Scientific Research Publishing

Journal of Software Engineering and Applications, 2020, 13, 245-257
https://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

Enabling Proactive Management of School Dropouts Using Neural Network

Khamisi Kalegele

ICT Department, Open University of Tanzania, Dar es Salaam, Tanzania
Email: kalegs03@gmail.com

How to cite this paper: Kalegele, K. (2020) Enabling Proactive Management of School Dropouts Using Neural Network. *Journal of Software Engineering and Applications*, **13**, 245-257.
https://doi.org/10.4236/jsea.2020.1310016

Received: September 2, 2020
Accepted: October 16, 2020
Published: October 19, 2020

Abstract

The growing need to use Artificial Intelligence (AI) technologies in addressing challenges in education sectors of developing countries is undermined by low awareness, limited skill and poor data quality. One particular persisting challenge, which can be addressed by AI, is school dropouts whereby hundreds of thousands of children drop annually in Africa. This article presents a data-driven approach to proactively predict likelihood of dropping from schools and enable effective management of dropouts. The approach is guided by a carefully crafted conceptual framework and new concepts of average absenteeism, current cumulative absenteeism and dropout risk appetite. In this study, a typical scenario of missing quality data is considered and for which synthetic data is generated to enable development of a functioning prediction model using neural network. The results show that, using the proposed approach, the levels of risk of dropping out of schools can be practically determined using data that is largely available in schools. Potentially, the study will inspire further research, encourage deployment of the technologies in real life, and inform processes of formulating or improving policies.

Keywords

Dropout, Machine Learning, School Management

1. Introduction

Globally, there are ever-growing excitements about revolutions and benefits that the Artificial Intelligence (AI) technologies can bring in development sectors [1] [2]. In developing countries, key players for the harnessing such benefits face mountains of limitations. One such limitation facing researchers and experts is inadequate domain understanding that would inspire assessment and performance optimization of the applicable and appropriate AI technologies. Another limitation is fitness for use (e.g. accuracy, completeness, consistency and integrity) of the existing data that would attract application of sophisticated AI technologies [3]. Due to such limitations, it is quite daunting for experts, developers and researchers in developing world to expeditiously promote the use of AI technologies and associated research to meet the growing needs. Using the problem of school dropouts in a typical scenario of developing countries, this article presents a machine learning approach to predict dropouts in schools for the purpose of increasing awareness, catalysing further research and providing guide to developers. In real life, the approach can be adopted by schools and developers in predicting the likelihood of students dropping out of schools. In the remainder of this section, dropout determinants from literature, situational analysis of dropout problem in developing countries, and the research problem are explained.

1.1. Dropout Determinants

Oxford dictionary defines dropout as a person who has abandoned a course of study or who has rejected conventional society to pursue an alternative lifestyle. In many countries, such definition holds but additionally, authorities extend it to specify the number of days when a student is absent before being officially declared a dropout. In Tanzania, for example, it used to be 90 days for a long time but that has recently come into scrutiny by some politicians. Using new concepts of absenteeism, in this article, we provide a new definition of dropout in the Methods Section. Studies depict that prior to dropping from school, students face lots of challenges and a number of undesirable but indicative life events happen. Most of the issues surrounding dropping students are quantifiable and form the basis of manageable determinants. Literature has established a number of such determinants [4] [5] [6] [7]. The relationships of such determinants to the likelihood of dropping from school have also been studied extensively [8] [9]. Moreover, a rich body of statistical techniques that can productively analyse the determinants exists.

Thoroughly, Kalinga *et al.* studied dropout determinants and summarized some of their relationships [8]. The status of one's family such as income, education and employment is closely linked to prospects of dropping out of school. Students from low income families are deemed likely to drop out of schools. The situation can be worse for some students under single parenting or orphans. Gender also matters a lot because girls suffer from menstruation related challenges and pregnancies while boys do not. Age is also sometimes considered because the older students face more problems both in school and in society including social responsibilities. Students living in urban areas are relatively highly exposed to undesirable affairs and therefore are far more tempted than their counterparts in semi-urban or rural. The issue of urban locality is subjective because in some other studies students in rural areas were found engaged in farm related activities which can lead to dropping out of schools. Student's sense of belonging is also linked to dropout, for instance, those who chose school because of its excellence are less likely to drop than those who chose because the school is in the neighbourhood. There are also academic and disciplinary factors whereby well-performing students are less likely to drop. This includes performance of student's peers as it has been observed that peer influence can be quite strong and decisive. Loop holes in school management (such as policies, and practices) are also a direct negative influence to many dropouts. Generally, dropout determinants are diverse and their relationships to local norms are sometimes sensitive. For instance, although it is common knowledge that some tribes care more about education than others, including tribe as a dropout determinant can be tricky and unwelcome.

