ISSN No:-2456-2165

# Forecasting Student Academic Performance in Kenyan Secondary Schools Using Data Mining

Terrence Njiru Kananda[1]
[1] Department of Computing,
Jomo Kenyatta University of Agriculture and Technology,
Nairobi, Kenya

Henry Mwangi[2]
[2] Department of Computing,
Jomo Kenyatta University of Agriculture and Technology,
Nairobi, Kenya

**Abstract:- Stakeholders in Kenyan education are concerned about student performance. Data mining has emerged as an alternate method for education stakeholders to employ in making decisions about student performance in their final year exam. Kenya's education sector provides a wealth of statistical data that might provide vital information about students. Information and communication technology collects and compiles low-cost data that can be used to forecast student performance. However, no meaningful information is extracted from this data by Kenyan educational institutions. In this paper, we propose and develop a prediction model for forecasting Kenya secondary school learner performance utilizing prior performance data from students, which will be transformed and cleaned before being used in training and testing the model. Our model employs data mining techniques to improve forecast accuracy. We will present the model theoretical framework, conceptual framework, and outcomes.**

***Keywords***:- *DM - Data Mining, EDM – Educational Data Mining, KCPE – Kenya Certificate of Primary Education, KCSE – Kenya Certificate of Secondary Education.*

## I. INTRODUCTION

Quality education with a high transition rate from one educational stage to the next is one of the country's most important responsibilities to its citizens. Quality education does not indicate a great level of information created, but rather that education is offered to students in an efficient and effective manner in accordance with education pedagogy. Student information generates a tremendous volume of data in the education sector. EDM has a wealth of useful information and provides a more complete perspective of students and their educational experiences[1]. The use of information and communication technology in educational institutions can yield useful data that can be used to provide better education. Educational data mining, a new research community characterized in [2] as perdition, Clustering, Mining for relationships Data distillation for human judgment, Model development. Educational data mining makes use of technological advances in the academic sector. It provides a new viewpoint on education, which will lead to higher educational standards in our community.

Data mining is a critical technique that identifies proper data from massive amounts of data and selects acceptable data based on user requirements. Educational institutions are under intense pressure to offer up-to-date information on institutional effectiveness which include student conduct, the decision-making framework, and student engagement [3]. In every educational institution, being able to forecast a student's achievement in the main examination is a vital milestone in child development. Finding new solutions to apply systematic and data mining approaches to educationally related data is a critical solution to this challenge. Data mining techniques [3], which are used to find hidden patterns and relationships that may be useful in decision making, present a promising tool for analyzing these elements. Reference [4] introduces data mining techniques such as clustering, decision trees, and neural networks.

This research aims to use data mining to create a model that predicts academic achievement of students in public secondary schools based on a range of variables. By identifying the nature of this analysis, this research is predicted to give to the development of useful data that is meant to be correct for education decision makers. This will persuade policymakers to approach education in a way that allows them to pursue viable interventions for improved workforce enlargement and enactment.

## II. MOTIVATION AND BACKGROUND

Data mining, also known as information discovery in databases, is the method for mining knowledge from massive amounts of data using machine learning techniques that may be applied to a wide range of complex problems [5]. Companies must filter through all of that data, which necessitates the use of technologies. Data mining tools can help organizations discover vital information to improve decision-making and enhance efficiency in the education sector. Data mining is a collection of tools and techniques for uncovering hidden data trends and associations [6]. Educational Data Mining (EDM) is an area of research that focuses on data mining applications in the field of education [7]. It examines data that originate from school setting using tools and techniques from machine learning, statistics, data mining, and data analysis to know students better.

ISSN No:-2456-2165

There is a common assumption that if a scholar is in high school, he or she would achieve well on their final exam regardless of their prior level marks. This presupposes that students are guided through all academic coursework, curriculum completion, and pedagogical learning [8]. Besides from prior achievement, there are other elements that influence student performance, which might be external or internal [9]. External factors are influences outside the classroom that can impact student academic performance, such as financial situation and social and emotional problems, whereas internal factors are influences within the classroom that can impact student academic performance, such as student competence and aptitude.

