3-2023

IDENTIFICATION OF STUDENTS AT RISK OF LOW PERFORMANCE BY COMBINING RULE-BASED MODELS, ENHANCED MACHINE LEARNING, AND KNOWLEDGE GRAPH TECHNIQUES

Balqis Mubarak Al Braiki
United Arab Emirates University, 200907523@uaeu.ac.ae

Recommended Citation

Al Braiki, Balqis Mubarak, "IDENTIFICATION OF STUDENTS AT RISK OF LOW PERFORMANCE BY COMBINING RULE-BASED MODELS, ENHANCED MACHINE LEARNING, AND KNOWLEDGE GRAPH TECHNIQUES" (2023). *Dissertations*. 194.
https://scholarworks.uaeu.ac.ae/all_dissertations/194

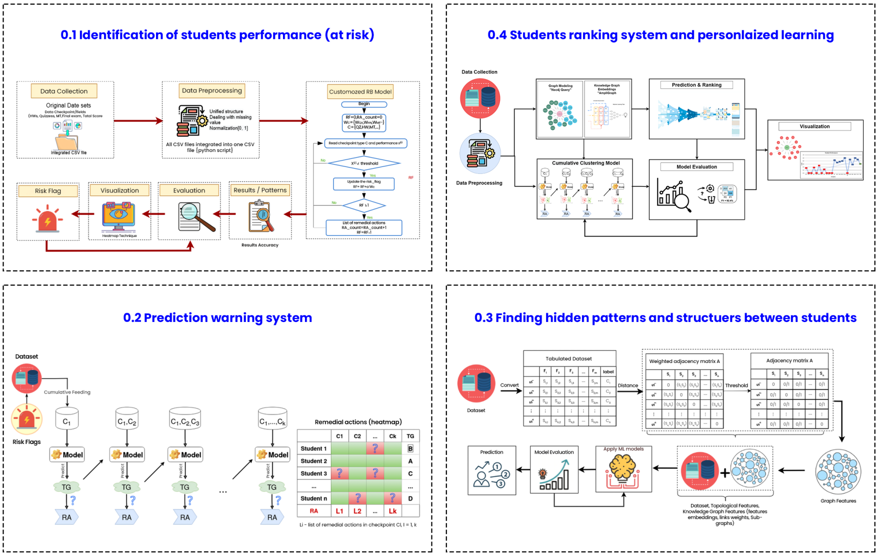
UAEU
جامعة الإمارات العربية المتحدة
United Arab Emirates University

DOCTORATE DISSERTATION NO. 2023:2

College of Information Technology

IDENTIFICATION OF STUDENTS AT RISK OF LOW PERFORMANCE BY COMBINING RULE-BASED MODELS, ENHANCED MACHINE LEARNING, AND KNOWLEDGE GRAPH TECHNIQUES

Balqis Albreiki



March 2023

United Arab Emirates University

College of Information Technology

IDENTIFICATION OF STUDENTS AT RISK OF LOW PERFORMANCE BY COMBINING RULE-BASED MODELS, ENHANCED MACHINE LEARNING, AND KNOWLEDGE GRAPH TECHNIQUES

Balqis Albreiki

This dissertation is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Informatics and Computing

March 2023

United Arab Emirates University Doctorate Dissertation

2023:2

Cover: Image regarding Information and Communication Technology in this study (Photo: By Balqis Albreiki)

Print: University Print Service, UAEU 2023

Declaration of Original Work

I, Balqis Albreiki, the undersigned, a graduate student at the United Arab Emirates University (UAEU), and the author of this dissertation, entitled *"Identification of Students at Risk of Low Performance by Combining Rule-based Models, Enhanced Machine Learning, and Knowledge Graph Techniques"*, hereby, solemnly declare that this dissertation is my own original research work that has been done and prepared by me under the supervision of Prof. Nazar Zaki, in the College of Information Technology at the UAEU. This work has not previously been presented or published, or formed the basis for the award of any academic degree, diploma or a similar title at this or any other university. Any materials borrowed from other sources (whether published or unpublished) and relied upon or included in my dissertation have been properly cited and acknowledged in accordance with appropriate academic conventions. I further declare that there is no potential conflict of interest with respect to the research, data collection, authorship, presentation and/or publication of this dissertation.

