2019-08-22

An Ensemble Predictive Model Based Prototype for Student Drop-out in Secondary Schools

Mduma, Neema

An Ensemble Predictive Model Based Prototype for Student Drop-out in Secondary Schools

Neema Mduma [1]*, Khamisi Kalegele [2], Dina Machuve [1]

[1] The Nelson Mandela African Institution of Science and Technology, TANZANIA
[2] Tanzania Commission for Science and Technology, TANZANIA

*Corresponding Author: mduman@nm-aist.ac.tz

ABSTRACT

When a student is absent from school for a continuous number of days as defined by the relevant authority, that student is considered to have dropped out of school. In Tanzania, for instance, drop-out is when a student is absent continuously for a period of 90 days. Despite the fact that several efforts have been made to improve the overall status of education at secondary level, the student drop-out problem still persists. Taking advantage of advancement in technology, several studies have used machine learning to address the student drop-out problem. However, most of the conducted studies have used datasets from developed countries, while developing countries are facing challenges on generating public datasets to be used to address this problem. Using a dataset from Tanzania which reflect a local scenario; this study presents an ensemble predictive model based prototype for student drop-out in secondary schools. The deployed model was developed by soft combining a tuned Logistic Regression and Multi-Layer Perceptron models. A feature engineering experiment was conducted to obtain the most important features for predicting student drop-out. Furthermore, a visualization module was integrated to assist interpretation of the machine learning results and we used flask framework in the development of the prototype.

Keywords: student drop-out, predictive model, machine learning, prototype, feature engineering experiment, visualization module

INTRODUCTION

Student drop-out continues to be a serious problem regardless of the fact that education has always been a national priority for successive Tanzanian governments since independence (Wizara ya Elimu na Mafunzo ya Ufundi, 2014). This problem affects both the progress of individuals and society (Kim and Kim, 2018). A total of 5.1 million children between the age of 7 and 17 are estimated to be out of school at the lower secondary level (Human Rights Watch, 2017). For many children, education ends after primary school; only three out of five Tanzanians adolescents or 52% of the eligible school population have been enrolled in lower-secondary education and fewer complete secondary education (Human Rights Watch, 2017). The implications for finding and implementing solutions to the drop-out problem go beyond the individual benefits for students. Furthermore, investing in future progress and better standards of living with multiplier effects requires enabling students to complete their education. Therefore, making efforts that will improve this situation demands a vibrant knowledge of the extent, reasons, circumstances and the response to policies that led to the student drop-out problem.

In response to the drop-out problem and other challenges that secondary schools are facing, the government of Tanzania introduced an Education Training Policy (ETP) and Education Sector Development Plan (ESDP)

(TAMISEMI, 2004). The aim was to place emphasis on the quality of education and improve access to secondary education. These goals are in line with the Sustainable Development Goals (SDGs), a United Nations initiative that puts a target for all countries to offer free, equitable and quality primary and secondary education to children by 2030 (Truta et al., 2018). The goals are also in line with Tanzania's international and regional human rights obligations to realize the right to primary and secondary education for all (Wizara ya Elimu na Mafunzo ya Ufundi, 2014). Despite the combined efforts for improving the status of secondary school education by improving access, capacity development, quality and secondary school direct funding, the student drop-out problem still seems to persist.

Recently, machine learning technologies have gained much attention in the fight against the school drop-out problem (Elbadrawy et al., 2016; Xu et al., 2017). The use of these advanced technologies can potentially facilitate the identification of at-risk students and enable timely planning for interventions (Fei and Yeung, 2015). However, most of the existing studies have focused only on developing predictive models without including mechanism to assist interpretation of machine learning results (Aulck et al., 2016; Hung et al., 2017; Liang et al., 2016; Santana et al., 2015). Taking advantage of an increase in the number of Internet users, which is about 23 million people - almost 45% of the Tanzanian population (Maginga et al., 2018), this study intends to develop an ensemble predictive model based prototype to enable authorities in identifying at-risk students and schools for early intervention. The study uses both student and school level datasets from a developing country to address the problem with consideration of the local context. The prototype requires Internet connection to support the flow of information between the interface and the server side. The specific focus was to come up with a prototype that allows a user to input student features with high contribution to the drop-out prediction based on the feature engineering experiment conducted. The developed prototype web-based system, which can automatically recognize students with high probability of dropping out, has been constructed by implementing an ensemble algorithm (Mduma et al., 2019b). Furthermore, the system was integrated with a visualization module to highlight schools with high drop-out rates in order to help the authorities to focus on school needs during planning and budgeting processes. The idea of developing a prototype was intended to support interpretability of machine learning results using an easier approach that will be understood by users with no knowledge of machine learning.

RELATED WORK

Machine learning approaches have been used for educational purposes including developing a system for an early identification of students at risk of dropping out (Berens et al., 2018). An Early Detection System (EDS) for predicting student success in tertiary education as a basis for a targeted intervention was developed. Regression analysis, Neural networks, Decision trees and the AdaBoost algorithms were used to point out students characteristics that distinguish potential dropouts from graduates. The developed methods was then implemented in every German university. This method uses student demographic and performance data which was collected and maintained by legal mandate.