1.2. Challenges Facing Dropout Management in Developing Countries

In many developing countries, there is now increasing level of awareness for data driven approaches to developmental challenges [10]. Largely the awareness is being driven by donor funded projects. Although there is significant progress in adopting the approaches, the overall buy-in is still quite low in many areas. Over decades, despite huge influx of funds to support ICT infrastructure projects, management of school attendance is still non-digital in many countries. The overall lack of digital platforms to manage school data is detrimental to efforts of using new technologies to ease management burden, improve efficiency and rescue students from dropping.

Policies require schools and authorities to collect and aggregate predefined school and student data. In Tanzania, for instance, schools maintain attendance books that are issued by the authorities and regularly provide reports. These reports are aggregated at national level by the ministry responsible for education [11]. Typical problems surrounding data management at national level are poor quality of data (as explained in Section 1) and inability to disaggregate data during analysis. Sometimes, low levels of data literacy lead to undesirable actions such as data cooking.

The whole end to end chain of managing school data, from school to national level, is characterized by low levels of data experts. At best, only descriptive analysis is conducted for reasons other than addressing the problem in question. College and University graduates have also been criticised for failure to perform as expected once employed.

Although Africa is still experiencing dropout problems, it is surprising that very little evidence from quality research is reported beside authorities' statistical figures (which are questionable in some cases). One can argue that, the shortage of evidence is primarily caused by lack of longitudinal studies to the problem. However, it was expected that reliable and comprehensive data, that presumably have been used for planning and projection purposes for years, be available. This article presents a dropout predictions scenario and uses it to present a novel approach to predict risk of dropouts. The approach involves a practical selection of determinants and use of neural network to adaptively make predictions. Despite the above highlighted challenges, schools and authorities will find it easy to deploy the proposed prediction approach in a production environment using a minimum of resources.

1.3. Research Problem

Sections 1.1 and 1.2 analyzed dropout situation in both a generic and developing country's contexts whereby despite the treats of dropouts, interventions are still reactive due to data use limitations. Although many data-driven approaches that utilize technologies exist, they have not been sufficiently tested in the context of developing countries. Prominent technologies, that are relevant to the dropout problem, commonly use machine learning algorithms. To most practitioners, it is not clear how such algorithms can be used, in the context presented above, to practically predict dropouts in ways that enable interventions. Thus, this article presents a complete approach, using neural network algorithm, for predicting dropouts with minimum of available data. The approach presents engineered dropout determinants using newly defined absenteeism concepts, and a guide on how to practically determine levels of risk of students to drop out of schools.

2. Methods

2.1. Conceptual Framework

This article is not about formulating new theories related to dropout or machine learning but rather testing existing ones and simulating findings. The various propositions described in the previous sections and machine learning theories are forged into a framework that guided this study, shown in Figure 1. Literature suggests that dropout determinants that can be categorized into six groups forming inputs to the framework as shown in the figure. Using statistical approaches, the inputs can be aggregated into single normalized values ready for machine learning based computation of dropout likelihood. Once dropout likelihood is known for a particular scenario, the level of risk can be determined and intervention decisions appropriately made by the school or parent.