Student's Signature: ____________________

Date: 16.03.2023

Advisory Committee

1) Advisor: Nazar Zaki

Title: Professor

Department of Computer Science and Software Engineering

College of Information Technology

2) Member: Walid Ibrahim

Title: Professor

Department Computer and Network Engineering

College of Information Technology

3) Member: Taoufik Zoubeidi

Title: Professor

Department of Analytics in the Digital Era

College of Business and Economics

Approval of the Doctorate Dissertation

This Doctorate Dissertation is approved by the following Examining Committee Members:

1) Advisor (Committee Chair): Dr. Nazar Zaki
Title: Professor
Department of Computer Science and Software Engineering
College of Information Technology
Signature ________________ Date 16.03.2023

2) Member: Dr. Hany Al Ashwal
Title: Associate Professor
Department of Computer Science and Software Engineering
College of Information Technology
Signature ________________ Date 16.03.2023

3) Member: Dr. Negmeldin Al Sheikh
Title: Associate Professor
Curriculum and Methds of Instructions
College of Education
Signature ________________ Date 16.03.2023

4) Member (External Examiner): Dr. Chu Kiong Loo
Title: Professor
Department of Artificial Intelligence
University of Malaya (Malaysia)
Signature ________________ Date 16.03.2023

This Doctorate Dissertation is accepted by:

Dean of the College of Information Technology: Professor Taieb Znati

Signature Taieb Znati

Date 12/05/2023

Dean of the College of Graduate Studies: Professor Ali Al-Marzouqi

Signature Ali Hassan

Date 15/05/2023

Abstract

Technologies and online learning platforms have changed the contemporary educational paradigm, giving institutions more alternatives in a complex and competitive environment. Online learning platforms, learning-based analytics, and data mining tools are increasingly complementing and replacing traditional education techniques. However, academic underachievement, graduation delays, and student dropouts remain common problems in educational institutions. One potential method of preventing these issues is by predicting student performance through the use of institution data and advanced technologies. However, to date, scholars have yet to develop a module that can accurately predict students' academic achievement and commitment. This dissertation attempts to bridge that gap by presenting a framework that allows instructors to achieve four goals: (1) track and monitor the performance of each student on their course, (2) identify at-risk students during the earliest stages of the course progression (3), enhance the accuracy with which at-risk student performance is predicted, and (4) improve the accuracy of student ranking and development of personalized learning interventions. These goals are achieved via four objectives. Objective One proposes a rule-based strategy and risk factor flag to warn instructors about at-risk students. Objective Two classifies at-risk students using an explainable ML-based model and rule-based approach. It also offers remedial strategies for at-risk students at each checkpoint to address their weaknesses. Objective Three uses ML-based models, GCNs, and knowledge graphs to enhance the prediction results. Objective Four predicts students' ranking using ML-based models and clustering-based KGEs with the aim of developing personalized learning interventions. It is anticipated that the solution presented in this dissertation will help educational institutions identify and analyze at-risk students on a course-by-course basis and, thereby, minimize course failure rates.

Keywords: Rule-based System; Prediction; At-risk Student; Machine Learning; Knowledge Graph; Personalized Learning.

Title and Abstract (in Arabic)

تحديد الطلاب المعرضين لخطر انخفاض الأداء الأكاديمي عن طريق دمج النماذج المستندة إلى القواعد والتعلم الآلي المعزز والمعرفة بتقنيات الرسم البياني

