Examining Data Mining Classification Techniques for Predicting Early Childhood Development in Nigeria

Aimufua Ikponmwosa, Narasimha Rao Vajjhala, Sandip Rakshit, and Olumide Longe

Abstract Early childhood is a critical part of a child's development as it involves physical, cognitive, and psychological development. In the educational domain, especially early childhood education, there is rich data available that we could leverage to determine the development stage of a child and hidden patterns of a child's learning ability or disability. This study investigates which data mining classification technique will be most suitable in building a predictive model that can identify the social, cognitive, and emotional stages of a child. The authors compared J48, Naïve Bayes, random forest, support vector machines (SVM), and k-nearest neighbors (KNN) classifiers using performance measures like Kappa statistics, receiver operating characteristic (ROC), root-mean-square error (RMSE), and mean absolute error (MAE) using a data mining analytical tool called WEKA. The authors also compared the accuracy measures like true positive (TP) rate, false positive (FP) rate, precision, recall, and *F*-measure. The results indicate that the J48 classifier has a better classification accuracy and prediction rating over other tested algorithms using the early childhood dataset.

Keywords Data mining · Childhood development · Classification · Nigeria · Prediction · Accuracy · Model · Random forest · Classifier · Support vector machine

A. Ikponmwosa · S. Rakshit · O. Longe
American University of Nigeria, Yola, Adamawa, Nigeria
e-mail: ikponmwosa.aimufua@aun.edu.ng

S. Rakshit
e-mail: sandip.rakshit@aun.edu.ng

O. Longe
e-mail: olumide.longe@aun.edu.ng

N. R. Vajjhala (✉)
University of New York Tirana, Kodra e Diellit, Tirana, Albania
e-mail: narasimharao@unyt.edu.al

S. Borah et al. (eds.), *Advances in Data Science and Management*, Lecture Notes on Data Engineering and Communications Technologies 86,
https://doi.org/10.1007/978-981-16-5685-9_6

1 Introduction

Over 200 million children under the age of 5 in low and middle-income countries, particularly in Africa and Asia, do not attain full developmental potential because of various factors, including poverty, poor nutrition, and other factors [1]. Early childhood development is considered as a predictor of adult health and productivity [2]. Research indicates that investment in early childhood development can help countries in improving human development, human capital formation, economic growth, and social progress [1, 2]. Also, the exposure of a child to an environment that is not so conducive for upbringing during the first few years of their life lowers the child's intelligence quotient (IQ). This can also lead to low academic achievement, increased anti-social behavior, and reduces economic productivity in adulthood [2].

2 Review of Literature

2.1 Data Mining

The advent of smartphones over the last decade, coupled with other technological advances, has led to large volumes of data. Health, manufacturing, and other leading industries use large data repositories to help design business strategies and analyze unstructured and structured data to gain useful knowledge. Data mining involves analyzing large-scale observation datasets and identifying previously unknown relationships and summarizing the data in a novel manner. Data mining algorithms include both predictive and descriptive algorithms [3]. Data mining algorithms have been successfully applied in several domains, including biomedical research, and decision making at various management levels [4].

Data mining has several application types, including classification, estimation, prediction, correlation analysis, and visualization. According to Kumar and Khatri [4], data mining comprises the analysis of large data to discover trends and meaningful information that can be converted into a form of intelligence. Data mining is mostly used to interpret knowledge and discover hidden patterns from various domains of expertise. Ming-Syan et al. [5] stated that the difference between data mining and traditional data analysis is the ability to mine information and discover knowledge and the premise of no clear assumption. Data mining uses automated data analysis techniques to uncover previously undetected relationships among data items [6].

2.2 Applications of Data Mining

There are several domains in which data mining can be applied, a domain like social media analytics, medical data mining, educational data mining, business intelligence, etc. In the medical and healthcare domain, there are issues. Healthcare domain is a large domain, especially as the healthcare information systems generate a huge amount of data regarding the patient's medical history, current ailment, and diagnoses. With this huge data being generated, the medical domain can watch out for repeated patterns and predict the outcome of such a result, leading to improved quality relating to the overall quality of patient services, early prediction and diagnosis of diseases [7]. Data mining techniques can be employed again in the medical field for clinical test analysis and its relationship pathology [8].