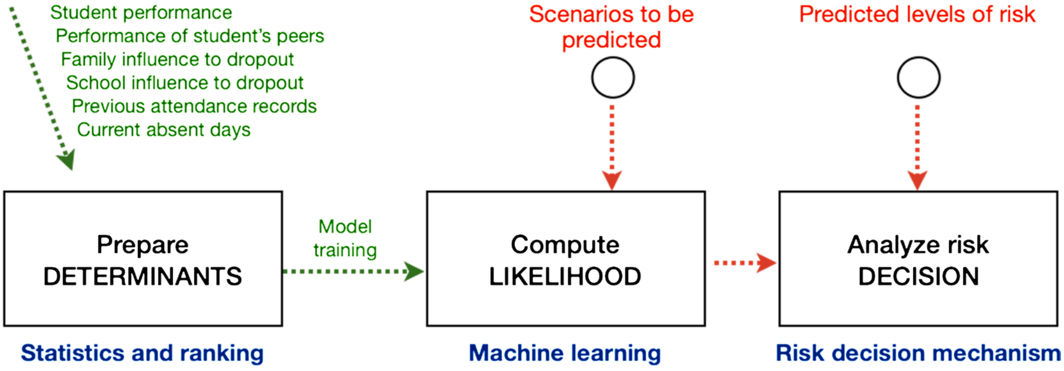

Figure 1. Conceptual Framework for the prediction of dropouts.

2.2. Selection of Determinants

In developing countries, the reality of school data management environment suggests that systems can hardly afford additional overhead cost to collect data. The choice of determinants in this study was primarily driven by the desire to use data that can easily be obtained by schools. Attendance, which underpins absenteeism, and academic performance data are selected because schools are already collecting them and there is enough body of evidence relating them to dropout. Academic performance data constitutes information about academics scores and disciplinary records of both a student and his/her peers. As presented in Section 1, studies have also suggested family and school related factors. The next Sub-Section presents frequency distributions of all determinants that were used.

2.3. Generation of Data

Statistics have shown that academic performance, family influence to dropouts, and school influence are normally distributed around certain average values. Assuming normalized values in percentage, synthetic data was generated using R statistical tools. The use of synthetic data has continued to be popular among computer scientists and researchers [12] [13] [14]. This resulted from huge number of failing projects due to missing quality data and also time cost associated with acquisition of data [15] [16]. For a dropout problem, synthetic data is only used to train the initial prediction model with prospects of optimizing it as actual data instances are seen. In this study, the generation of data considered two absenteeism related attributes which constitute new concepts: average absenteeism and current cumulative absenteeism.

Definition 2.1: Average absenteeism is the average number of all previous absent days accrued per semester or term *i.e.* sum of total number of semester (or term) absent days divide by number of semesters or terms.

Definition 2.2: Current cumulative absenteeism is the total number of ongoing continuous absent days. This is zero for all present students in any particular day.

Definition 2.3: A dropout (based on current cumulative absenteeism) is a student whose current cumulative absenteeism is equal or larger than the threshold set by the relevant authorities.

A typical class of Sub-Saharan Africa of 201 students was assumed and the authority's dropout threshold was set at 9 days for convenience (any arbitrary number can be used). For each student, a random pattern of current cumulative absenteeism over attendance days was generated as shown in **Figure 2**. Randomly, 19 students (9.5%) were deliberately set to cross the dropout threshold as shown in **Figure 3**. **Figure 4** shows distributions of all the selected determinants for both dropping and non-dropping students.

2.4. Preparation of Datasets

The generated data, in Section 2.3, consisted of records for 201 students or

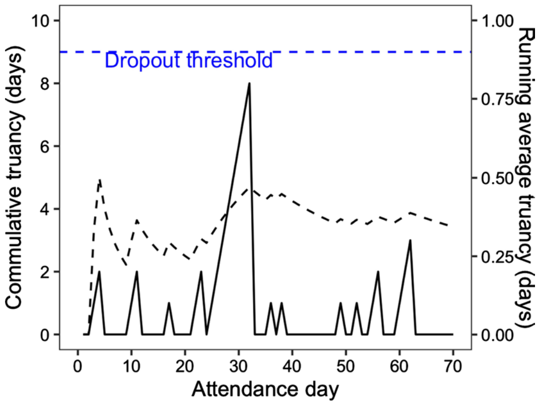

Figure 2. Absenteeism—student A.

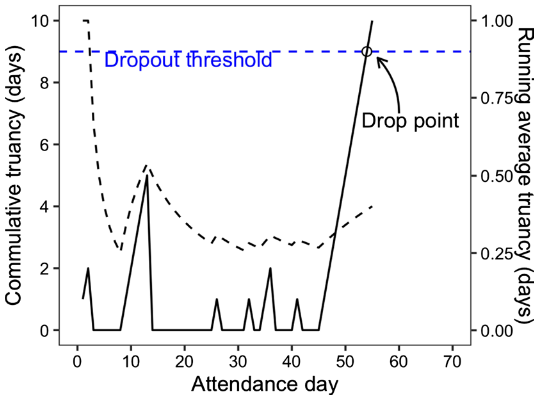

Figure 3. Absenteeism—student B.