لقد غيرت التقنيات ومنصات التعليم والتعلم عبر الإنترنت من مناهج التعليم المعاصرة مما اعطى لهذه المؤسسات التعليمية المزيد من الخيارات والبدائل وضمن بيئة معقدة وتنافسية. فقد حلت منصات التعليم عبر الإنترنت وتحليل بيانات التعلم وأدوات استخراج وتنقيب البيانات محل التقنيات التعليمية التقليدية واستبدلتها او اكملتها وبشكل متزايد وعلى أحدث وجه. ومع ذلك، يظل الفشل الأكاديمي وتأخر تخرج الطلبة وتسربهم من الدراسة من الأمور الشائعة في المؤسسات التعليمية. وتتمثل إحدى الطرق الممكنة لمنع هذه المعوقات والصعوبات من خلال إمكانية التنبؤ بأداء الطلبة الأكاديمي باستخدام بيانات المؤسسات التعليمية والتقنيات المتطورة. ومع ذلك، لم يتمكن العلماء حتى الآن من تطوير وسائل تستطيع القيام بالتنبؤ بأداء الطلبة ومتابعتهم بدقة عالية. تحاول هذه الرسالة سد هذه الفجوة عن طريق تقديم إطار يسمح للمدرسين بتحقيق أربعة أهداف: (١) تتبع ومراقبة أداء كل طالب في مساقاته الدراسية ، (٢) التعرف على الطلبة ذو المستوى الأكاديمي المتدني وتشخيصهم أثناء المراحل المبكرة من التحاقهم بالمساق الدراسي. (٣) تحسين دقة التنبؤ بأداء الطلبة ذو المستوى الأكاديمي المتدني. (٤) تحسين دقة تصنيف الطلبة وتطوير برامج تقوية تعليمية خاصة بهم تناسب احتياجاتهم. حيث أنه يتم تحقيقهم عن طريق أربعة أهداف. يقترح الهدف الأول وضع استراتيجية قائمة على القواعد عن طريق وضع مؤشرات عامل الخطر لتحذير المدرسين بشأن الطلبة المعرضون لخطورة تدني المستوى الأكاديمي. بينما يقوم الهدف الثاني بتصنيف الطلبة ذو المستوى الأكاديمي المنخفض باستخدام نموذج يعتمد على دمج نظام تعلم الآلة القابل للتنبؤ وطريقة الاعتماد على القواعد. كما يوفر خططا تصحيحية وتحسينية للطلبة ذو المستوى الأكاديمي المنخفض عند كل نقطة تدقيق لأجل معالجة نقاط ضعفهم. الهدف الثالث يستخدم نماذج التعلم الآلي والشبكة

العصبية لالتفاف الرسم البياني والرسوم البيانية المعرفية لتحسين نتائج دقة التنبؤ. اما الهدف الرابع فيقوم بالتنبؤ وتصنيف الطلبة باستخدام نماذج التعلم الآلي والرسوم البيانية المعرفية القائمة على وضع الطلبة بمجموعات بهدف اعتماد تدخلات تعليمية خاصة لكل طالب. ونأمل ان تساعد الحلول التي تطرحها هذه الاطروحة المؤسسات التعليمية على تشخيص وتحليل حالات الطلبة المعرضون للمستوى الأكاديمي المنخفض وعلى أساس كل مساق وبشكل منفرد لكي يتم تقليص معدلات الفشل الأكاديمي بين الطلبة.

مفاهيم البحث الرئيسية: تقنيه نظم القواعد، التنبؤ ، الطالب في خطر التدني الأكاديمي ، تعلم الآلة ، الرسم البياني المعرفي ، التعلم الشخصي.

List of Publications

This dissertation is based on the work presented in the following papers, referred to by Roman numerals:

I. Albreiki, B., Habuza, T., Shuqfa, Z., Serhani, M. A., Zaki, N., & Harous, S. (2021). Customized Rule-Based Model to Identify At-Risk Students and Propose Rational Remedial Actions. Big Data and Cognitive Computing, 5(4), 71. https://www.mdpi.com/2504-2289/5/4/71

II. Albreiki, B., Habuza, T., & Zaki, N. (2022). Framework for automatically suggesting remedial actions to help students at risk based on explainable ML and rule-based models. International Journal of Educational Technology in Higher Education, 19(1), 1-26.https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-022-00354-6

III. Albreiki, B., Habuza, T., & Zaki, N. (2023). Extracting topological features to identify at risk students using machine learning and graph convolutional network models. International Journal of Educational Technology in Higher Education, 20(1), 1-22. https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-023-00389-3

IV. Albreiki, B., Habuza, T., Palakkal. N.& Zaki, N. (2023). Clustering-based knowledge graphs and entity-relation representation improves the detection of at risk students. Education and Information Technologies. (Manuscript Under-review)

Author’s Contribution

The contribution of Balqis Albreiki to the papers included in this dissertation was as follows:

I. Identified and outlined the dissertation goals and developed an associated solution for the early detection and prediction of low-performing students in higher education contexts

II. Participated in planning of the work, had main responsibility for the data collection and processing, evaluation of results and manuscript writing.

III. Participated in planning of the work, had main responsibility for the experimental work, data collection and processing, evaluation of results and manuscript writing.