Komi et al. [9] conducted a study on applying data mining methods in diabetes prediction. The authors emphasized using data mining methods in shedding more light on the prediction of diabetes mellitus. In their report, the authors comparatively analyzed five different mining techniques using MATLAB tools, including Gaussian mixture modeling (GMM), extreme learning machines (ELM), support vector machines (SVM), logistic regression, and artificial neural networks (ANN), to propose a technique that will be most effective in the early prediction of diabetes. This technique model was trained and validated against a test dataset. The experiment's effect proved that provided with the diabetes dataset ANN technique provided the highest accuracy.

Kumar and Khatri [4] compared different classification techniques and their prediction accuracy for a specific dataset (chronic kidney disease). Kumar and Khatri [4] used WEKA as a data mining tool to analyze five classification techniques, namely J48, Naïve Bayes, random forest, SVM, and KNN, using performance measures like TP rate, FP rate, and precision. Olukunle and Ehikioya [10] proposed using a fast association rule mining (ARM) algorithm to be used on medical image dataset. They established that ARM aims at discovering strong, interesting patterns between items in a vast dataset. Olukunle and Ehikioya [10] further proposed the use of the frequency pattern (FP) growth algorithm, a standard ARM algorithm that is efficient for mining a large dataset from frequency pattern. Olukunle and Ehikioya [10] suggested that his approach will have a compactable parallel data representation scheme for input and output structure. Further, Olukunle and Ehikioya [10] experimented with showing that FP growth has the desirable features in handling large medical data images. In their conclusion, Olukunle and Ehikioya [10] indicated that it is necessary to mine medical images because of the vast information available for knowledge support and to apply the ARM technique because it is simple and explanative. As the education data mining gets larger into early childhood education by so doing having different characteristics, different data mining techniques will have their predictive efficiency. Nigeria is the most populous African country. However, there is limited literature predicting early childhood development challenges using various data mining classification techniques, even though there is significant data available in this domain. In this study, the authors comprehensively studied and

compared other data classification techniques and their prediction accuracy for early childhood datasets.

Shouman et al. [11] used decision tree algorithm techniques in diagnosing heart disease. The authors investigate applying a range of techniques to a different type of decision tree seeking better performance in heart disease diagnosis. Shouman et al. [11] proposed a model that enhances the decision tree accuracy by integrating a multiple classifier voting technique with different types of discretization methods and different decision tree types. The research seeks to improve diagnosis accuracy by applying the multi-interval discretization method, multiple classifier voting, and reduced error pruning to the decision tree. The experiment was conducted using Microsoft Visual Studio 2019, and it involved systematically testing different discretization techniques, multiple classifiers, voting techniques, and various decision tree types in the diagnosis. Shouman et al. [11] concluded by saying that by applying multi-interval equal frequency with nine voting gain ratios, and the accuracy of the decision tree will be improved by a percentage of 84.1%.

Kumar and Pal [12] investigated engineering students' performance improvement using some data mining techniques. A 17-attribute dataset was created to be run by WEKA open-source analytical tool to enable the user to apply classification and regression on the resulting dataset over a tenfold cross-validation, thereby estimating the predictive model's accuracy. The authors used the Iterative Dichotomizer 3 (ID3), C4.5, and Classification and Regression Tree (CART) algorithms in the classification model to test the predictive model. In their study, Shouman et al. [11] found that the C4.5 technique had the highest predictive accuracy in identifying students who were more likely to fail than other methods.

Comendador et al. [13] examined the students' history of accessing a university learning management system (LMS) data, applying some data mining techniques to build a model that will predict the user's learning behavior. The study's objective was to categorize the typical online behavior of distance education and identify influencers for learning outcomes. In their study, the authors used a dataset from the University history to access the Polytechnic University of the Philippines (PUP) and choose a dataset of 248 student records. Comendador et al. [13] conducted an experiment using WEKA and understudied reduced error punning tree (REPTree), CART, and J48 tree algorithm to evaluate the appropriate classification algorithm that can be utilized for predicting student finals based on usage data in the LMS. Comendador et al. [13] further applied discretization and tested the algorithms on the provided dataset using tenfold cross-validation in WEKA. After obtaining the final attribute, Comendador et al. [13] further applied a three-feature selection technique CH, GR, and IG in the classification of student performance in an E-Learning environment. The final result showed that the score obtained from participation in the online activity was the most valuable influencer to completing the program [13].