Similarly, a mobile academic performance prediction system was developed with the aim of predicting students that require early intervention (Mgala, 2016). The study used datasets of primary schools collected in Kenya. Logistic regression, Multilayer perceptron, Sequential minimal optimization algorithm (SMO), Bayesian network classifiers, Naive Bayes classifier, Lazy learners, Random forest classifier and J48 algorithm were used to build the model. However, a simple Logistic regression classifier achieved the best results. Therefore, it was used in the implementation of this mobile system.

Another study outlined an extensive framework that uses machine learning approaches to identify students who are at risk of not finishing high school on time (Lakkaraju et al., 2015). The study was done in the United States and it aimed at giving both students and schools hands-on tools based on their needs, and to assist schools in identifying and prioritizing students that are at risk of adverse academic outcomes.

In another study, a survival analysis based framework was developed to identify at-risk students (Ameri et al., 2016). A Time-dependent Cox (TD-Cox) model was applied to capture time-varying factors and to leverage this information to provide more accurate prediction of student drop-out. The framework was proposed to predict which students were likely to drop-out including the semester when the drop-out was expected to occur. This method was evaluated on real student data collected at Wayne State University.

Another example involved the use of students data gathered from the University of Barcelona (UB) to implement visualization tools for predicting academic grades and student drop-out (Rovira et al., 2017). The developed tools allowed interpretation of drop-out prediction errors based on the grades distribution.

Furthermore, one study developed a deep learning based prototype system for automated eye gaze following, that estimated where each person in a classroom was looking (Aung et al., 2018). The study aimed at helping teachers to give attention at the right thing or to the right students within classrooms. Since the focus was on
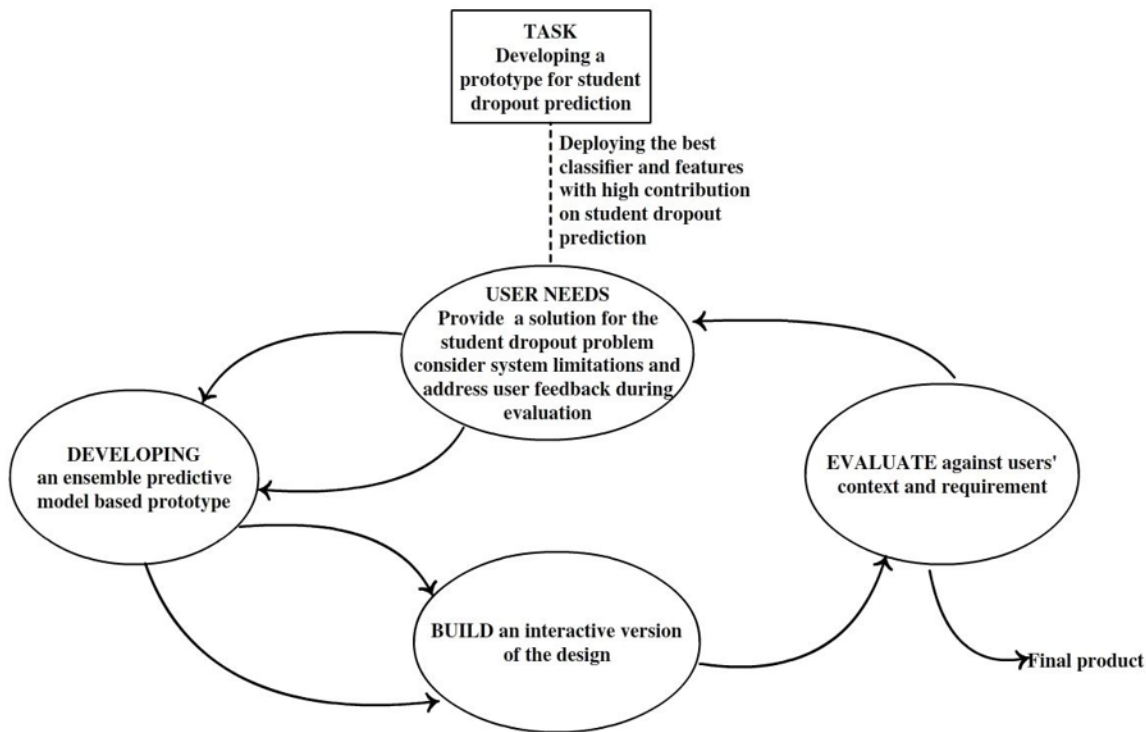
Figure 1. Prototyping software development approach

classroom observation videos, a dataset of publicly available classroom sessions from YouTube videos were collected.

Our study is based on the earlier works done in the educational field as presented in this section. However, in this study both student and school-level datasets from a developing country were used to reflect local context. As it has been observed, many existing studies focused only on student-level datasets and did not consider school-level datasets for addressing this problem (Mduma et al., 2019a). Logistic Regression, Multilayer Perceptron, Random Forest and K-Nearest-Neighbors were used to build the model. The results showed that Logistic Regression and Multi-Layer Perceptron achieved the highest performance. Furthermore, hyper-parameter tuning was performed to improve the predictive power of the well performing models and an ensemble classifier which was developed by soft combining the best performing models attained the best performance (Mduma et al., 2019b). We therefore implemented an ensemble predictive model for this prototype.