IV. Sole responsibility for planning the research, conducting the experiments, and preparing the manuscript.

Author Profile

Balqis Albreiki received her bachelor's degree in Information Technology and Master's degree in Engineering Management from United Arab Emirates University, Al Ain, UAE in 2014 and 2016, respectively. Her research interest is in the fields of artificial intelligence, machine learning, and predictive models. In 2018, having a strong interest in Data Science and research, she returned to academia and joined the Computer Science Ph.D. program at the UAEU. Currently, she mainly focuses on using advanced machine learning techniques to solve educational problems. She has more than five years of research experience. As a result of her current research activity, Balqis was awarded for best research work in the Chancellor Innovation Award under the Administrative category in 2021. During her PhD studies she published more than 9 articles in Q1 and Q2 journals. She has also worked in big data analytics center at UAE university.

Contact information:

Email	b.albraiki@hotmail.com
Phone	+971503399834
PO box	15551
Department	Department of CS and Software Engineering
College	College of IT, UAEU

Acknowledgements

I am deeply grateful to my supervisor, Prof. Nazar Zaki, for his unwavering support throughout my PhD studies and related research. I am grateful for his patience, motivation, and extensive knowledge, which was invaluable in guiding me through my research and completing this dissertation. I cannot imagine having a better mentor for my PhD studies. I also wish to extend my thanks to the Acting Vice Chancellor, Prof. Ghaleb Alhadrami, Prof. Ateeq Almansori, Mr. Khalfan Al Dhaheri, and Prof. Wissam Ibrahim for their support, encouragement, and providing me with a favorable environment to excel in my studies. I also thank my advisory committee members, Prof. Taoufik Zoubeidi and Prof. Walid Ibrahim, and all the faculty members of the College of Information Technology at UAE University for their support during my studies. This work would not have been possible without the generous support of the UAE University. I am grateful to my PhD colleagues and friends, especially Tetiana Habuza, for their support, discussions, and suggestions. It was an honor to be a part of our PhD cohort. Finally, I would like to acknowledge my family, especially my wonderful parents, sisters, and brother, for their unwavering belief in me. Their support has kept my spirits high throughout this journey. I would also like to thank my friends, particularly Mariam Alshamsi and Hessa Alshamsi, for their continuous support and encouragement. I could not have accomplished this work without their emotional and intellectual support. Lastly, I would like to thank everyone who has been there to support me academically and emotionally throughout my coursework.

Dedication

To my great parents, my lovely family and friends

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Acc	Accuracy
AI	Artificial Intelligence
AUC	Area Under the Curve
BAC	Balanced Accuracy
CI	Confidence Interval
CNN	Convolutional Neural Network
CSV	Comma-Separated Value
DL	Deep Learning
DS	Dataset Features
FPR	False Positive Rate
GCN	Graph Convolution Network
GE	Graph embeddings
GF	Graph topological Features
HW	Homework assignment
ITS	Intelligent Tutoring Systems
LR	Linear Regression
MAE	Mean Absolute Error
ML	Machine Learning
MOOC	Massive Open Online Courses
MT	Mid-Term exam
Qz	Quiz
ROC	Receiver Operating Characteristic Curve
RF	Random Forest
Sens	Sensitivity
Spec	Specificity
SVM	Support Vector Machine
TG	Total Grade
TPR	True Positive Rate
XGB	eXtreme Gradient Boosting

Chapter 1: Introduction

Recent technological developments have fundamentally changed contemporary life. One context in which changes have been particularly evident is education [1–3]. The advanced technologies applied in the education field have resulted in the development of several online learning platforms, such as Tutee, Intelligent Tutor, and Learning Partner [4]. Despite the recent developments in various technology-assisted educational platforms, higher education institutions continue to experience increased rates of poor performance among the student population [5]. According to findings presented in a ThinkImpact report, in college dropout rates average 40% for undergraduate students in United States of America (USA) in 2020 [6] (see Figure 1.1). Furthermore, only 41% of college students graduate in four years or less, however 44% of students who enroll in a four-year college program succeed in finishing their studies within the first six years. Similarly, the Organization for Economic Co-operation and Development (OECD) countries have reported an average student dropout rate of approximately 33% for undergraduate students in 2022 [7]. Significant low student performance leads to economic loss, loss of human capital, and loss of potential contributions to society. Therefore, it is crucial for higher education institutions to address drop-out rates and identify ways to improve student retention and performance.

