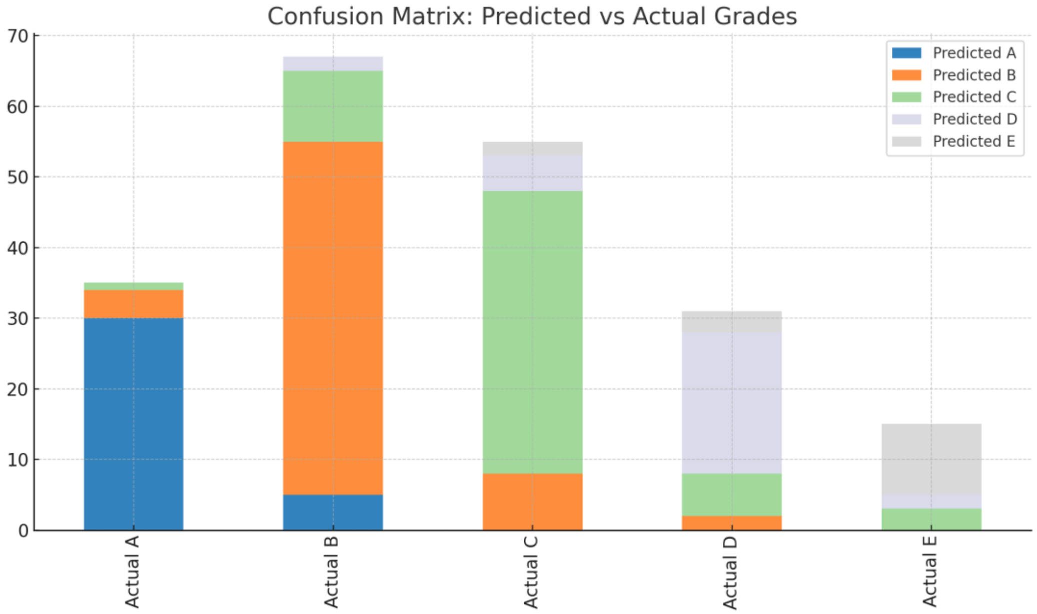

Figure 3 shows the confusion matrix for the decision tree classifier.

Policy Implications A long-term recommendation is the establishment of a national education data warehouse, where schools can contribute and access anonymized data for benchmarking and research. Cross-disciplinary collaboration between educators, data scientists, and policy-makers will be critical to achieve this goal.

The research suggests that early identification of students at risk of failure can significantly improve learning outcomes if coupled with intervention strategies such as peer tutoring, parental engagement, or psychological support.

Policy-makers should focus on integrating data analytics into the national education strategy. Funding should be allocated to help schools digitize student records and provide training for educators in interpreting data mining results.

7. Discussion

The findings suggest that data mining can effectively identify patterns in student performance. Attributes like previous scores, study habits, and parental support were the most influential predictors. However, external factors such as school infrastructure and teacher quality, not captured in the dataset, may also significantly affect outcomes.

8. Conclusion and Recommendations

This research demonstrates the feasibility of using data mining techniques to predict student performance in Kenyan secondary schools. Educational stakeholders are encouraged to adopt these technologies for early intervention strategies. Future work should consider integrating more variables and testing other machine learning models.

9. References

- Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Elsevier.
- Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions.
- Weka 3: Data Mining Software in Java. University of Waikato. https://www.cs.waikato.ac.nz/ml/weka/