PROACTED: An AI-Driven Course Recommender, Academic Performance Monitoring and Intervention Model

Cynthia Amol
Department of Computer Science, Maseno University, Kenya
Maseno Centre for Applied Artificial Intelligence, Kenya
cynthia@maseno.ac.ke

Benaiah Galavu
Department of Pure and Applied Mathematics, Maseno University, Kenya

Mercy Okeyo
Department of Computer Science, Maseno University, Kenya

Simon Muthungu
Department of Computer Science, Maseno University, Kenya

Vitalis Amakalu
Department of Computer Science, Maseno University, Kenya

Abstract

The use of data analysis techniques such as Education Data Mining (EDM) with predictive Machine Learning (ML) models in course recommendation and performance monitoring of students helps improve completion rates of students in institutions of learning. Most models focus on monitoring student performance using reactive parameters such as end-of-semester marks unlike proactive parameters such as student attendance and continuous assessment tests which provide metrics for educator intervention before the students are at risk of dropping out. Coupled with accurate interest and academic capability-based course recommender modules, the proactive performance monitoring and interactive intervention model developed by this study enables students to self-assess. It also facilitates educators with data-driven insights to allow them to take active measures in monitoring student performance and intervening based on need. The Sentence-BERT (S-BERT)-based recommender model offers a selection of relevant courses for students to pursue based on interests and academic capabilities with an accuracy of 0.96. The predictive Logistic Regression (LR)-based performance monitoring model provides completion probabilities with an accuracy of 0.97.

1 Introduction

On average, only 6 out of 10 students enrolled in institutions of higher learning in sub-Saharan Africa manage to complete their studies (World Bank, 2021). Uninformed course selection and lack of proper student performance monitoring and intervention systems are some of the major contributors to these low completion rates. Most universities, especially in developing countries, do not take

6th Deep Learning Indaba Conference (DLI 2024).

advantage of EDM to understand the student learning environment (Albreiki et al., 2021) to improve teaching and learning (Hicham et al., 2020). Valuable data gained by exploration of EDM techniques (Kovalev et al., 2020) such as personal, academic and faculty statistics can be used not only in self-evaluation by the students themselves but also in designing educational intervention mechanisms by educators (López Zambrano et al., 2021).

Previous studies discussed in *section 2* propose the use of predictive modelling using both classical machine learning algorithms such as Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT) and Artificial Neural Networks such as Multi-Layer Perceptron (MLP) to gauge student success or completion rates. These models use student and faculty data such as student age, faculty curriculum, student frequency of attendance, weighted grade points from transcripts and student finances (Ahmed and Elaraby, 2014) as input and output the probability of completing a course or graduate. However, only a few of these attributes are used which often paints an unclear picture of the situation in schools. Additionally, most of these studies are focused on the end goal, completion, and not proactively monitoring student performance and active intervention by educators to boost completion rates.

The PROACTED model developed by this study addresses three key challenges in academic performance monitoring and intervention: course recommendation, student performance monitoring and intervention by educators. The ML-based course recommender model provides a unique platform for pre-university students to select courses based on personal interests and academic capabilities (Esteban et al., 2020). The model was SBERT-based (Reimers and Gurevych, 2019). It generated semantically meaningful sentence embeddings (Chowdhury, 2022) - vector representations of the text from input sentences (data on student interests and academic performance). The embeddings were compared using cosine similarity (Pardos and Jiang, 2020) to generate the most similar courses as output. The performance monitoring module utilised a predictive LR-based model to give a probability score of whether the student would pass the program, offering insights for educators to set up intervention measures. Use of of ML algorithms, according to Bird et al. (2021), is a better way of identifying at-risk students compared to traditional monitoring of final marks alone.

This study's contributions are:

1. An integrated ML-driven course recommender model that offers course selections based on student interests and academic capacity.

2. A ML-driven student performance prediction model that uses proactive metrics to evaluate performance for intervention.

2 Related Work

According to the report by the World Bank on higher education performance in Kenya (Bank, 2019), the completion rate for public universities is around 55%. This is due to, among other factors, inadequate academic preparedness and limited institutional support in terms of academic advising and support structures. Current ad-hoc academic advising structures, especially in over-populated public institutions are insufficient. Exploration of data mining techniques supported by Artificial Intelligence (AI) (Bursać et al., 2019) offers a better, more reliable solution to student performance monitoring and intervention.