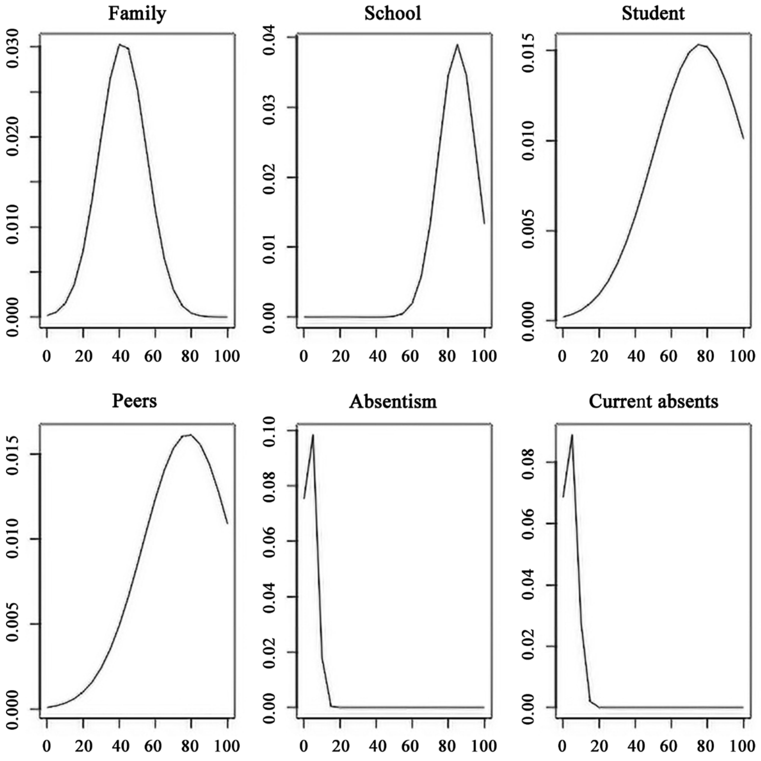

Figure 4. Distributions of individual attributes from generated data.

instances. A sample of randomly generated (seed = 80) 80% of the instances were made into a training set and the remaining percentage became the test set. This partitioning of the dataset was informed by experimentation with varying sizes of training sets as shown in Figure 6. In the figure, various sizes of training set which produces minimum of errors are shown.

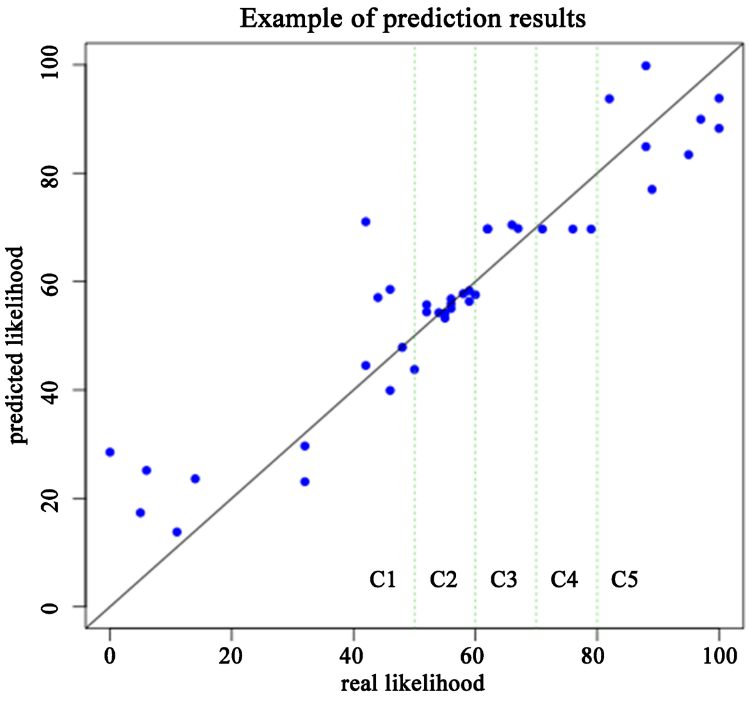

Figure 5. Performance of prediction model.