de Paula Santos et al. [14] proposed an evaluation model using educational data mining techniques to analyze students' responses during an institutional teaching evaluation. The mining data process begins to identify the categories of analysis that students find most important in the teaching practices and identify the semantics orientation of student response to the instructor, whether positive or negative. de

Paula Santos et al. [14] further categorized the model into five stages and compared it to existing Higher Technological Education, in which the statistical models and data mining were not used. They aim to promote EDM use in particular sentiment analysis, to identify which teaching practice is good and not considered from the student perspective. Their research's relevance is student-centered, making students cease from being mere spectators and becoming the protagonist in reconstructing the teaching model and roles within the institution.

3 Methodology

In trying to understudy the educational data mining (EDM) of early childhood, we came up with this approach on:

- Which classification technique (J48, random forest, Naïve Bayes, SVM, and KNN) will be most appropriate in developing a predictive model for early childhood education system in Nigeria?
- What are the relevant factors that influence early childhood development in Nigeria?

Classification can be classified as a learning technique that organizes items in collection to target categories which aims to accurately predict the large dataset into classes. Kesavaraj and Sukumaran [15] indicated in their research that classification can be used to predict group membership of same instance and can be used to classify item into a set of classes. According to Umadevi and Marseline [16], classification techniques are classified into two groups, the supervised learning and the unsupervised learning. In the supervised learning, the classified data are grouped into classes based on insight of different classes, while the unsupervised classification data are not predicted by the user.

Clustering technique in data mining is the process of grouping a collection of same datasets into classes of similarity in order that data of same object type are grouped differently from object of another cluster. The number of similarities between the object of a cluster is calculated by the use of similarity function. Clustering is very useful for document organization, it helps in data recovery technology, and it also increases the efficiency of a database system [17]. In analyzing the clustering technique, the first step involves computing "proximity indices" between particular groups in relating to interest area. Having known the proximity indices, a clustering algorithm can then be applied to group with similarity in object data. There are factors for selecting that leads to choosing an appropriate clustering method which includes the nature of the data (continuous or nominal) and the size of the data matrix [18]. In this phase, we obtained early childhood learning data from Word of Faith College (nursery section) in Benin City, Nigeria and created an online questionnaire (Google form) which the links to the questionnaire was sent to Church Day care

centers, social media platforms, students, staff, and members of American University of Nigeria with the help of the Student Government Association (SGA). A survey instrument was used to collect the data containing eighteen (18) questions.

4 Analysis and Findings

Descriptive statistics involve summarizing the figures of data collected. Descriptive analysis includes numbers in average and/or percentage, graph, and tables [19]. For this research thesis, the target population are individuals who have children or give care to children within the early childhood development stage.

Table 1 shows that majority of the respondent were mothers and they made up of 42.2% of the whole 279 population of respondent but 9 responses in general was incomplete, while Father's has a total of 6.6%. From our WEKA analysis, we overlaid the environment in which the child is growing up with the type of responses given, and we could deduce a factor that matter to the overall development of early childhood. From Table 2, we can notice that children in the averagely conducive environment and that of the very conducive environment have higher and better number of positive responses than that of the children living in the not so convenient environment.

In Table 3, five major classifiers, namely J48, random forest, Naïve Bayes, SVM, and KNN were considered for use in this thesis research using WEKA. Various performance measures relating to the aims and objective of the thesis have been measured in Table 3.

Figure 1 shows the level of accuracy for all classifiers. It shows that J48 and random forest have better reading of accuracy, while K-nearest neighbor reveals the lowest. We can deduce that J48 has slightly better accuracy than random forest algorithm.