This paper focuses on constructing and evaluating a data mining model suitable for predict student academic success. By identifying the nature of this analysis, this research is predicted to contribute to the development of useful information that is meant to be accurate to education decision makers. Since admission marks and final performance are used in the modeling process, there will be a shift in technique in calculating performance level determinants. Instructors can implement policies to guarantee that the entire project meets the required objectives, taking into account the grade level reached in the KCPE, location, age, and other factors. This will persuade policymakers to approach education in such a way that viable interventions for higher employee growth and achievement may be pursued.

## III. MODELLING FRAMEWORK

The research will be directed at recent high school graduates. In this regard, the main group partaking in the trial will be KCSE. Previous academic achievement (KCPE score), living location (rural or urban), student age, parents (both or single), and primary school type are independent factors for this paper as shown in table 1 .KCSE performance is a dependent variable.

Table 1 Subset of Data Utilized in Model Training and Testing

| S No | ADM | Primary | Parents | Location | Age Bracket | Stream | KCPE | KCSE |
|---|---|---|---|---|---|---|---|---|
| 1 | 43365 | Private | Both | Rural | 18 And Above | X | Average | Average |
| 2 | 44277 | Public | Both | Rural | 18 And Above | Y | Average | Good |
| 3 | 44182 | Public | Single | Rural | Below 18 | Y | Average | Good |
| 4 | 44284 | Public | Single | Rural | 18 And Below | X | Good | Good |

➢ *Theoretical framework of the study*

Figure 1 illustrates the theoretical framework of our model. Data is gathered from secondary schools, academic offices, and libraries, among other places. Following data collection, Microsoft Excel is utilized to store the data for subsequent processing stages. The collected data will be analyzed further, and any errors, duplicate data, and so on will be deleted. Data is translated in Data Transformation based on usage or tools for evaluating data Weka to.csv files. In Association Rule Mining, we identify strong rules in databases by employing several measures of support and confidence. The Apriori algorithm will be used in rule creation to construct the relevant association rules. In Evaluation, we will test our extracted knowledge and announce the percentage of efficiency of such information. Finally, by identifying any weaknesses or inefficiencies in the retrieved information, it can be utilized to forecast educational performance in a variety of applications.