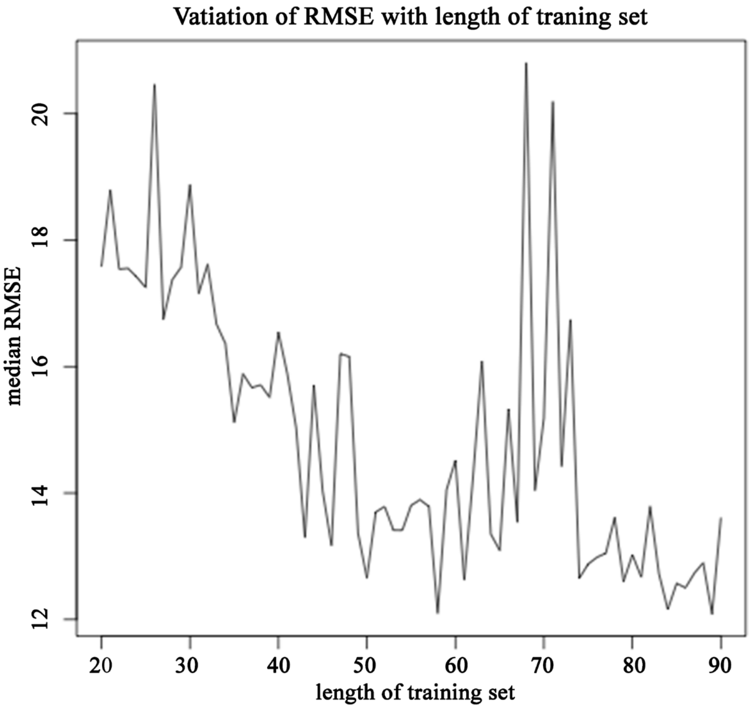

Figure 6. Data sizee and learning performance.

2.5. Neural Network Learning

Neural network technique was preferred for the study because of two main reasons. First, neural networks can learn non-linear and complex relationships which are expected of school dropout problem. Second, neural networks do not restrict input variables which represent dropout determinants. Nonetheless, any other relevant machine learning algorithm could have been used for demonstration. In this study, a standard practice of learning neural networks was followed. The six attributes based on selected determinants formed inputs to a neural network after normalization. For experimentation and demonstration purposes, the neural network was set to have 3 hidden layers. Two activation functions were tested—linear and logistic. The output of the neural network is the dropout likelihood in percentage. Once trained, the performance of the neural network was measured using accuracy of prediction of known cases and learning error based on mean square error.

2.6. Decision Mechanism

A threshold based mechanism of tracking and categorizing students is used whereby a cut-off likelihood level for dropping out is defined. This is a point of highest risk of dropping out and constitutes a new concept hereby referred to as Dropout Risk Appetite (DRA), *Definition* 2.4 below. Before reaching DRA, students pass through various risk stages as summarized in **Table 1** using experimental threshold values that were arbitrarily selected. In practice, these are configurable values defined by either schools or education authorities.

Definition 2.4: Dropout Risk Appetite is the level of risk that schools can ignore when tracking student's likelihood of dropping before deliberate proactive intervention becomes necessary.

2.7. Scenario of Incomplete Student Data

The developed prediction model was tested for production environment where available student data is incomplete e.g. environments with scenarios for newly transferred students with missing absenteeism data. Two such data scenarios for testing the predictor were designed as follows (i) scenario 1: student with average absenteeism and family ranking data only (ii) scenario 2: student with current cumulative absenteeism, peer's performance and student performance data

Table 1. Dropout risk level and thresholds.

Stage	Description	Threshold
None	Students with zero or reasonable absenteeism	C1 = 50
New	First time abnormal absenteeism	C2 = 60
Medium	Regular absenteeism	C3 = 70
High	Students who crossed the dropout risk appetite	C4 = 80
Dropped	Students who have dropped	C5 (NA)

only. For each scenario, the performance of the predictor was tested using accuracy. When testing the scenarios, all missing data attributes were set to default values (shown in brackets, giving benefit of doubt) as follows: family ranking (1), school ranking (1), student performance (1), peer performance (1), average absenteeism (0) and current cumulative absenteeism (0).