Table 1 Descriptive analysis of collected data

Category	Percentage %
Mothers	42.2
Class teacher	39.6
School administrator	1.9
Fathers	6.6
Care giver	3
Other (brothers, sisters, relatives)	6.7

Table 2 Environmental condition of children

Count	Environment of child
78	Very conducive (color code red)
154	Averagely conducive (color code blue)
38	Not so conducive (turquoise blue)

Table 3 Experimental result of all algorithms

Algorithm	Accuracy	ROC	Kappa statistics	RMSE	Mean absolute error	Time to build model (s)
J48	71.85	0.727	0.476	0.3994	0.247	0.06
Random forest	70.37	0.811	0.451	0.3668	0.274	0.19
Naïve Bayes	61.11	0.767	0.377	0.4509	0.266	0.01
SVM	68.14	0.733	0.425	0.3892	0.301	0.24
KNN	57.77	0.619	0.226	0.5072	0.286	0

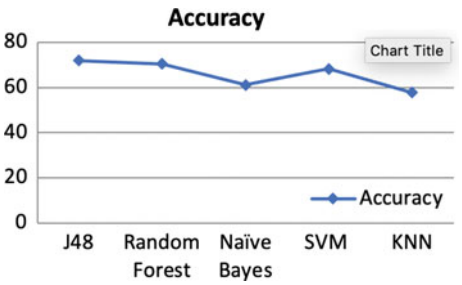

Fig. 1 Classification accuracy value of algorithm

In Fig. 2, ROC curves reveal good result for random forest and Naïve Bayes, a little fair outcome for SVM and J48, and a decreased performance for KNN. However, the time it took to build models is less in the case of KNN, SVM, and J48 compared over random forest and Naïve Bayes.

The *F*-measure represents the combining of precision and recall. We can then put together that a classifier that has high precision and low recall is adequate and most accurate as shown in Fig. 3.

In Table 4, J48 revealed a significantly high figure in the true positive (TP) rate, the precision, recall, and the *F*-measure which indicates a good performance of the algorithm and a low figure in the false positive (FP) rate which indicates that the algorithm can handle amount of false positive attribute/numbers and reduce the outcome of false result.

5 Conclusion

This study uses data mining classification techniques to predict and provide a better understanding of early childhood to develop to understand the teaching methods to meet the children's needs. A predictive model was developed that could be used to collect, process, and review hidden information. This information can help in predicting the challenges that children face as part of early childhood development. Previous studies in the education domain had mainly focused on higher education, so this predictive model should help educators and policymakers better understand the

