Interpreting Ensemble Machine Learning Models Local and Global Predictions by Leveraging Explainable Artificial Intelligence

By

Muhammad Karim

Registration No: CS320211002

Supervisor-I: Dr. M. Irfan Uddin — Institute of Computing, KUST, Kohat — Signature

Supervisor-II: Dr. Muhammad Adnan — Institute of Computing, KUST, Kohat — Signature

Director: Dr. Amjad Mehmood — Institute of Computing, KUST, Kohat — Signature

Institute of Computing
Kohat University of Science and Technology, Kohat-26000
Khyber Pakhtunkhwa, Pakistan

Introduction

With the popularity of Artificial Intelligence (AI), almost every organization is trying to incorporate AI techniques to make better decisions. As more and more organizations try to embed AI to automate their business processes, there is a need for transparency while harnessing the effectiveness and efficiency of AI [1]. Traditional machine learning (ML) and deep learning (DL) models follow the black-box approach where the decision-making process is hidden from humans, thus, is not interpretable by humans [2]. Explainable Artificial Intelligence (XAI) [3], a subfield of AI, tries to explain how input features are connected with the output features and how a machine learning model makes a particular decision.

To solve complicated problems, machines and human beings must collaborate clearly and effectively. The future of AI requires good communication, clarity, trust, and understanding with the machines. EAI tries to address such questions by combining the best practices of AI and machine learning [4]. Predicting and interpreting online students' performance at various stages of course lengths is a challenging task. The earliest possible prediction and interpretation will help university stakeholders such as instructors and administrators to perform interventions at optimal times thus motivating and guiding students to keep on the right path.

In this research study, we propose the XAIPI model which will predict and interpret students' performance at various stages of course lengths i.e., at 20%, 40%, 60%, 80%, and 100%. The XAIPI model will mainly work in two phases. In phase 1, various ensemble machine learning models will be used to predict students' performance at 20%, 40%, 60%, 80%, and 100% course lengths. Each ensemble model will be evaluated by using metrics such as precision, recall, accuracy, and f-score. Moreover, Receiver Operating Curve (ROC) Area Under Curve (AUC), and Precision-Recall AUC will be used to determine the overall performance of various models. In phase 2, we will use explainable AI to interpret and understand how a model makes the final decision. Unlike the black box approach, where the internal working of the model is hidden from the user, explainable AI tools and techniques will be used to make the working of ML models transparent and interpretable to the users. We will use Local Interpretable Model-agnostic Explanations (LIME) to interpret ML model [5] prediction on a single observation whereas Shapely Additive Explanations (SHAP) values will be employed for the model's global explainability.

The anonymous dataset used in this study is of Open University, UK called Open University Learning Analytic Dataset (OULAD). This dataset consists of seven files that are connected using a unique identifier. All these seven files are stored in CSV format. The dataset contains information about courses, students, assessments, and their interactions with Virtual Learning Environment (VLE) [6]. The dataset mainly contains demographic data, assessment scores, and click-stream data of the students. Overall, the dataset contains information about 22 courses, 32,593 students, their assessment

results, and daily summaries of the students' clicks (10,655,280 entries). The students are enrolled in two sessions per year. One session starts in February which is represented by "B" and another session starts in October which is represented by the letter "J". The three types of students' assessment scores stored in the dataset are Tutor Marked assessment (TMA), Computer Marked Assessment (CMA), and Final Exam(Exam).