The United Arab Emirates (UAE) Ministry of Education's (MOE's) 2020 strategy[1] comprises a series of five-year plans that are aimed at enhancing the quality of education offered in the UAE with a specific

[1] https://www.moe.gov.ae/Arabic/Docs/MOE%20_Strategy.pdf

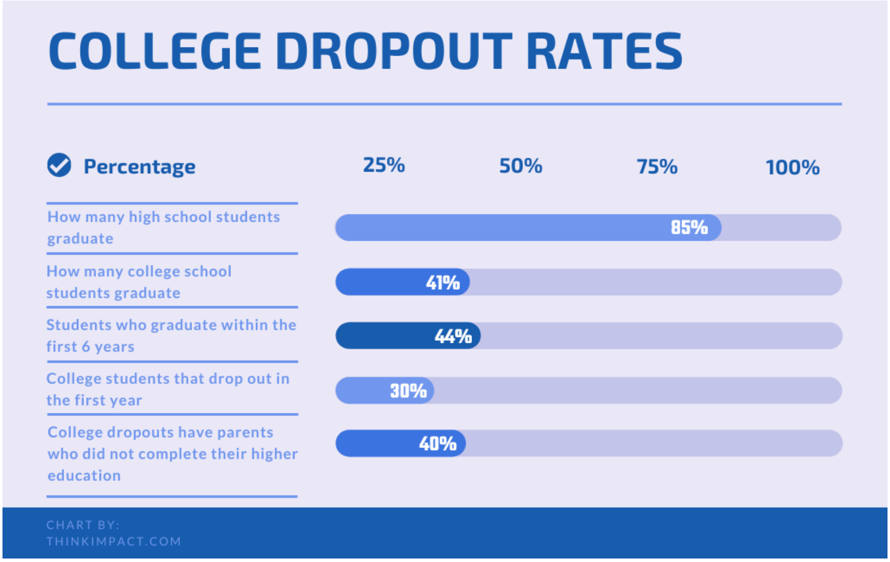


Figure 1.1: College dropout rates

emphasis on the different teaching strategies and learning capabilities of teachers and students respectively [8]. One fundamental aspect of these plans involves addressing and mitigating students' dropout rates and improving student performance. Both these variables have a considerable impact on educational institutions' reputations, the country's economy, and the overall development of society. Therefore, it is vital for the MOE to implement effective strategies and interventions to enhance student retention and performance. It is anticipated that research in this area can provide valuable insights that enable the MOE to better understand the factors that contribute to poor performance and high dropout rates and to develop targeted interventions to address these issues. This thesis focuses on the identification and strategic management of students who are predicted to be at risk of failing. The remainder of this thesis is organized as follows. Chapter 1 introduces the topic, followed by the problem statement and research objectives. Furthermore, an extensive literature review is provided to generate meaningful insights and identify research gaps. Chapter 2 defines the methods and results. The strengths and limitations of the research are outlined in Chapter 3. Chapter 4 presents the conclusion along with recommendations for future research directions.

1.1 Problem Statement

Accurately predicting student academic performance is complex, necessitating the development of novel approaches that take into account the evolving factors and conditions that influence student academic performance. It is possible that the impact of these factors and circumstances may differ from one batch of students, program, and educational institution to another. One major challenge that higher education institutions are currently encountering is high rates of student failure. However, it is difficult to determine the different variables that affect student performance. In recent years, extensive research efforts have been invested in assessing and predicting student performance in educational institutions. Learning analytics explore student data to investigate their activities and behaviors and provide relevant recommendations [9, 10]. The different learning-based models that are available range from simple statistical techniques and educational data mining methods to advanced machine learning algorithms [11, 12]. The various statistical methods that have been used to examine students' academic performance may include correlation, regression, structural equation modelling (SEM), and analysis of variance (ANOVA) [13]. Different ML-based approaches have also been leveraged in more recent times to predict the academic performance of students. Some of the different ML-based models utilized for predicting future academic performance of students include regression models [14], decision trees [15], collaborative filtering [16], support vector machine [17], and artificial neural networks [18–21]. However, most of these techniques suffer from low prediction accuracy [12, 17].

Student performance can be measured in a variety of ways. Most academics consider a range of factors when assessing performance, from final grades to students' job prospects. In the contemporary era of digitization and Internet-based learning, many educational institutions use online learning systems and courses, such as Intelligent Tutoring Systems (ITS), Learning Management Systems (LMS), and Massive Open Online Courses (MOOCs). In addition, many automated systems have been developed in different higher education institutions for the purpose of identifying students who are at risk of failing or dropping out of a course of study [11, 18]. These systems store significant amounts of students' data that can potentially be used to understand the factors contributing to students' dropout rates, academic underachievement, and graduation delays, among other issues [12, 22]. However, these capabilities have yet to be fully leveraged within automated solutions that can effectively analyze student performance and identify the factors contributing to students' underachievement and high dropout rates. Many of the aforementioned studies have been unable to clearly delineate the factors that have a direct impact on performance. This can not be achieved without solving prevailing issues associated with data quality, quantity, and complexity within ML models.