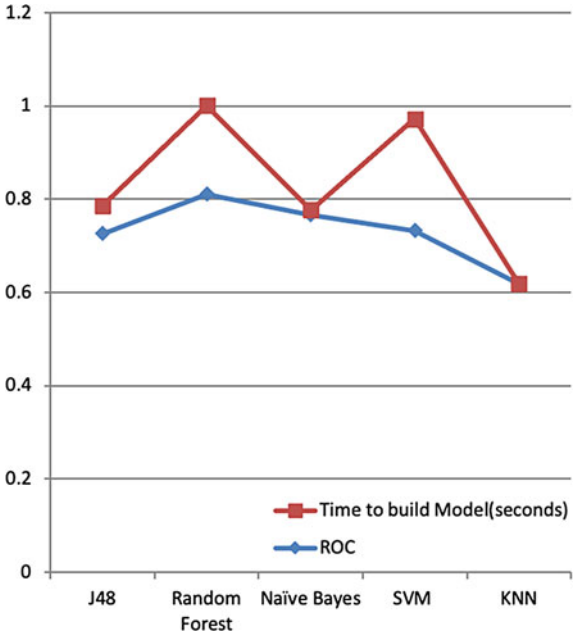

Fig. 2 ROC versus time to build

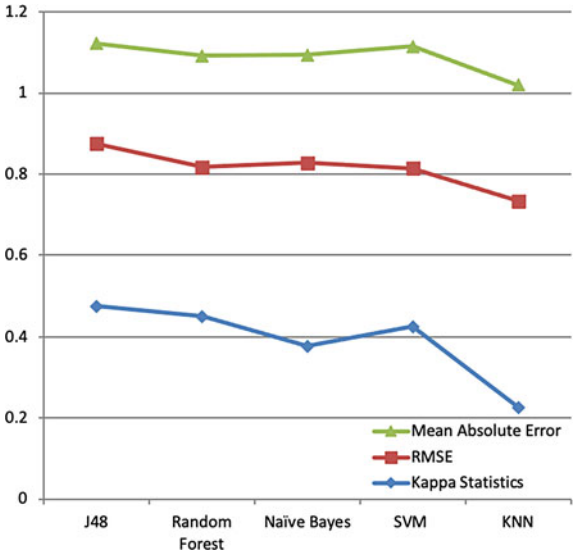

Fig. 3 Result of *K*-value versus error

Table 4 Major accuracy measure value

Algorithm	TP rate	FP rate	Precision	Recall	*F*-measure
J48	0.719	0.280	0.719	0.719	0.711
Random forest	0.704	0.292	0.702	0.704	0.694
Naïve Bayes	0.611	0.235	0.644	0.611	0.613
SVM	0.681	0.289	0.678	0.681	0.677
KNN	0.578	0.390	0.566	0.578	0.567

challenges encountered during early childhood development. Several classification techniques along with the key performance indicators were analyzed in this study in the context of early childhood development.

References

1. Baker-Henningham H (2014) The role of early childhood education programmes in the promotion of child and adolescent mental health in low- and middle-income countries. Int J Epidemiol 43(2):407–433
2. Masterov D (2007) The productivity argument for investing in young children. Rev Agric Econ 29:446–493
3. Wright MOD, Masten AS (2005) Resilience processes in development. In: Goldstein S, Brooks RB (eds) Handbook of resilience in children. Springer US, Boston, MA, pp 17–37
4. Kumar N, Khatri S (2017) Implementing WEKA for medical data classification and early disease prediction. In: 2017 3rd international conference on computational intelligence & communication technology (CICT). IEEE, Ghaziabad
5. Ming-Syan C, Jiawei H, Yu PS (1996) Data mining: an overview from a database perspective. IEEE Trans Knowl Data Eng 8(6):866–883
6. Sahu H, Shrma S, Gondhalakar S (2011) A brief overview on data mining survey. Int J Comput Technol Electron Eng (IJCTEE) 1(3):189–207
7. Aigbovo O (2019) Trend and pattern of economic and financial crimes statutes in Nigeria. J Financ Crime 26(4):969–977
8. Podgorelec V, Hericko M, Rozman I (2005) Improving mining of medical data by outliers prediction. In: 18th IEEE symposium on computer-based medical systems (CBMS'05). IEEE, Dublin
9. Komi M et al (2017) Application of data mining methods in diabetes prediction. In: 2017 2nd international conference on image, vision and computing (ICIVC). IEEE, Chengdu
10. Olukunle A, Ehikioya S (2002) A fast algorithm for mining association rules in medical image data. In: IEEE CCECE2002. Canadian conference on electrical and computer engineering. Conference proceedings (Cat. No. 02CH37373). IEEE, Winnipeg, Manitoba
11. Shouman M, Turner T, Stocker R (2011) Using decision tree for diagnosing heart disease patients, vol 121, pp 23–30
12. Kumar S, Pal S (2012) Data mining: a prediction for performance improvement of engineering students using classification. World Comput Sci Inf Technol J 2:51–56
13. Comendador BEV, Rabago LW, Tanguilig BT (2016) An educational model based on knowledge discovery in databases (KDD) to predict learner's behavior using classification techniques. In: 2016 IEEE international conference on signal processing, communications and computing (ICSPCC). IEEE, Hong Kong

14. de Paula Santos F, Lechugo CP, Silveira-Mackenzie IF (2016) "Speak well" or "complain" about your teacher: a contribution of education data mining in the evaluation of teaching practices. In: 2016 international symposium on computers in education (SIIE). IEEE, Salamanca
15. Kesavaraj G, Sukumaran S (2013) A study on classification techniques in data mining. In: 2013 fourth international conference on computing, communications and networking technologies (ICCCNT). Tiruchengode, IEEE, pp 1–7
16. Umadevi S, Marseline KSJ (2017) A survey on data mining classification algorithms. In: 2017 international conference on signal processing and communication (ICSPC). IEEE, Coimbatore
17. Patel D, Modi R, Sarvakar K (2014) A comparative study of clustering data mining: techniques and research challenges, vol iii, pp 67–70
18. Antonenko PD, Toy S, Niederhauser DS (2012) Using cluster analysis for data mining in educational technology research. Educ Technol Res Dev 60(3):383–398
19. Agresti A (2018) Statistical methods for the social sciences, 5th edn. Pearson Inc., Boston, MA