3. Results

3.1. Prediction Model

An example of a learned neural network based predictor of dropout likelihood is shown in Figure 7. This is a typical neural network with inputs matching the selected determinants and output value that signify the desired likelihood in percentage. The used predictor has 3 hidden layers, 21 neurons, associated weights and 4 bias. Any size of neural network could have served the purpose of this paper as long as it provides the desired level of performance. The weights in the model are automatically generated during learning process. The output in percentage allows applications to make prediction decision using the threshold mechanism presented in the Methods Section.

3.2. Performance of the Neural Network

The learned neural network was validated using the test dataset and the decision mechanism. Table 2 presents a sample of the student likelihood data used for

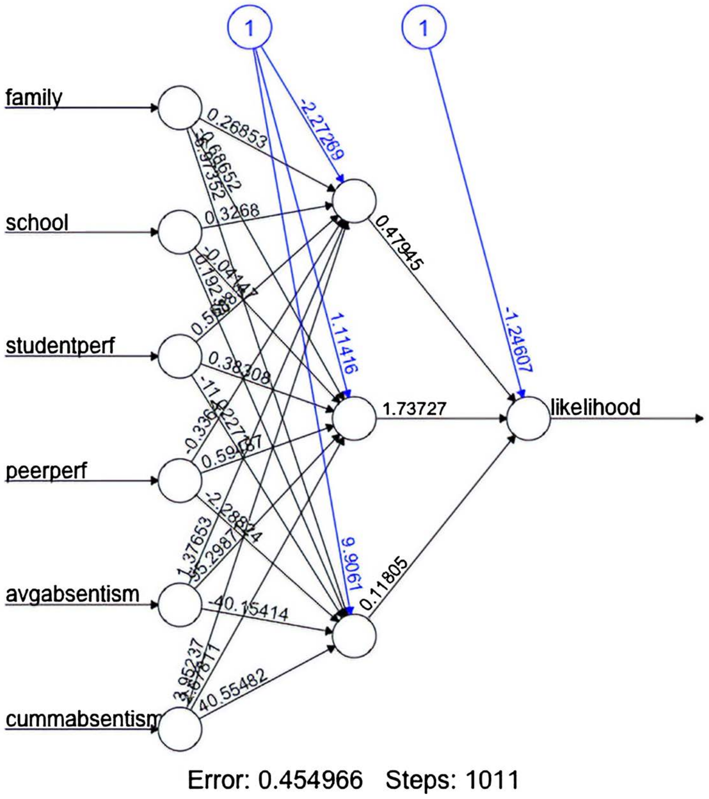

Figure 7. Neural network prediction model.

Table 2. Examples of prediction outcomes.

Student	actual likelihood	predicted likelihood	actual status	predicted status
5	82	93.73259	Dropped	Dropped
13	46	39.91505	None	None
17	50	43.77739	New	None
23	52	54.39903	New	New
27	32	29.65741	None	None
30	14	23.61449	None	None
32	76	69.69127	High	High
38	79	69.69103	High	High
39	0	28.52225	None	None
40	44	57.05550	None	New
41	60	57.56848	Medium	New

validation. The results were graphically shown in Figure 5 where the data items fall along the diagonal of an XY plot of predicted and actual likelihood values. The prediction decision mechanism is indicated by the dotted vertical green lines separating zones for thresholds C1, C2, C3, C4 and C5 shown in Table 1. Students in the region C5 are the dropouts. This performance of the predictor is equivalent to an accuracy of 75.61%.

3.3. Students with Incomplete Data

The learned neural network was also tested for use in scenarios where there is missing determinants. The results of the two scenarios which were tested are shown in Figure 8 and Figure 9 with accuracies 58.54% and 41.46% respectively. Scenario 1 considered a student with data on average absenteeism and family influence only. Scenario 2 considered a student with data on current cumulative absenteeism, peers' academic performance and student's academic performance.

4. Discussion

The results presented in the preceding section validate the proposed approach and provide a guide to researchers and practitioners. Users of the approach will still need to optimize few things in order to achieve their desired level of learning and prediction performance which was not the scope of the current study. This study has depicted that schools with no data can initially use well crafted artificial data to build machine learning models. In every semester or term, as new data comes in, the models can be improved. More studies are needed on how to update the models with minimum efforts. It is also important to note that there are limitation surrounding this study as follows:

- It is not clear how the results will be when real data is used. Therefore, in future, once adequate data is obtained, these results will have to be validated again.