Although the recently advanced ML algorithms can be regarded as quite accurate, they are frequently considered black box models. Explaining the reasons behind a particular decision is crucial when using a model for decision-making purposes. Understanding the impact and significance of different features is essential to boosting the accuracy of model predictions. To achieve this, a model that can be understood by non-experts, including teachers and instructors, is required. This will allow teaching strategies to be adjusted in accordance with the capacities of the students. The current works are also lacking in terms of system performance (measured using different metrics, e.g., accuracy) when predicting students' academic performance [23]. There is a lack of literature describing the prediction of student performance in the early stages of a course due to the many challenges associated with this task. Interventions to address the problem of low academic performance rely on the development of early-warning detection systems that can accurately predict student performance during the learning process [24, 25]. However, researchers have yet to develop an effective solution that can flag at-risk students (i.e., students at risk of failing or leaving a course) at an early stage so that educators can devise adequate remedial actions and strategies to improve student performance. Predicting students' outcomes at an early stage of study can help academic institutions select students for grants and scholarships [1]. It can also help students avoid academic failures and promote retention in educational institutions. Additionally, the early identification of at-risk students can allow for the implementation of targeted interventions, such as additional support and resources, to improve the student's chances of success. To summarize the problem statement, there is a pressing need to establish a concrete system using robust and state-of-the-art technology within the data-driven domain that can effectively and accurately identify which students within a given cohort are at risk of failing.

1.2 Research Objectives

The work in this dissertation addresses the challenges and factors associated with predicting student performance during the early stages of their courses. Fortunately, the massive data stored in the current learning platforms can facilitate a regular analysis of students' learning and performance. The analytical tools available within contemporary platforms can help identify at-risk students and facilitate the development of different remediation strategies to improve their performance and avoid failure in the short and long term. Consequently, these systems can provide considerable benefits to students, educational institutions, instructors, and societies at-large. This research aims to address the challenges described above through an exploration of four frameworks:

- A rule-based model that can classify at-risk students (i.e., students at risk of failure) with the underlying objective of developing effective remediation strategies.
- An ML framework that can predict student performance and inform the development of effective remediation strategies by identifying the different factors that lead to students being classified as at-risk.
- A knowledge graph to extract meaningful attributes (topological graph features) for the classification of at-risk students.
- A framework that can leverage the insights generated by the knowledge graph to inform the development of personalized learning

plans to address the needs of at-risk students.

The following objectives are proposed in this dissertation:

- Objective One: To develop a customized rule-based warning system that accurately identifies the performance of at-risk students. The system will take the various factors that may contribute to a student's poor performance into account and identify those who are at risk of failure.
- Objective Two: To develop an explainable ML-based system capable of predicting the performance of students during the early stages of their course and identifying factors influencing the student performance with the underlying aim of predicting which students are at risk of failure and subsequently planning and implementing the appropriate remedial action for at-risk students.
- Objective Three: To improve the accuracy of the proposed performance prediction system by combining features extracted using knowledge graph-based methods with original data features. This will involve the use of graph-based techniques to extract features from the data.
- Objective Four: To propose an improved student performance ranking system and provide personalized learning methods based on knowledge graphs. This will involve the use of the system developed via the previous objectives to rank students based on their performance and subsequently provide personalized learning methods for each student based on their individual performance and capabilities.

1.3 Literature Review

Many studies have developed solutions to predict student performance based on the data available on learning platforms. However, due to the education paradigm shift and transformation in the education landscape that has been observed in recent years, early assessment and prediction of student performance have become critical for educational institutes [22]. The lack of intelligent automated systems for predicting student performance has resulted in many institutions experiencing high student failure and dropout rates. Several studies have proposed predictors for assessing student performance based on students' grades and course assessments. Following the advancement in e-learning platforms, several other indicators have also been used to analyze student performance. Various techniques have been developed that leverage different features for analyzing and predicting student performance. These frameworks will be examined in the following sub-sections.