MATERIALS AND METHODS

The development of this system followed a prototyping software development approach. This approach was created to receive feedback from users for refining the final product (Nacheva, 2017; Yu, 2018). It presents the analysis, design and implementation phases so as to develop a simplified version of the system and provide users with the evaluation and feedback (Iqbal, 2017). The prototype was then improved following feedbacks from the users. The improved prototype was given back to the users for further evaluation, and the cycle continued until the users were satisfied with the final prototype as shown in **Figure 1**.

Since the system was designed primarily to help educational stakeholders in identifying at-risk students and schools; education officers, parents, teachers and information systems development experts were involved in the process of prototype development. Education stakeholders from five selected districts were involved in the focus group discussion during data collection. The technical feedback from information systems development experts were used to improve the prototype.

The study includes both functional and non-functional requirements. The functional requirements indicate what a user needs from the system, while the non-functional requirements refer to the system architecture (Alsaleh and Haron, 2016). The functional requirements for developing this prototype cover:

- The issues of predicting whether a student will drop-out or not.
- The use of features with high contribution to the drop-out prediction.
- The use of the best classifier - an ensemble algorithm developed by soft combining the tuned Logistic Regression and Multi-Layer Perceptron models.
- The issue of visualizing school drop-out.

The non-functional requirements of the system cover the issues of:
- Scalability
- Usability
- Performance
- Accessibility
- Consistency

Datasets Description

There is a dearth of studies focused on addressing student drop-out using machine learning in developing countries (Mduma et al., 2019a) and publicly available datasets addressing this problem are difficult to find. This study used datasets from Tanzania which reflect the context of a specific developing countries. We used the Uwezo data [1] collected in 2015 at the country level to develop an ensemble predictive model. This student-level dataset collected by Twaweza was assembled with the aim of evaluating children learning levels across hundreds of thousands of households in East Africa. The dataset consists of 61,340 samples of student records and 18 features:
- Parent who check his/her child's exercise book once in a week (PCCB)
- Student gender (Sex)
- Household meals per day (MLPD)
- Main source of household income (Income)
- School has girl's privacy room (SGR)
- Region
- District
- Village
- Student who did read any book with his/her parent in last week (SPB)
- Parent who discuss his/her child's progress with teacher last term (PTD)
- Student age (Age)
- Enumeration Area type (EAarea)
- Boy's Pupil Latrines Ratio (BPLR)
- Household size (HHsize)
- Pupil Teacher Ratio (PTR)
- Girl's Pupil Latrines Ratio (GPLR)
- Parent Teacher Meeting Ratio (PTMR)
- Pupil Classroom Ratio (PCR)

Data preprocessing as an important step for cleansing the dataset was done (Basu et al., 2019). Furthermore, several approaches on handling numeric values, missing values, and outliers were identified (Shahul et al., 2016). In this study, Principle Component Analysis (PCA) was performed with the purpose of diminishing the number of dimensions without losing too much information (Jiang et al., 2016).

A school-level dataset collected by the Presidents Office Regional Administration and Local Government in Tanzania (PORALG) was integrated with publicly available data accessed through Government Open Data Portal[2] to support visualization. The dataset consists of 11 features:
- Region
- District
- Ward
- School name
- Dropout Male
- Dropout Female
- Pupil Teacher Ratio (PTR)
- Pupil Qualified Teacher Ratio (PQTR)
- Pupil Classroom Ratio (PCR)
- Boys Pupil Latrine Ratio (BPLR)
- Girls Pupil Latrine Ratio (GPLR)