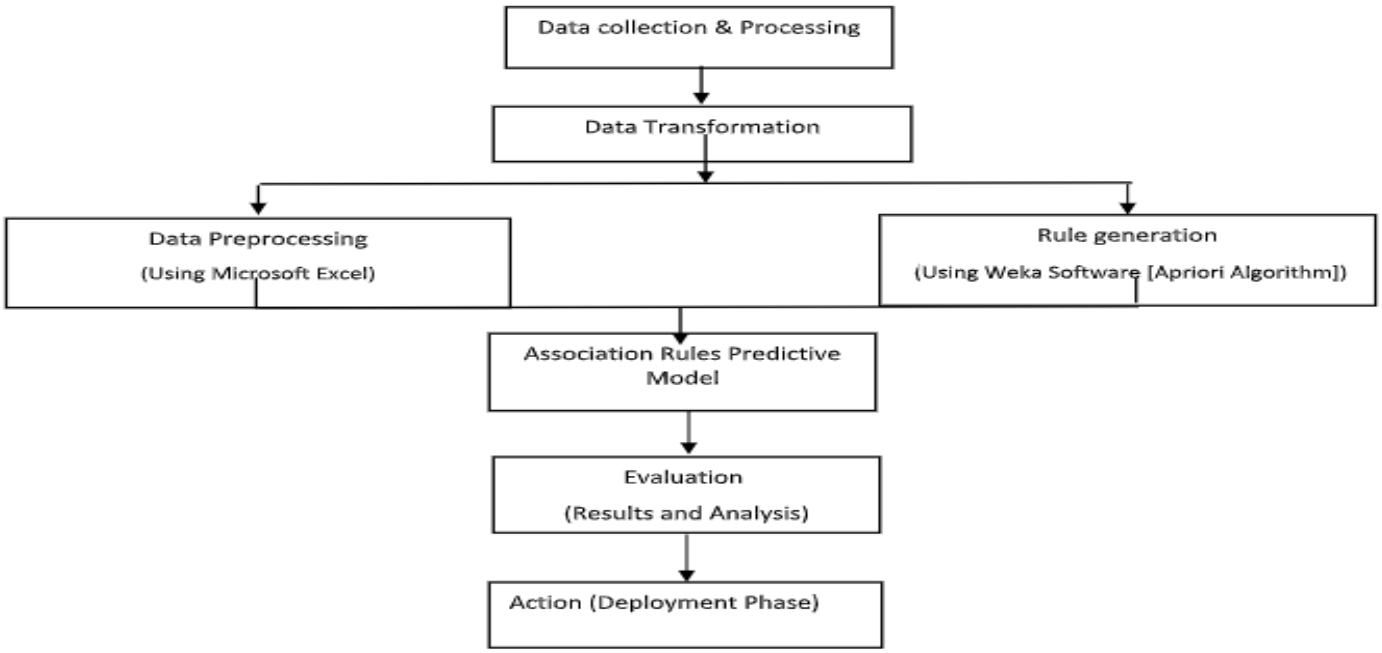


Fig 1 Theoretical Framework of the Model

ISSN No:-2456-2165

➢ *Conceptual Framework of the Model*

Figure 2 depicts the model's Conceptual Framework. Association rules are used to mine data in order to discover hidden relationships. This will aid in determining how each set of rules influences performance by examining support and confidence levels. The Apriori technique is used to mine all frequently occurring item sets in a database. Initially, the algorithm counts each item in the database to find the frequency of 1-itemsets with only one item. The frequency of 1-itemsets is used to find the itemsets in 2-itemsets, which is then used to find 3-itemsets, and so on until there are no more k-itemsets. If an itemset is not frequent, any significant subset of it is also not frequent; this condition prunes the search space of the database.