Literature Review

In these modern days, E-Learning, also known as Electronic Learning, has become an essential part of internet-based education [7]. During the Covid-19 pandemic [8], E-Learning has been used throughout the world due to its flexibility, high temporal, and highly effective learning resources. Although, in this type of learning process, teachers cannot easily perceive the learning status of their learners; which led us to pose questions about the quality of E-Learning. The study which assists teachers to predict learning performance gives a basis for them (the teachers) to adjust their teaching methods for students who may have difficulties by predicting students' performance on the future exam, minimizing the risk of failing to pass the course, and make sure that E-Learning is highly effective. Through a large number of experiential studies that examine the relationship between E-Learning behavior and learning performance, learners' E-Learning behavior has a significant effect on learning performance. This is why, in recent years, learning performance prediction which is based on learning process data has caught considerable attention throughout the world. The use of measurement, collection and analysis of learning process data to get learning performance prediction can help teachers modify teaching strategies in time and start during students' learning processes using the role of supervision and early warning.

Online learning, also referred to as E-Learning or virtual learning [9] is delivered through a website on the Internet where multiple applications and technical equipment are used for this purpose. Until now, E-Learning has beefed up into various offerings with the use of different types of resources, applications, and other technologies. It has popularised rapidly in a very short time.

Online courses are commonly supported by Web technologies such as Learning Management Systems (LMS) [10] which creates behoove, precise and relevant produced daily by students, teachers, and administrators. The information that we get from it provides the opportunity that analyze to and deeply examines student performance (SP) by allowing permitting the detection of students at risk of failure as well as modifying quality issues. Thus, a precise, representative, and early prediction model of SP based on data recorded by an LMS could assist in minimizing failure rates by allowing the timely implementation of corrective actions.

Online courses and e-degrees have caught attention in the last ten years/decades even then

it has been presenting since mid-1990. In addition, the Coronavirus disease compelled the world (e.g. Italy, the US, China, and other developed countries) to use e-sources for education. Universities as well as academies adopted digital technologies for teaching more broadly. But they will have to comprehend what likely ways of assessing and effectively teaching will be in this scenario. In illustrated overview, along with the utility and ever-present access to the educational platforms of online courses, details enrolments in a large number of students. This is why being able to deal with the trend shifting of student interactions with the course platforms in real-time has become of paramount importance. Student dropout definition, Input modeling, Underlying machine, and deep learning techniques, Evaluation measures, Datasets, and privacy concerns.

With the help of digital learning, environment [11], we are regulating the learning process distantly and running online courses without direct teacher-student interaction. Thus, course maintenance is restricted to keeping the content up-to-date throughout and after the course as well as providing student counseling services.

Problem Statement

Student dropouts and failure in online learning environments have become major issues [12]. Due to a lack of guidance, motivation, and timely intervention most students are not able to complete their degrees thus numerous students are dropout or cannot attain passing grades. Moreover, as there is no direct and physical interaction between the instructors and students, instructors are unaware of the strength and weaknesses of students. Additionally, current machine learning models implemented and integrated with online learning environments follow a black-box approach [13], where the internal working and decision-making processes are hidden from the instructors. Thus, the instructors are unaware of how students are classified into various performance groups by the ML models. Finally, most ML models interpret and predict students' performance at the end of the course thus weak students are at risk of dropout during the course length.

Objectives

In this study, we propose the XAIPI model to perform the earliest possible student performance prediction and interpretation.

- To predict and interpret students' performance at various percentages of course duration i.e., 20%, 40%, 60%, 80%, and 100%.
- To facilitate instructors for local and global interpretation of students' performance at different

course lengths.

- To use various evaluation metrics such as the ROC-AUC curve, and PR-curve will be used to evaluate the XAIP model performance.
- To propose various persuasion strategies that will help instructors in motivating and encouraging students to improve their performance.

Methodology

The methodology part consists of the following steps:

Data Preprocessing

To make a more accurate and efficient ML model, the OULA dataset will be preprocessed for removing nulls, incomplete values, missing data, and outliers. For example, the assessment table has a feature called date. It is important to make sure that no incomplete or missing values are present in the date feature. As our aim is to predict and interpret the ML model at various stages of course length, therefore date feature plays a key role in dividing and understanding the course into various parts. The missing and incomplete values in the date feature are replaced by the date mean value. Similarly, the date feature values in other files such as the assessment file, studentVle file, student registration file, and in vle file are replaced by proper date values. Moreover, outliers, missing data, and incomplete data in other features are deleted or replaced by their mean values. Once we make sure that we have a complete, comprehensive, and clean dataset, we will then apply feature engineering and feature merging steps to our dataset. Table 1 presents the important features that will be used for ML training and testing at different times of course length.