The use of EDM, which employs data mining tools and techniques (Aleem and Gore, 2020), statistics and ML algorithms (Romero and Ventura, 2010) has been fueled by digitization of educational materials. With the advent of big data generated by both virtual and physical learning, EDM methods are often used to analyse learning and learners (Romero and Ventura, 2010). Studies by Hooda et al. (2022); Mubarak et al. (2021); Hashim et al. (2020) have utilised EDM and Learning Analytics (LA) with classical and deep learning algorithms to predict student performance.

Recommender systems (RS) are integral in providing relevant suggestions on various aspects of life (Afsar et al., 2022). Educational Recommender Systems (ERS) are often used in tailoring education content to provide a more personalised learning approach (da Silva et al., 2023). The use of ERS in content recommendations to students (Esteban et al., 2020), mostly used on online learning platforms (Ashraf et al., 2021), is useful in ensuring that students choose the right courses according to their personal interests and academic capabilities.

In previous studies, (Pallathadka et al., 2023; Zeineddine et al., 2021; Tsiakmaki et al., 2020; Miguéis et al., 2018; Wang et al., 2018) predictive modelling has been used to monitor the performance of students. These studies use EDMs and other data mining techniques to curate datasets with both student and faculty data from institution databases (Albreiki et al., 2021) or online Learner Management Systems (LMS) (Cardona et al., 2023).

Metrics such as frequency of attendance, level interaction with learning materials, gender, race, GPA, income and grades (Ahmed and Elaraby, 2014) were used as inputs to regression or classification models. These models use various traditional ML algorithms such as LR, Random Forest (RF) and deep learning models. For the development of baseline models that are easy to implement and interpret (Bird et al., 2021), especially with limited data, simpler LR and RF-based models are preferred. The output of these models were probabilities of the students completing relevant courses or programs.

3 Methodology

A mixed research design was used in this study: an exploratory research design was employed to curate the dataset for input data to the course recommender and prediction models, ensuring comprehensive and representative data collection. This involved generating synthetic data based on anonymized samples to maintain student privacy while capturing essential student attributes.

Following this, an experimental research design was implemented to develop and fine-tune the models, systematically testing and optimizing various algorithms to achieve accurate and reliable predictions and recommendations. This approach allowed for a thorough exploration of data characteristics and a rigorous evaluation of model performance. The methodology process is summarised in *Figure 1.*

In the figure, the recommender model's input is a combination of descriptions of student interest and academic qualification. The model produces course recommendations as output. Once the student selects a relevant program and is enrolled, the prediction model which tracks student performance uses this as one of the inputs to predict the student's probability of passing.

3.1 Dataset

Due to privacy concerns, access to real student data was not provided by the institution. Instead, we synthetically generated data using anonymized samples collected from the institution.

3.1.1 Data Processing

Standard preprocessing steps were employed to ensure high-quality data for model training. We verified the consistency and correctness of the data entries and removed duplicate values, ensuring that numerical values fell within the expected ranges. The dataset was divided into training and testing sets in an 80:20 ratio, providing a robust framework for evaluating the models' performance without bias.

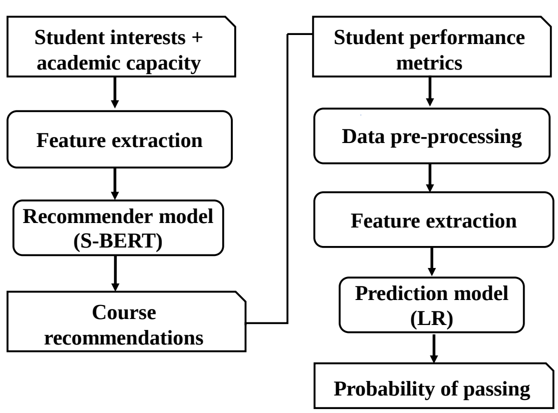

Figure 1: Methodology

Data Normalization

The numerical features in the dataset were standardized using the *StandardScaler* function. This transformation involved subtracting the mean and scaling to unit variance for each feature, which is essential for algorithms sensitive to the scale of input data.

Handling Imbalanced Data

Given the potential imbalance between the number of students who passed and those who did not, the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) was employed. SMOTE over-sampled the minority class in the training dataset, ensuring the model was not biased toward the majority class during training.

3.1.2 Feature Selection

The attributes used included: *the percentage of lessons attended compared to the total number of lessons in the stated period, aggregate points, homework submission rate, activity on online learning platforms, Continuous Assessment Tests (CATs) marks, and percentage of deadline adherence.* These features cover a broad spectrum of academic and behavioural metrics, providing a holistic view of student performance. The large size and diversity of the dataset ensured that the models could learn from varied patterns of student behavior and performance, enhancing their predictive accuracy and generalisability.