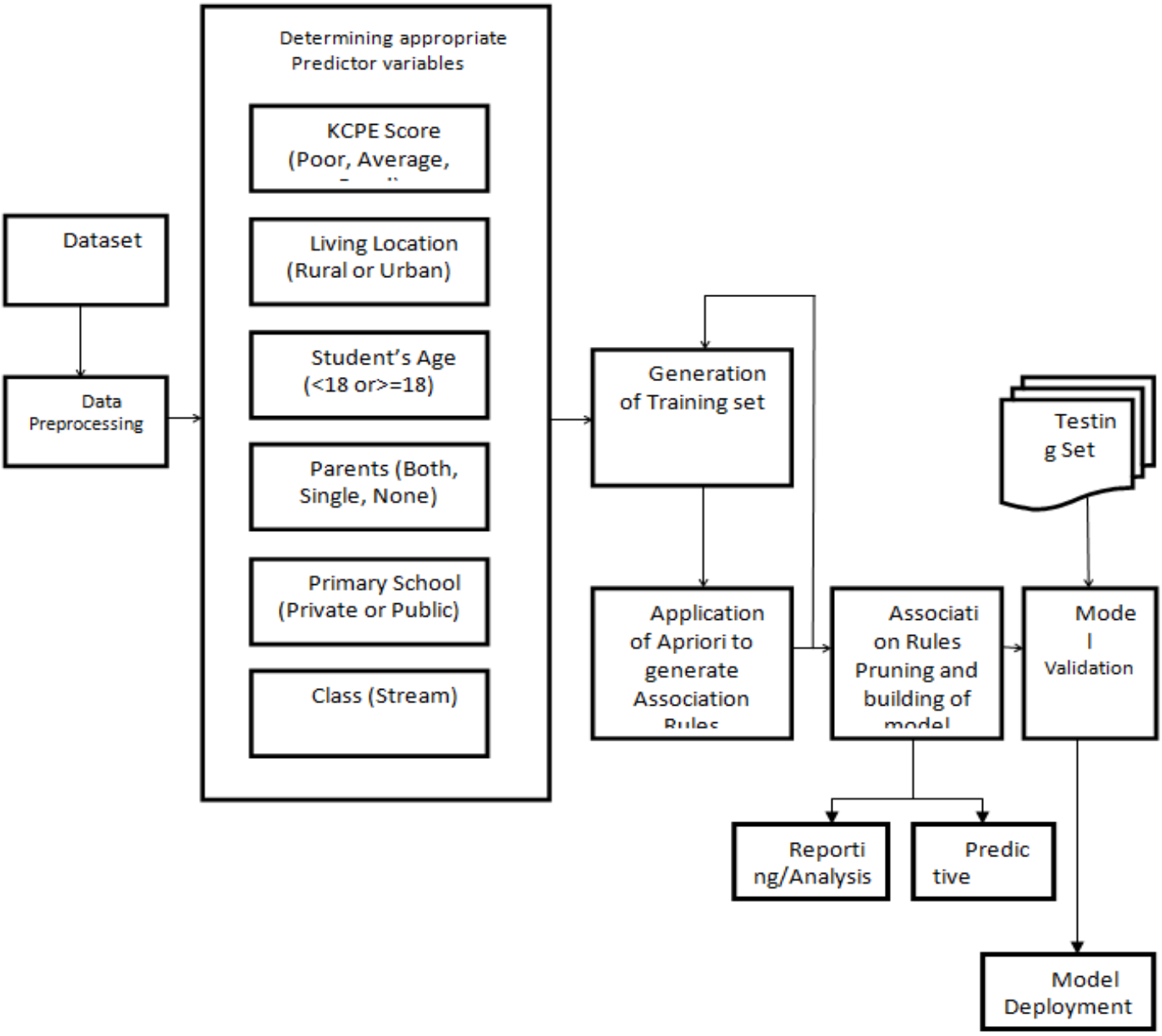


Fig 2 Conceptual Framework of the Mode

➢ *Model Output*

Figure 3 depicts a sample of the model's output. The first step in creating the Apriori model was to construct all frequent itemsets with the least amount of support. The following are the minimum levels of support and confidence: Minimum support: 0.2 (44 instances) and Minimum metric <confidence>: 0.8. The Apriori algorithm is then used to determine the most often occurring itemset. We may deduce from the model output that this study contributed to the development of accurate information for education decision makers on student performance predicting. This will persuade policymakers to approach education in a way that allows them to pursue viable interventions for improved workforce development and performance. We may deduce from the model output that this study contributed to the enlargement of accurate information for education decision makers on student performance predicting. This will persuade policymakers to approach education in a way that allows them to pursue viable interventions for improved workforce development and performance.