In the feature engineering step, we will generate more features related to assessment scores, clickstreams, and final performance as we will train the ML model at various percentages of course length. We will generate relative assessment features such as AS20, AS40, AS60, AS80, and AS100 at different course lengths. Next, we will generate late assessment features values such as LS20, LS40, LS60, LS80, and LS100 at various course length. Similarly, new features for representing students' clickstream information will be generated at various course lengths. Finally, if needed we will perform a feature merging step on the final performance score where the Fail and Withdrawn classes will be merged into one class namely the Fail class, and Pass and Distinction grades will be merged into one grade namely the Pass grade. The feature merging operation will help us in increasing the performance of the ML model.

Table 1: Features generated after features engineering process

Demographic Features	**Student ID, Gender, Immigration Band, Highest Education, Age Band, Number of previous attempts, credit hours already studied, Region, disability, Code Module Registered, Code Presentation Registered, Final Result.**
Relative Assessment Features at various length of course	AS20, AS40, AS60, AS80, AS100
Late Assessment Features score at various length of course	LS20, LS40, LS60, LS80, LS100
Total/Sum Clickstreams	SC20, SC40, SC60, SC80, SC100
Average Clickstreams	AC20, AC40, AC60, AC80, AC100
Final Performance Score	Final performance score in the form of Pass, Fail, Withdrawn, Distinction

Methods and Experiments

The two major goals of the experiment will be training, testing, and evaluating the ML model at various course lengths and interpreting ML model performance at various course lengths. During the ML model training process, we will first employ various ensemble ML algorithms such as gradient boosting, adaptive boosting, extreme gradient boosting, light gradient boosting, and stacking algorithms to determine how they perform to model the students' learning behavior at various course lengths. After training and evaluating ensemble ML models, we will also leverage neural network models such as Artificial Neural Networks (ANN) and Feed-forward Neural Networks to model the students' online learning behavior at different course lengths. To compare the performance of ensemble models with neural networks, various metrics such as the ROC-AUC curve, Precision-Recall curve, accuracy, recall, precision, and f-score will be utilized. Once it is determined that will ML model gives the best performance, that ML model will be selected for interpreting students' learning behavior at different course lengths by leveraging various explainable AI (XAI) methods and techniques. The XAI techniques will help all the stakeholders involved in the online learning process such as university administers, instructors, and students in knowing and interpreting in an easy and understandable way, how the performance of students was predicted, which are the crucial features affecting students' performance and how ML model makes the decision in predicting students grades at different course lengths.

The advantage of integrating XAI techniques with ML models is that unlike the black box approach (which is used by most ML models), the working and decision-making process is transparent to humans, and it will allow both instructors and students in trusting the ML models. Furthermore, the student's behavior will be interpreted globally and locally through various XAI tools, methods, and techniques. The global explainability will allow the instructors to understand how the ML model

is performing while predicting and classifying the performance of all the students. On the other hand, local explainability will facilitate instructors in determining how the ML model predicts the performance of the individual student. Global explainability will help instructors in providing help and guidance needed for all the students whereas local explainability will help instructors in recommending adaptive and tailored help, feedback, and guidance to the individual student which will facilitate students in remaining on the right path thus reducing the risk of student failure and dropout. We will use the following XAI tools to evaluate and interpret students' performance at different course lengths.

Our methodology consisted of the following steps. Moreover, the complete workflow of the methodology is presented in figure 1.