A correlation matrix was used to assess the relationships between features. Recursive Feature Elimination (RFE) was applied to identify and rank features based on their importance to the model. RFE is a feature selection algorithm that applies gradual elimination of unimportant features (Priyatno et al., 2024) in tasks such as prediction.

3.1.3 Dataset Analysis

The dataset used for model training and testing consisted of approximately 100,000 instances, each representing an individual student's record.

3.2 Modelling

Two models were developed: the course recommender model and the predictive model for performance evaluation.

3.2.1 Recommender Model

The SBERT (Reimers and Gurevych, 2019) model was used for the recommendation task. SBERT is a pre-trained model designed to derive semantically meaningful sentence embeddings, making it well-suited for NLP tasks such as similarity detection. In the recommender model, SBERT was fine-tuned to adapt it to the specific requirements of the recommendation system. This process involved additional training on our specific dataset comprising student responses about their interests, activities, and academic performance. Fine-tuning the model ensured that the embeddings generated were highly relevant to the context of educational recommendations, enhancing the quality and accuracy of the course suggestions.

The embeddings generated by SBERT were then used in the recommendation algorithm to match students with courses that align with their stated interests and previous academic performance. Key hyperparameters and configurations for using SBERT included:

Embedding Dimensions: The dimensionality of the sentence embeddings generated by SBERT was 300.

Batch Size: The number of sentences processed together during inference was optimized for performance.

Similarity Measure: Cosine similarity was used to compare sentence embeddings, determining the relevance of courses based on student responses. By fine-tuning SBERT, the system efficiently utilised advanced NLP capabilities to enhance the recommendation process, ensuring that students received personalized and meaningful course suggestions. The results are shown in *Table 1.* A sample of the recommender model input and resulting output are shown in *Figure 2.*

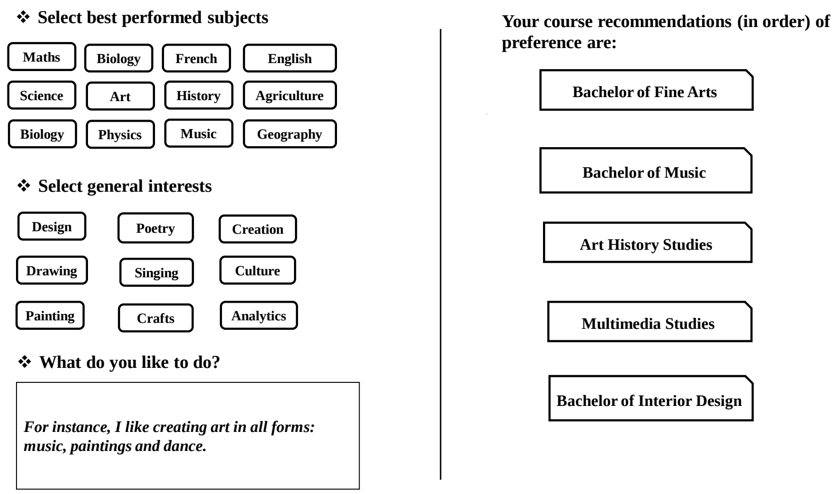

Figure 2: Sample Recommender System Input/Output

3.2.2 Prediction Model

The LR model was trained to predict the likelihood of students completing school based on various academic and engagement metrics. The training process involved fitting the model to the training data by minimizing the Mean Squared Error (MSE) between the predicted and actual values. This was achieved using gradient descent, which iteratively adjusted the model parameters (slope and intercept) to reduce the error.

Monitoring the training progress involved tracking the loss (MSE) over iterations, resulting in a loss curve that illustrated how the error decreased as the model learned from the data. The loss curve helped us ensure that the model was converging and learning effectively. Additionally, we evaluated the model's performance using several metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R-squared. These metrics provided a comprehensive understanding of the model's accuracy and its ability to generalise to new, unseen data. The results are shown in *Table 2*.

4 Results and Discussion

The SBERT-based recommender model, fine-tuned for course recommendations, provided highly relevant and personalized suggestions based on student profiles. The model had an accuracy of 0.96, demonstrating its ability to provide relevant course options based on input. Compared to an XLNet (Yang et al., 2019) model trained on the same dataset that achieved an accuracy of 0.89, the recommender model developed by this study shows improved performance. This can be attributed to SBERT's ability to generate semantically meaningful sentence embeddings (Reimers and Gurevych, 2019) from the input on student interests and academic capacity that were in the form of descriptive sentences.