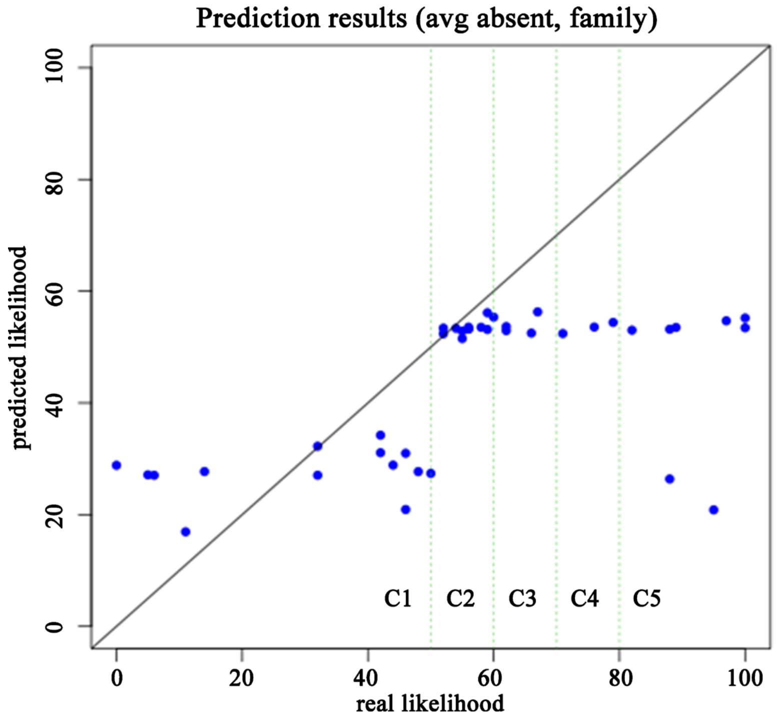

Figure 8. Prediction using past attendance records and family influence alone.

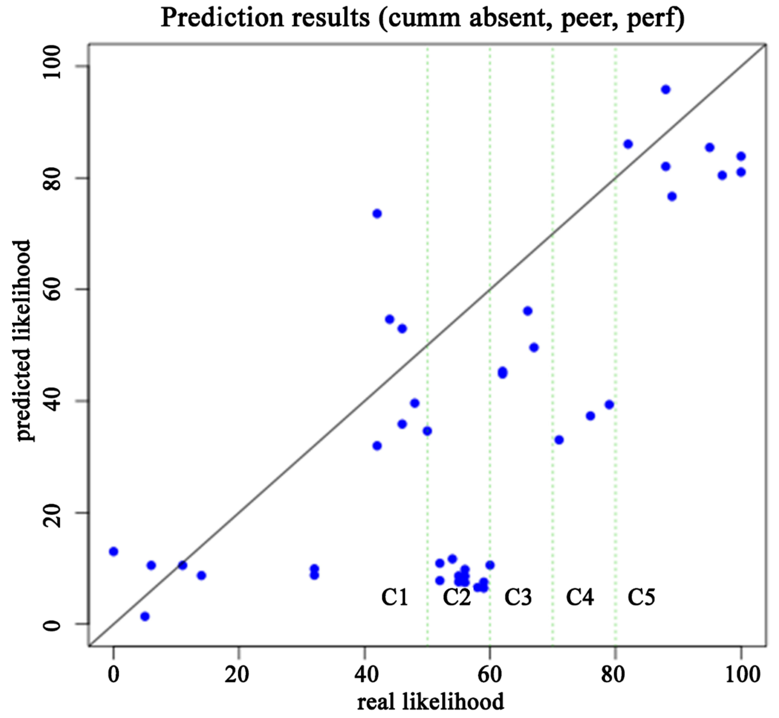

Figure 9. Prediction using current attendance data, and academic performanceonly.

- There was no optimization of parameters used to learn neural network model in order to achieve best performance. In this study, a need to thoroughly optimize the model was deemed less important because of the use of synthetic data. For instance, linear activation function was used while there are evidence that functions such as rectified linear unit works better in similar scenarios. Moreover, learning errors can be reduced by employing back propagation techniques to optimize the weights of the neuron.