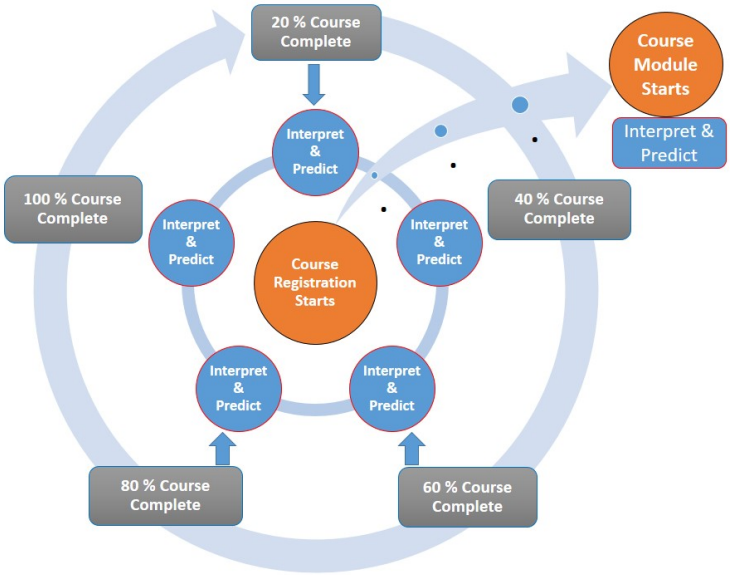

Figure 1: Complete workflow Diagram illustrating each step during the methodology steps

Time Frame

The following timeframe diagram depicts the time required to complete each synopsis phase.

S.No.	Research Components	Proposed Time
1	Literature Review	02 months
2	Ensemble AI techniques	03 months
3	Experimentation	03 months
4	Thesis writing	02 months

Table 2: Proposed study timeline

References

[1] R. van de Schoot, J. de Bruin, R. Schram, P. Zahedi, J. de Boer, F. Weijdema, B. Kramer, M. Huijts, M. Hoogerwerf, G. Ferdinands *et al.*, "An open source machine learning framework for efficient and transparent systematic reviews," *Nature Machine Intelligence*, vol. 3, no. 2, pp. 125–133, 2021.

[2] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[3] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[4] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[5] N. B. Kumarakulasinghe, T. Blomberg, J. Liu, A. S. Leao, and P. Papapetrou, "Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2020, pp. 7–12.

[6] S.-U. Hassan, H. Waheed, N. R. Aljohani, M. Ali, S. Ventura, and F. Herrera, "Virtual learning environment to predict withdrawal by leveraging deep learning," *International Journal of Intelligent Systems*, vol. 34, no. 8, pp. 1935–1952, 2019.

[7] V. Belaya, "The use of e-learning in vocational education and training (vet): Systematization of existing theoretical approaches." *Journal of Education and Learning*, vol. 7, no. 5, pp. 92–101, 2018.

[8] A. M. Maatuk, E. K. Elberkawi, S. Aljawarneh, H. Rashaideh, and H. Alharbi, "The covid-19 pandemic and e-learning: challenges and opportunities from the perspective of students and instructors," *Journal of Computing in Higher Education*, vol. 34, no. 1, pp. 21–38, 2022.

[9] R. Williams, "An academic review of virtual learning environments," *ICRRD Quality Index Research Journal*, vol. 3, no. 2, pp. 143–145, 2022.

[10] O. Yakit and R. Ismailova, "Learning management system implementation. case study on user interface configurations." *MANAS Journal of Engineering*, vol. 6, no. 2, pp. 164–176, 2018.

[11] S. C. Harris and V. Kumar, "Identifying student difficulty in a digital learning environment," in *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 2018, pp. 199–201.

[12] A. A. Mubarak, H. Cao, and W. Zhang, "Prediction of students' early dropout based on their interaction logs in online learning environment," *Interactive Learning Environments*, vol. 30, no. 8, pp. 1414–1433, 2022.

[13] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Information Fusion*, vol. 77, pp. 29–52, 2022.