1.3.1 ML-based Student Performance Prediction

Several ML-based studies have been conducted to predict student performance [21, 26–35]. Some researchers have proposed ensemble or hybrid approaches based on multiple ML techniques for predicting student performance using different datasets [36–40]. For example, Kuzilek et al. [41] explored student activities using General Unary Hypotheses Automation (GUHA) and Markov Chain Model (MCM) in Virtual Learning Environment (VLE) frameworks. Students' assignment marks and VLE activity logs under 13 scenarios were used as features for the ML models. The findings revealed that GUHA and Markov chain methods could effectively predict student performance. He et al. [42] examined the identification of at-risk students in MOOCs using two transfer learning algorithms, sequentially Smoothed Logistic Regression (LR-SEQ) and Simultaneously Smoothed Logistic Regression (LR-SIM), based on two different datasets (DisOpt1 and DisOpt2). They concluded that LR-SIM outperformed the LR-SEQ in AUC and the baseline Logistic Regression (LR) algorithm.

Osmanbegovic et al. [26] analyzed Naive Bayes (NB), Decision Trees (DT), and Multilayer perception (MLP) algorithms to predict students' success based on data collected from the University of Tuzla between 2010 and 2011. The proposed frameworks were evaluated based on accuracy, learning time, and error rate [26]. The NB model achieved a high accuracy score of 76.65% with a training time of less than 1 s; however, it also exhibited high error rates [26]. Similarly, Marquez-Vera et al. [28] analyzed the performance of Jrip, NNge, OneR, Prison, Ridor, ADTree, J48, Random Tree, REPTree, and Simple CART, and three variants of the genetic algorithm. Lakkaraju et al. [33] used Support Vector Machine (SVM), Random Forest, Logistic Regression, Adaboost, and Decision Tree to identify at-risk students. The different algorithms were evaluated using precision, recall, accuracy, and AUC for binary classification [33]. These studies highlight how different ML-based algorithms have been used to predict student performance with varying degrees of success.

Different features have been used in the extant literature to predict student performance. For example, Kovacic [43] analyzed the early prediction of student success using machine learning techniques based on socio-demographic features (e.g., gender, education, disability, work status, etc.) in combination with course characteristics (e.g., course block, course program, etc.). Craige et al. [44] found that Student Evaluation of Teaching Effectiveness (SETE) represented a general and effective indicator for student learning on the online platform based on results using different statistical approaches, Nearest Neighbors, and Bayesian approaches. Some researchers have used students' demographic features, e-learning system logs, academic data, and admission information to predict student performance [32]. Hu et al. [29] used time-dependent variables to develop an early warning system for predicting student success in online learning. In contrast, Aggarwal et al. [45] applied eight different ML algorithms using different academic and non-academic features (such as demographic information) to effectively predict students' academic performance. Acharya et al. [31] suggested an early prediction system using an ML-based classification algorithm that leveraged embedded feature selection approaches to reduce the size of the feature set.

Some researchers have attempted to extract rules from students' datasets to predict their performance. For example, Baradwaj and Pal [27] investigated the accuracy of decision trees to extract valuable rules from the dataset. Many researchers have analyzed the MOOCs platform for early detection of potential student dropouts [46–49]. For instance, Wang et al. [46] highlighted the significance of temporal features for effectively predicting student dropout rates. They used data from quiz scores and material received from discussion forums via the Canvas API to create temporal features that reflected student performance [46]. The features used in their study included dropout week, number of forum views, number of discussion posts, number of module views, number of quiz views, social network degree, and active days [46]. The two classification approaches employed in this study were General Bayesian Network (GBN) and Decision Tree (DT) [46].

Nagrecha et al. [48] investigated the use of deep learning models that could automatically extract features from raw MOOC data. They introduced ConRec Network, a deep learning framework that integrated Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN) to accomplish feature extraction at the pooling layer. The performance of ConRec was measured using accuracy, recall, F-score, and AUC values, such that all the reported metrics demonstrated high overall performance. Liang et al. [49] used data from students' learning activities to estimate the likelihood of a student dropping out. Data was gathered from the XuetangX platform. The proposed framework was comprised of data pre-processing, feature extraction, feature selection, and machine learning algorithms [49]. The XuetangX online learning dataset included 39 Open eDX courses with data spanning 40 days of students' activity records [49]. The log data was pre-processed before being utilized for training machine learning algorithms [49]. A total of 122 features were retrieved and divided into three groups: user features, course features, and enrollment features. The dataset was split into training and testing sets, each with 120,054 and 80,360 occurrences. Gradient Boosting (GBT), Support Vector Machines (SVM), and Random Forest (RF) classifiers were employed, with GBT demonstrating the highest average AUC score [49].