The LR-based predictor model achieved a test accuracy of 0.97, indicating strong predictive power for determining students' likelihood of passing. We compared the predictor model to a Random Forest (Breiman, 2001) (RF) model trained on the same dataset. The RF model achieved an R-squared score of 0.86 upon evaluation, proving better generalisability of the PROACTED model developed by this study. Table 2 shows performance, with high accuracy and low error metrics, demonstrating the model's reliability in predicting student outcomes. Figure 3 shows sample output of the prediction model.

Table 1: Recommender Model Performance

Metric	Accuracy	Precision	Recall	F-Score
Score	0.96	0.81	0.97	0.90

Table 2: Prediction Model Performance

Metric	MSE	RMSE	MAE	R^2
Score	0.25	0.50	0.50	0.97

5 Ethical Considerations

Student privacy protection through data de-identification, secure storage of data and transparency in data handling are some of the ethical data-driven (Cerratto Pargman and McGrath, 2021) measures that were taken to ensure ethical model development (Kitto and Knight, 2019). These strategies were essential in not only protecting student privacy but also promoting the trustworthiness (Mathrani et al., 2021) of the PROACTED model.

6 Conclusion and Future Work

The course recommendations from the recommender model are highly relevant and tailored to the students' academic profiles and interests. These recommendations support the study's goal of guiding students towards courses that match their skills and academic capabilities, thus enhancing their academic success and satisfaction.

The predicted probabilities align well with the study's goal of performance evaluation, showing the probability of a student passing at any point thus enabling targeted interventions. The model's performance metrics suggest robustness and reliability in a practical educational setting.

We, however, recognise that certain factors that affect student performance such as psychological, financial and certain social that concern mental health, fee and societal support that are often not quantifiable might also impact student performance. Research into these factors and ways of including them in the model, though out of the scope of this study, is paramount in representing the real world and will form part of future work.

Further work will also involve integration with well anonymized, real-world data to further enhance model accuracy and relevance. This data will be ethically collected in partnership with institutions of higher learning with informed consent from relevant stakeholders such as the students and institution administration to ensure fairness and transparency of the resulting model. Additionally, the implementation of a feedback mechanism where students can rate the relevance of recommendations will enable continuous model improvement based on real user experiences.

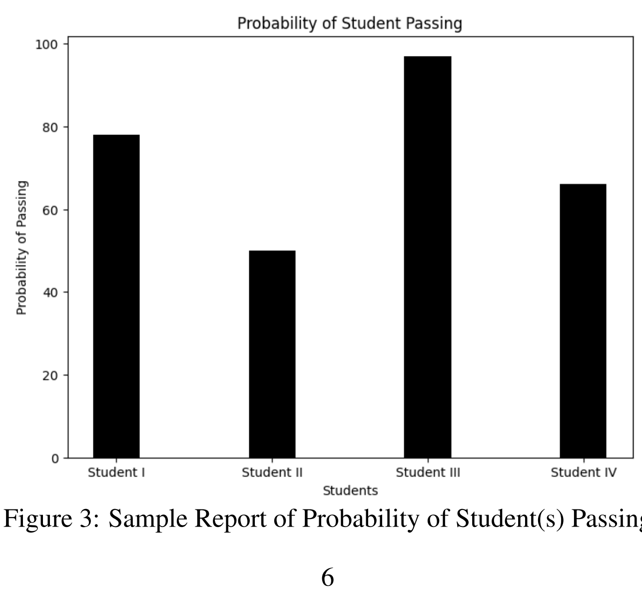

Figure 3: Sample Report of Probability of Student(s) Passing

7 Acknowledgement

We acknowledge the support of the students and faculty members of the School of Computing and Informatics of Maseno University through initiatives like ICT innovation forums which foster innovation. We also thank Maseno Centre for Applied Artificial Intelligence (MCAAI) and Dr. Lilian Wanzare for guidance.

References

Afsar, M.M., Crump, T., Far, B., 2022. Reinforcement learning based recommender systems: A survey. ACM Computing Surveys 55, 1–38.

Ahmed, A., Elaraby, I.S., 2014. Data mining: A prediction for student's performance using classification method. World Journal of Computer Application and Technology 2, 43–47.

Albreiki, B., Zaki, N., Alashwal, H., 2021. A systematic literature review of student'performance prediction using machine learning techniques. Education Sciences 11, 552.

Aleem, A., Gore, M.M., 2020. Educational data mining methods: A survey, in: 2020 ieee 9th international conference on communication systems and network technologies (csnt), IEEE. pp. 182–188.

Ashraf, E., Manickam, S., Karuppayah, S., 2021. A comprehensive review of course recommender systems in e-learning. Journal of Educators Online 18.

Bank, W., 2019. Improving higher education performance in Kenya: A policy report. World Bank.