---

[1] http://www.twaweza.org/go/uwezo-datasets
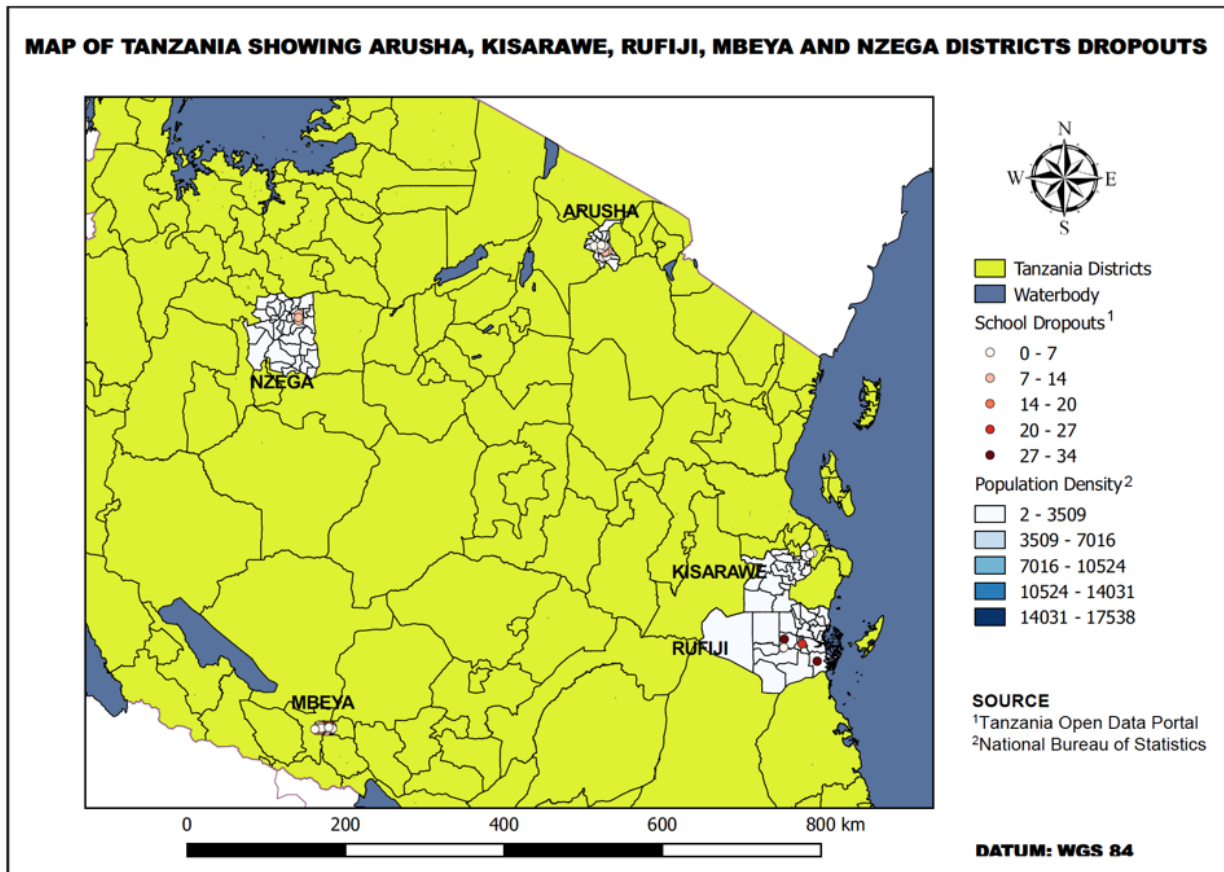[2] http://opendata.go.tz/dataset

Figure 2. Tanzania map indicating secondary schools dropout – 2016

The data is a sample of 145 secondary schools in five districts which are Mbeya, Nzega, Rufiji, Kisarawe and Arusha districts in 2016 as shown in **Figure 2**.