ISSN No:-2456-2165

➢ *Best Rules found:*

- PARENTS=BOTH LOCATION=RURAL AGE BRACKET=18 and Above KCPE PERFORMANCE=Good 51 ==> KCSE PERFORMANCE=Average 51 conf:(1)
- LOCATION=RURAL AGE BRACKET=18 and Above KCPE PERFORMANCE=Good 62 ==> KCSE PERFORMANCE=Average 60 conf:(0.97)
- PRIMARY=PUBLIC PARENTS=BOTH LOCATION=RURAL AGE BRACKET=18 and Above KCPE PERFORMANCE=Average 54 ==> KCSE PERFORMANCE=Good 49 conf:(0.91)
- PRIMARY=PUBLIC PARENTS=BOTH LOCATION=RURAL KCPE PERFORMANCE=Average 62 ==> KCSE PERFORMANCE=Good 56 conf:(0.9)
- PARENTS=BOTH AGE BRACKET=18 and Above KCPE PERFORMANCE=Good 65 ==> KCSE PERFORMANCE=Average 57 conf:(0.88)
- PRIMARY=PUBLIC PARENTS=BOTH KCPE PERFORMANCE=Average 64 ==> KCSE PERFORMANCE=Good 56 conf:(0.88)
- PRIMARY=PUBLIC PARENTS=BOTH AGE BRACKET=18 and Above KCPE PERFORMANCE=Average 56 ==> KCSE PERFORMANCE=Good 49 conf:(0.88)
- PARENTS=BOTH LOCATION=RURAL KCPE PERFORMANCE=Good 62 ==> KCSE PERFORMANCE=Average 54 conf:(0.87)
- PRIMARY=PRIVATE LOCATION=RURAL 52 ==> KCSE PERFORMANCE=Average 45conf:(0.87)
- LOCATION=RURAL KCPE PERFORMANCE=Good 77 ==> KCSE PERFORMANCE=Average 65conf:(0.84)

## IV. CONCLUSION

This paper describes a Data Mining Model to Predict Student Academic Achievement in Kenya. The acquired data is further evaluated and cleansed. Data is transformed to.csv files in Data Transformation based on usage or tools for evaluating data Weka. An Association Rule Mining identifies strong rules in databases by using a variety of support and confidence criteria. In rule creation, the Apriori algorithm is utilized to generate the relevant association rules. The model is then tested using extracted knowledge to determine the efficiency of such data. Ultimately, it can be used to forecast educational success in a range of applications by finding any gaps or inefficiencies in the collected information.

## ACKNOWLEDGMENT

I would like to thank my supervisors, Dr. Mwangi and Dr. Kimwele, for their inspirational guidance throughout this project.

## REFERENCES

[1] J. Mostow and J. Beck, “Some useful tactics to modify, map and mine data from intelligent tutors,” *Nat. Lang. Eng.*, vol. 12, pp. 195–208, Jun. 2006, doi: 10.1017/S1351324906004153.

[2] A. Algarni, “Data Mining in Education,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, Jun. 2016, doi: 10.14569/IJACSA.2016.070659.

[3] Prince Sattam Bin Abdulaziz University, M. Osman Hegazi, and M. A. Abugroon, “The State of the Art on Educational Data Mining in Higher Education,” *Int. J. Comput. Trends Technol.*, vol. 31, no. 1, pp. 46–56, Jan. 2016, doi: 10.14445/22312803/IJCTT-V31P109.

[4] A. S. OSMAN, “Data Mining Techniques: Review,” *Int. J. Data Sci. Res.*, vol. 2, no. 1, Art. no. 1, Jul. 2019.

[5] M. Gheisari *et al.*, “Data Mining Techniques for Web Mining: A Survey,” *Artif. Intell. Appl.*, vol. 1, no. 1, Art. no. 1, 2023, doi: 10.47852/bonviewAIA2202290.

[6] “(PDF) Data mining tools | Sithiphong Padungbuth - Academia.edu.” https://www.academia.edu /32619580/Data_ mining_tools (accessed Mar. 18, 2023).

[7] M. Goyal and R. Vohra, “Applications of Data Mining in Higher Education,” *Int. J. Comput. Sci. Issues*, vol. 9, Mar. 2012.

[8] J. Choi, J.-H. Lee, and B. Kim, “How does learner-centered education affect teacher self-efficacy? The case of project-based learning in Korea,” *Teach. Teach. Educ.*, vol. 85, pp. 45–57, Oct. 2019, doi: 10.1016/j.tate.2019.05.005.

[9] I. Jabor Al-Muslimawi and A. Hamid, “External and Internal Factors Affecting Student’s Academic Performance,” *Soc. Sci.*, vol. 14, pp. 155–168, Oct. 2019, doi: 10.36478/sscience.2019.155.168.