- This study has assumed that schools and relevant authorities have approaches to aggregate data and produce ranking from the dropout perspective. For instance, schools can have a mechanism to assign scores (out of 100%) to families using such criteria as single parenting, parent education, and employment status, to mention a few.

5. Conclusion

This article promotes the use of and advocates further study on data-driven approaches to manage school dropout problems. An approach to use synthetic data to develop machine learning model that can be used in real life was validated. The approach was based on a newly proposed conceptual framework underlaying new concepts of average absenteeism, current cumulative absenteeism and dropout risk appetite. Both of these concepts have been defined and tested in this article. The article has also presented limitations of the current study in order to catalyse more research on the topic. For sub-saharan African countries where dropout problem persists and previous ICT infrastructure investments have strengthened data collection, this article saves to provide awareness to stakeholders involved in formulation of relevant polices at national, local and school levels.

Acknowledgements

I would like to express sincere gratitude to NAMA Foundation for their financial support to project activities that led to the publication, and EagleAnalytics Company Limited for hosting the study.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Raaijmakers, S. (2019) Artificial Intelligence for Law Enforcement: Challenges and Opportunities. *IEEE Security and Privacy*, **17**, 74-77. https://doi.org/10.1109/MSEC.2019.2925649

[2] du Boulay, B. (2016) Artificial Intelligence as an Effective Classroom Assistant. *IEEE Intelligent Systems*, **31**, 76-81. https://doi.org/10.1109/MIS.2016.93

[3] Cai, L. and Zhu, Y. (2015) The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, **14**, 2. https://doi.org/10.5334/dsj-2015-002

[4] Theunissen, M.-J., Bosma, H., Verdonk, P. and Feron, F. (2015) Why Wait? Early Determinants of School Dropout in Preventive Pediatric Primary Care. *PLoS ONE*, **10**, e0142315. https://doi.org/10.1371/journal.pone.0142315

[5] Lee, S. and Chung, J. (2019) The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. *Applied Sciences*, **9**, 3093. https://doi.org/10.3390/app9153093

[6] Tan, M. and Shao, P. (2015) Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method. *International Journal of Emerging Technologies in Learning* (*iJET*), **10**, 11-17. https://doi.org/10.3991/ijet.v10i1.4189

[7] Sivakumar, S., Venkataraman, S. and Selvaraj, R. (2016) Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree. *Indian Journal of Science and Technology*, **9**, 1-5. https://doi.org/10.17485/ijst/2016/v9i4/87032

[8] Kalinga, T. (2013) Causes of the Dropout in Secondary School in Tanzania: The Case Study of Mbeya, Dar es Salaam and Kilimanjaro Regions. Master's Thesis, The Open University of Tanzania, Dar es Salaam.

[9] Cortez, P. and Silva, A. (2008) Using Data Mining to Predict Secondary School Student Performance. *Proceedings of* 5*th Annual Future Business Technology Conference*, Porto, 9-11 April 2008, 5-12.

[10] Mundell, J. (2017) Africa's Data Revolution: Accelerating Development through Data-Driven Decision-Making. https://www.inonafrica.com/2017/10/02/africas-data-revolution-accelerating-development-data-driven-decision-making/

[11] PORALG (2016) Pre-Primary, Primary and Secondary Education Statistics in Brief. The Government of the United Republic of Tanzania, Dodoma.

[12] Jaderberg, M., Simonyan, K., Vedaldi, A. and Zisserman, A. (2014) Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition.

[13] Gupta, A., Vedaldi, A. and Zisserman, A. (2016) Synthetic Data for Text Localisation in Natural Images. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, 27-30 June 2016, 2315-2324. https://doi.org/10.1109/CVPR.2016.254

[14] Le, T., Baydin, A., Zinkov, R. and Wood, F. (2017) Using Synthetic Data to Train Neural Networks Is Model-Based Reasoning. 2017 *International Joint Conference on Neural Networks* (*IJCNN*), Anchorage, AK, 14-19 May 2017, 3514-3521. https://doi.org/10.1109/IJCNN.2017.7966298

[15] Nicholson, C. (2019) A.I. Wiki-Open Datasets. https://pathmind.com/wiki/open-datasets

[16] Gonfalonieri, A. (2019) Do You Need Synthetic Data for Your AI Project? https://towardsdatascience.com/do-you-need-synthetic-data-for-your-ai-project-e7ecc2072d6b