Model Development and Proposed Solution

The model was formulated after comprehensive analysis of widely used machine learning algorithms which represent linear, neural network, ensemble and instance models. Since data imbalance was observed during the pre-processing stage, the Synthetic Minority Oversampling Technique and Edited Nearest Neighbor (SMOTE-ENN) approach were applied to handle the problem. The dataset was split into training (60%), validation (20%) and testing (20%) sets. The sampling approach was applied only to the train set. The model was built using train and validation sets and evaluation was done using an unseen test set in order to observe model behavior in a real environment which is imbalance, the overall experimental procedure is summarized in **Figure 3**.
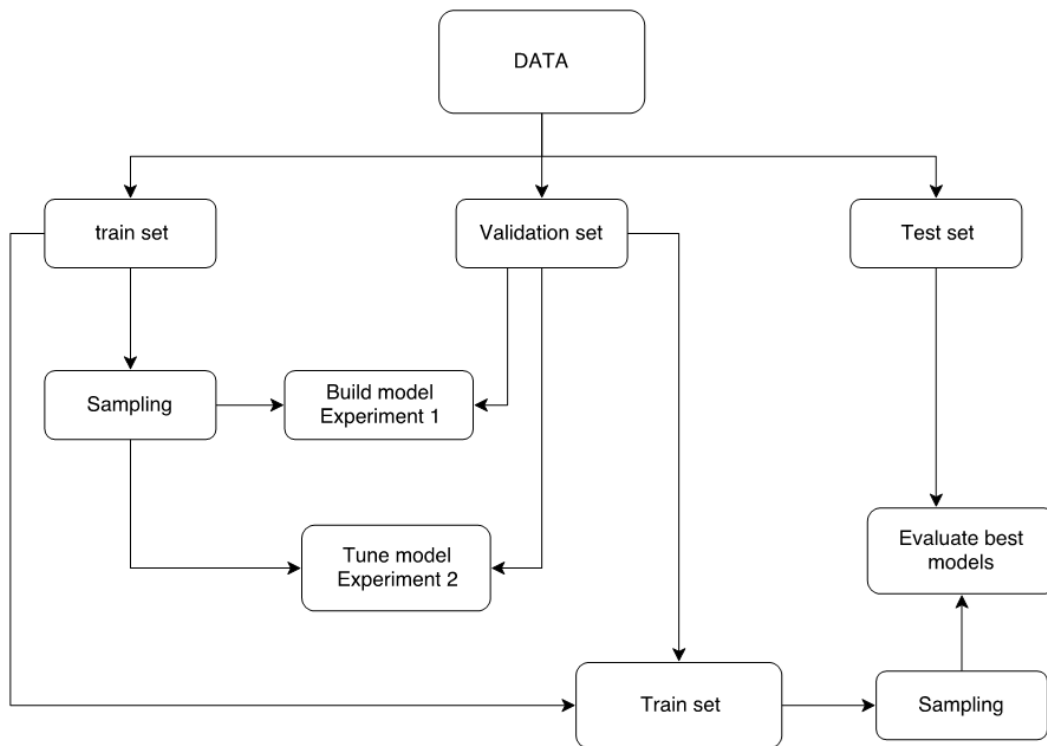
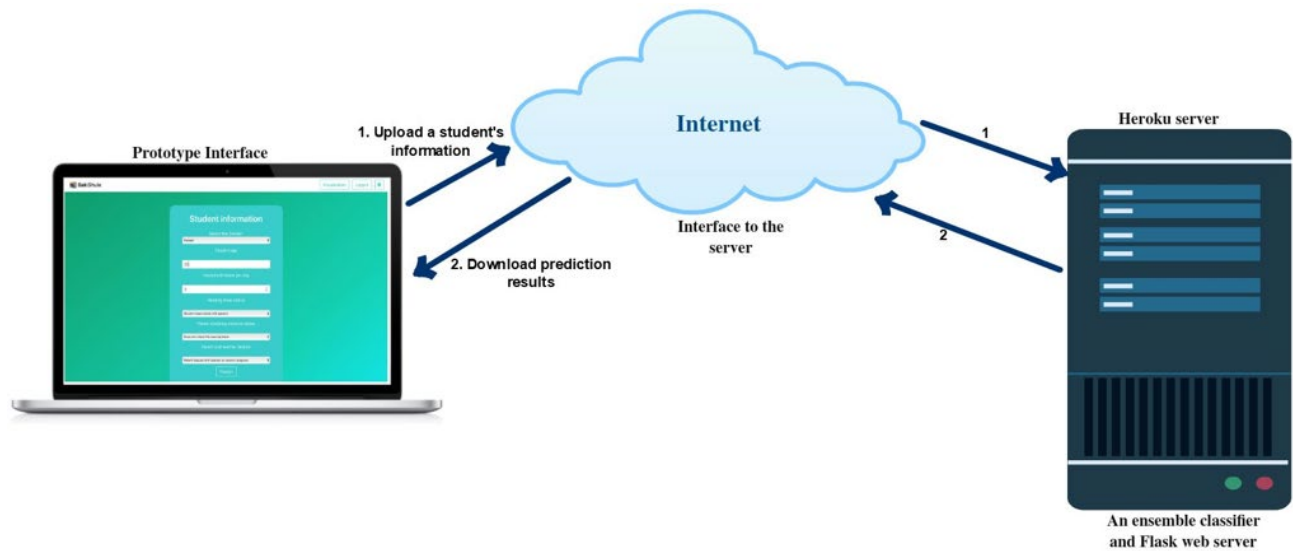**Figure 3.** Model development experimental procedure

Figure 4. Diagram of the system's architecture

F-measure (F_m), Geometric Mean (G_m) and Adjacent Geometric Mean (AG_m) metrics were used to evaluate the model and stratified 5-fold out-of-bag overall cross validation was used in the execution of the experiment.

From the architecture diagram in **Figure 4**, the prototype interface was linked to the server via the Internet. The developed prototype on the client side allows input of the students' information, comprised of features with high contribution to the drop-out prediction. The features were selected using a feature engineering experiment. In data preparation for machine learning, this approach is conducted to construct suitable features for improving predictive performance (Nargesian et al., 2017; Naz et al., 2019). This was attained by evaluating permutation of the feature importance score. The score was anticipated to measure the impact of an individual feature on the model performance by permuting values of each feature and evaluating how much the permutation decreases the model performance. The server contained an ensemble algorithm which was developed by soft combining the tuned Logistic regression and Multi-Layer Perceptron models, earlier recognized as the best model. This model was then implemented in python using Scikit-learn (Mitchell, 2015).