Bird, K.A., Castleman, B.L., Mabel, Z., Song, Y., 2021. Bringing transparency to predictive analytics: A systematic comparison of predictive modeling methods in higher education. AERA Open 7, 23328584211037630.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Bursać, M., Blagojević, M., Milošević, D., 2019. Early prediction of student success based on data mining and artificial neural network, in: International Conference on Human Centered Computing, Springer. pp. 26–31.

Cardona, T., Cudney, E.A., Hoerl, R., Snyder, J., 2023. Data mining and machine learning retention models in higher education. Journal of College Student Retention: Research, Theory & Practice 25, 51–75.

Cerratto Pargman, T., McGrath, C., 2021. Mapping the ethics of learning analytics in higher education: A systematic literature review of empirical research. Journal of Learning Analytics 8, 123–139.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.

Chowdhury, S., 2022. Evaluating cold-start in recommendation systems using a hybrid model based on factorization machines and sbert embeddings.

Esteban, A., Zafra, A., Romero, C., 2020. Helping university students to choose elective courses by using a hybrid multi-criteria recommendation system with genetic optimization. Knowledge-Based Systems 194, 105385.

Hashim, A.S., Awadh, W.A., Hamoud, A.K., 2020. Student performance prediction model based on supervised machine learning algorithms, in: IOP conference series: materials science and engineering, IOP Publishing. p. 032019.

Hicham, A., Jeghal, A., Sabri, A., Tairi, H., 2020. A survey on educational data mining [2014-2019], in: 2020 international conference on intelligent systems and computer vision (ISCV), IEEE. pp. 1–6.

Hooda, M., Rana, C., Dahiya, O., Shet, J.P., Singh, B.K., 2022. Integrating la and edm for improving students success in higher education using fcn algorithm. Mathematical Problems in Engineering 2022, 7690103.

Kitto, K., Knight, S., 2019. Practical ethics for building learning analytics. British Journal of Educational Technology 50, 2855–2870.

Kovalev, S., Kolodenkova, A., Muntyan, E., 2020. Educational data mining: current problems and solutions, in: 2020 V International Conference on Information Technologies in Engineering Education (Inforino), IEEE. pp. 1–5.

López Zambrano, J., Lara Torralbo, J.A., Romero Morales, C., et al., 2021. Early prediction of student learning performance through data mining: A systematic review. Psicothema .

Mathrani, A., Susnjak, T., Ramaswami, G., Barczak, A., 2021. Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. Computers and Education Open 2, 100060.

Miguéis, V.L., Freitas, A., Garcia, P.J., Silva, A., 2018. Early segmentation of students according to their academic performance: A predictive modelling approach. Decision Support Systems 115, 36–51.

Mubarak, A.A., Cao, H., Hezam, I.M., 2021. Deep analytic model for student dropout prediction in massive open online courses. Computers & Electrical Engineering 93, 107271.

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., Phasinam, K., 2023. Classification and prediction of student performance data using various machine learning algorithms. Materials today: proceedings 80, 3782–3785.

Pardos, Z.A., Jiang, W., 2020. Designing for serendipity in a university course recommendation system, in: Proceedings of the tenth international conference on learning analytics & knowledge, pp. 350–359.

Priyatno, A.M., Widiyaningtyas, T., et al., 2024. A systematic literature review: Recursive feature elimination algorithms. JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer) 9, 196–207.

Reimers, N., Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 .

Romero, C., Ventura, S., 2010. Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews) 40, 601–618.

da Silva, F.L., Slodkowski, B.K., da Silva, K.K.A., Cazella, S.C., 2023. A systematic literature review on educational recommender systems for teaching and learning: research trends, limitations and opportunities. Education and Information Technologies 28, 3289–3328.

Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., Ragos, O., 2020. Transfer learning from deep neural networks for predicting student performance. Applied Sciences 10, 2145.

Wang, Z., Zhu, C., Ying, Z., Zhang, Y., Wang, B., Jin, X., Yang, H., 2018. Design and implementation of early warning system based on educational big data, in: 2018 5th International Conference on Systems and Informatics (icsai), IEEE. pp. 549–553.

World Bank, 2021. One africa: Technology, education, and covid-19. URL: `https://thedocs.worldbank.org/en/doc/908af3404023a2c31ef34853bba4fe60-0200022022/original/One-Africa-TE-and-COVID-19-11102021.pdf`. accessed: date-of-access.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems 32.

Zeineddine, H., Braendle, U., Farah, A., 2021. Enhancing prediction of student success: Automated machine learning approach. Computers & Electrical Engineering 89, 106903.