The server interface used Flask framework. Flask was preferred in this study due to it popularity and ability to make the core functionality simple but extensible in terms of development. It also saves time needed to build web applications (Armash et al., 2015). The developed system transferred a students information entered through the
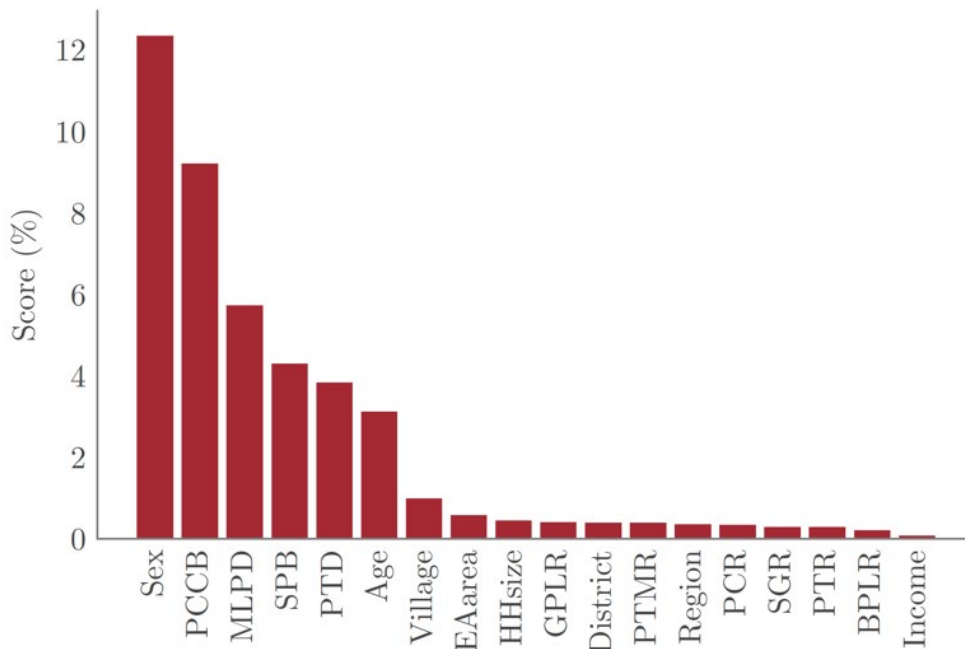
Figure 5. Feature selection experimental results

system interface via the Internet to an ensemble algorithm on the server. On the server, the deployed model predicts the result for this new entry. The result is next transferred via the Internet to the prototype interface. Flask web server facilitated the record transfer to the server and the result from the server to the system interface. For this prototype, Heroku server was used as the hosting platform to support deployment of the developed system.

RESULTS

Feature Engineering Results

The results demonstrated in **Figure 5**, indicate that Student gender (Sex), Parent who check his/her child's exercise book once in a week (PCCB), Household meals per day (MLPD), Student who did read any book with his/her parent in last week (SPB), Parent who discuss his/her child's progress with teacher last term (PTD) and Student age (Age) have high contribution on the drop-out prediction performance. These features were included in the developed prototype to serve as an input for student information.

Drop-out Prediction Interface

The interface allows the system to connect and exchange information by acting as the bridge between a user and the system (Iftikhar et al., 2018). The drop-out prediction module allows users to input student information as shown in **Figure 6** and prediction was given based on the provided information. The system then provided prediction results to indicate whether a given student will drop-out or not. This module was developed to assist parents and teachers on identifying at risks student who are in most need of help.

Drop-out Visualization Interface

Visualization has been recognized as an important approach to understanding data (Xin et al., 2018). This technique has been used to support interpretation of machine learning results. The school-level dataset was visualized to highlight school drop-out within selected districts as shown in **Figure 7**. The intention was to assist education stakeholders on identifying at-risk schools in order to provide requirements based on the school needs.
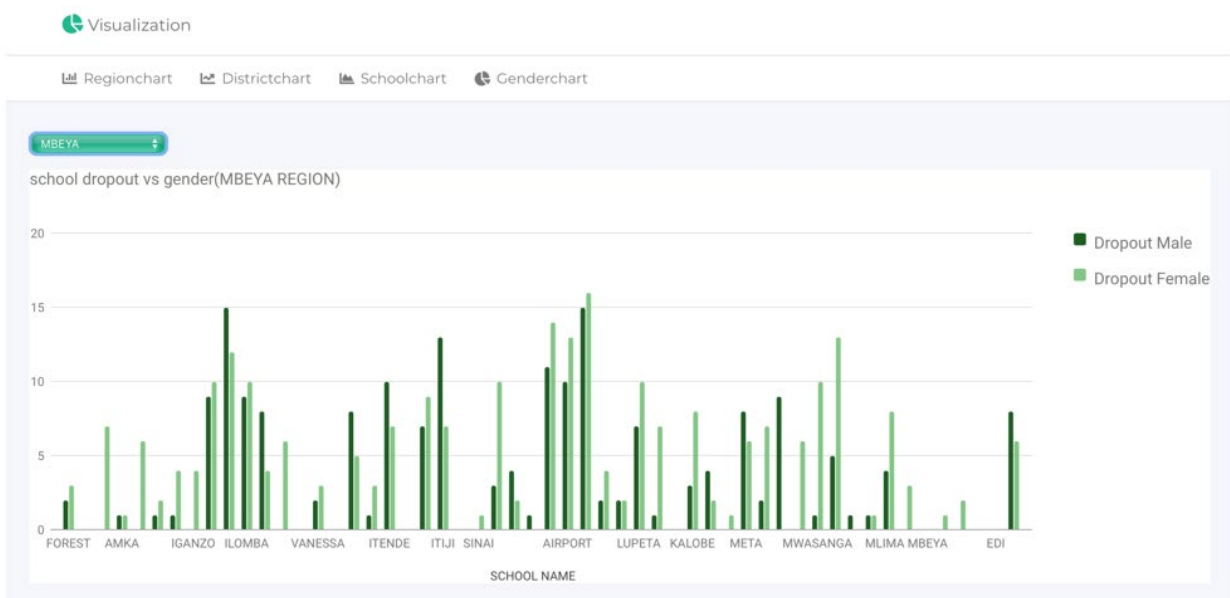
Figure 6. Drop-out prediction interface

Figure 7. Drop-out visualization interface

DISCUSSION AND CONCLUSION

An ensemble predictive model based prototype has been developed to predict student drop-out as declared in this study. The developed system whose requirements specifications were narrated in this paper helps in the identification of at-risk students for early intervention. By taking advantage of Internet penetration within the country and the use of machine learning technology, this system is directly going to benefit education stakeholders in identifying at-risk students and schools. Authorities will be able to use the developed system to facilitate the planning and budgeting process in order to provide school needs based on the requirements. The development of this system considered both users with/without knowledge of basic computer skills.

Several studies in developed countries have applied machine learning techniques to tackle student drop-out (Elbadrawy et al., 2016; Fei and Yeung, 2015; Xu et al., 2017). However, few studies focused on developing a prototypes to assist education stakeholders in the interpretation of machine learning results (Aung et al., 2018; Rovira et al., 2017; Berens et al., 2018). A mobile based tool was developed in a developing country to fight against

the drop-out problem; however, the study focused only on student-level dataset which is not publicly available (Mgala, 2016). Furthermore, feature engineering results that show student gender has high contribution to the drop-out prediction support researchers' findings on drop-out rate with gender association (Kim and Kim, 2018). Therefore, focus should be directed not only on developing predictive models on addressing the problem but also on providing a room for intended users to be able to interact with the developed approach. This can be achieved by implementing the developed models in the systems for easy understanding. Furthermore, evaluation of the developed systems must be taken into consideration to ensure that the systems address the users needs. Additionally, cost and time limitations should be considered when generating new datasets to be used in addressing this problem. This can be achieved by emphasizing the identification of available datasets as done in this study in order to attract other researchers in the education field to provide solutions needed to address the student drop-out problem.

This paper presents an ensemble predictive model based prototype to help education stakeholders in the early detection of student drop-out in Tanzania. An ensemble classifier which was obtained by soft combining the tuned Logistic Regression and Multi-layer Perceptron was implemented in this prototype. Six features with significant contributions to the drop-out prediction were used as inputs for student information. Furthermore, the prototype was integrated with a visualization module to facilitate interpretation of machine learning results. In particular, the developed system predicted whether a given student will drop-out or not and visualized schools with high drop-out risks. Therefore, this study is limited on identifying at risk students and schools using a web-based approach. Inclusion of other components such as ranking and forecasting mechanisms will be an added advantage on facilitating a more robust and comprehensive early warning systems for students dropout.

Publicly available datasets have been identified to provide room for other researchers in the field of education to apply different approaches to solve the problem of student drop-out. This work proves the value of machine learning approaches on addressing drop-out prediction. The study complements previous research done by other researchers in developed countries using developed countries datasets. Regarding educational implications, the developed system can be extremely useful for education stakeholders, that will be able to recognize earlier which students and schools need help. This information will assist them in providing early intervention. Future directions of this study will be focused on evaluating performance of the developed system and developing a mobile application based on the developed prototype.

ACKNOWLEDGEMENT

REFERENCES

Alsaleh, S. and Haron, H. (2016). The Most Important Functional and Non-Functional Requirements of Knowledge Sharing System at Public Academic Institutions: A Case Study. *Lecture Notes on Software Engineering, 4*(2), 157–161. https://doi.org/10.7763/LNSE.2016.V4.242

Ameri, S., Fard, M. J., Chinnam, R. B. and Reddy, C. K. (2016). *Survival Analysis based Framework for Early Prediction of Student Dropouts.* https://doi.org/10.1145/2983323.2983351

Armash Aslam, F., Nabeel Mohammed, H., Jummal Musab, M. and Munir Murade Aaraf Gulamgaus, H. (2015). Efficient Way of Web Development Using Python and Flask. *International Journal of Advanced Research in Computer Science, 6*(2), 54–57.

Aulck, L., Velagapudi, N., Blumenstock, J. and West, J. (2016). Predicting Student Dropout in Higher Education.

Aung, A. M., Ramakrishnan, A. and Whitehill, J. R. (2018). Who are they looking at? Automatic Eye Gaze Following for Classroom Observation Video Analysis. *Educational Data Mining.*

Basu, K., Basu, T., Buckmire, R. and Lal, N. (2019). Predictive Models of Student College Commitment Decisions Using Machine Learning. *Data, 4*(2), 65. https://doi.org/10.3390/data4020065

Berens, J., Oster, S., Schneider, K. and Burghoff, J. (2018). Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods. *Schumpeter School of Business and Economics*, pp. 0–32.

Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G. and Rangwala, H. (2016). Predicting student performance using personalized analytics. *Computer, 49*(4), 61–69. https://doi.org/10.1109/MC.2016.119

Fei, M. and Yeung, D. Y. (2015). Temporal models for predicting student dropout in massive open online courses. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 256–263. https://doi.org/10.1109/ICDMW.2015.174

Human Rights Watch (2017). I Had a Dream to Finish School: Barriers to Secondary Education in Tanzania.

Hung, J. L., Wang, M. C., Wang, S., Abdelrasoul, M., Li, Y. and He, W. (2017). Identifying At-Risk Students for Early Interventions - A Time-Series Clustering Approach. *IEEE Transactions on Emerging Topics in Computing, 5*(1), 45–55. https://doi.org/10.1109/TETC.2015.2504239

Iftikhar, W., Sheraz, M., Malik, A., Tariq, S., Saad, M., Ahmad, J. and Shareef, F. (2018). User Interface Design Issues In HCI. *IJCSNS International Journal of Computer Science and Network Security, 18*(8), 153–157.

Iqbal, S. Z. (2017). Z-SDLC Model A New Model For Software Development Life Cycle (SDLC). *International Journal of Engineering and Advanced Research Technology (IJEART), 3*(2), 8.

Jiang, Y. H., Javaad, S. S., and Golab, L. (2016). Data mining of undergraduate course evaluations. *Informatics in Education, 15*(3), 85–102. https://doi.org/10.15388/infedu.2016.05

Kim, D. and Kim, S. (2018). Sustainable Education: Analyzing the Determinants of University Student Dropout by Nonlinear Panel Data Models. *Sustainability, 10*(4), 954. https://doi.org/10.3390/su10040954

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R. and Addison, K. L. (2015). A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. *Kdd*, 1909–1918. https://doi.org/10.1145/2783258.2788620

Liang, J., Li, C. and Zheng, L. (2016). Machine learning application in MOOCs: Dropout prediction. *ICCSE 2016 - 11th International Conference on Computer Science and Education, (Iccse)*, pp. 52–57. https://doi.org/10.1109/ICCSE.2016.7581554

Maginga, T. J., Nordey, T. and Ally, M. (2018). Extension System for Improving the Management of Vegetable Cropping Systems. *Journal of Information Systems Engineering & Management, 3*(4). https://doi.org/10.20897/jisem/3940

Mduma, N., Kalegele, K. and Machuve, D. (2019a). A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction. *Data Science Journal, 18*, 1–10. https://doi.org/10.5334/dsj-2019-014

Mduma, N., Kalegele, K. and Machuve, D. (2019b). Machine learning approach for reducing students dropout rates. *International Journal of Advanced Computer Research, 9*(42). https://doi.org/10.19101/IJACR.2018.839045

Mgala, M. (2016). *Investigating Prediction Modelling of Academic Performance for Students in Rural Schools in Kenya*. PhD thesis, University of Cape Town.

Mitchell, T. (2015). *Machine Learning With Python Scikit-Learn-Application to the Estimation of Occupancy and Human Activities*. Number July.

Nacheva, R. (2017). Prototyping Approach in User Interface. In *2nd Conference on Innovative Teaching Methods*, number June, pp. 80–87, Bulgaria.

Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B. and Turaga, D. (2017). Learning feature engineering for classification. *IJCAI International Joint Conference on Artificial Intelligence*, (August), pp. 2529–2535. https://doi.org/10.24963/ijcai.2017/352

Naz, Zafar and Khan (2019). Ensemble Based Classification of Sentiments Using Forest Optimization Algorithm. *Data, 4*(2), 76. https://doi.org/10.3390/data4020076

Rovira, S., Puertas, E. and Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PLOS ONE, 12*(2), 1–21. https://doi.org/10.1371/journal.pone.0171207

Santana, M. A., Costa, E. B., Neto, B. F. S., Silva, I. C. L., and Rego, J. B. A. (2015). A predictive model for identifying students with dropout profiles in online courses. *CEUR Workshop Proceedings*, 1446.

Shahul, S., Suneel, S., Rahaman, M. A. and Swathi, J. N. (2016). A Study of Data Pre-Processing Techniques for Machine Learning Algorithm to Predict Software Effort Estimation. *Imperial Journal of Interdisciplinary Research, 2*(6), 2454–1362.

TAMISEMI (2004). The United Republic of Tanzania Ministry of Education and Culture. Pp. 2004–2009.

Truta, C., Parv, L. and Topala, I. (2018). Academic engagement and intention to drop out: Levers for sustainability in higher education. *Sustainability, 10*(12), 1–11. https://doi.org/10.3390/su10124637

Wizara ya Elimu na Mafunzo ya Ufundi (2014). Sera ya Elimu na Mafunzo. Technical report.

Xin, Y., Claude, B., Rob, V., Bertoni, A. and Phillipe, E. (2018). Data Visualization in Conceptual De- sign: Developing a Prototype to Support Decision Making. In 12th International Conference on Modeling, Optimization and SIMulation - MOSIM'18, number July, page 12.

Xu, J., Moon, K. H. and van der Schaar, M. (2017). A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs. *IEEE Journal of Selected Topics in Signal Processing, 11*(5), 742–753. https://doi.org/10.1109/JSTSP.2017.2692560

Yu, J. (2018). Research Process on Software Development Model. *IOP Conference Series: Materials Science and Engineering, 394*(3).