Some researchers have developed hybrid and ensemble-based approaches for predicting student performance. For example, Al-Obeidat et al. [39] proposed a hybrid approach by combining PROAFTAN, a multi-criteria classifier, with a Decision Trees (DT) classifier. The proposed algorithm worked in three stages. First, the C4.5 algorithm was applied to the dataset with discretization. Second, the data was filtered and pre-processed. Third, the data was enhanced using C4.5 with PROAFTAN with attribute selection and discretization. The researchers used the same UCI dataset as that employed in [50]. Similarly, Kotsiantis et al. [37] proposed the combinational incremental ensemble of classifiers for student performance prediction. They combined three classifiers for calculating the prediction output using a voting method [37].

In a recent study, Niyogisubizo et al. [51] proposed a novel stacking ensemble based on a hybrid of Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB), and Feed-forward Neural Networks (FNN) to predict students' dropout from university classes. The proposed method was evaluated using a dataset containing university students' data from 2016 to 2020, and the outcomes revealed that the solution was more effective than the base models. However, it should be noted that this study did not present solutions for remedial action for at-risk students.

Many researchers have focused on determining the factors that contribute to student performance using Explainable ML and DL models. For example, Chen et al. [52] developed an intelligent framework for explainable student performance prediction (ESPP). They used a time-series weekly student activity data set and addressed the issues of imbalanced data within a virtual learning environment through the use of a hybrid data sampling method. Different spatial and temporal features were used to train and validate the convolutional neural network (CNN) and long short-term memory (LSTM) for developing an early prediction DL model [52]. The results were validated against baseline algorithms, LSTM, support vector machine (SVM), and logistic regression (LR) models. Guleria and Sood [53] developed a framework that integrated the features of machine learning (ML) and explainable AI (XAI) to analyze the educational factors that had an impact on student's career placements and growth. The study leveraged ML-based white and black box models to analyze the educational dataset comprising students' academic records and employability attributes that were deemed important for job placements and skill development. They validated and compared the performance of the proposed framework by combining Naive Bayes with other approaches (e.g., Logistic Regression, Decision Tree, SVM, KNN, and Ensemble models) to achieve Recall and F-Measure values of 91.2% and 90.7%, respectively. It is evident from the extant literature that various machine-learning techniques have been leveraged to predict at-risk students [54]. These studies have demonstrated the effectiveness of utilizing both static and dynamic data in predicting at-risk students. However, there remains an opportunity for further advancement in the field with regard to identifying the underlying factors that contribute to poor student performance and implementing remedial solutions to provide timely feedback to students, instructors, and educators.

1.3.2 Enhancing ML-based Student Performance Prediction

The proliferation of online learning platforms has led to the generation of vast quantities of educational data. Analysis of this data has the potential to facilitate reductions in student dropout ratios [55], enhancements to the learning platforms [56], and tracking of students' academic performance. Several endeavors have been undertaken in this field, including the development and implementation of frameworks at higher education institutions [57], spanning functionality such as student behavior prediction [58], user intention analysis system [59], at-risk student prediction system [60], and knowledge tracing [61]. In recent years, machine learning and data mining techniques have been utilized with great success in predicting student performance in higher education [62]. These techniques are instrumental in identifying trends related to student performance and the teaching-learning process. As noted by [62], no single technique can meet all the requirements of an educational solution, particularly with regard to predicting students' academic performance. However, the integration of technologies with existing e-learning platforms can assist students, instructors, and institutions in assessing student performance and identifying at-risk students.

Yadav et al. [63] analyzed machine learning-based predictive models for students' retention assessment. The study concluded that decision trees could effectively generate interpretable output related to students' retention in educational institutions [63]. Experimental results demonstrate the effectiveness of machine learning-based predictive models for predicting student retention, identifying at-risk students with considerable accuracy, and reducing student dropout rates [63].

Kolo et al. [64] analyzed the performance of four classification algorithms—J48, SimpleCart, Rep tree, neural networks, and decision trees—for assessing students' academic performance. Their findings revealed that neural networks provided better precision than other machine learning methods [64]. The authors proposed the use of multiple machine-learning techniques to complement the weaknesses of individual frameworks in predicting students' academic performance [64].

Similarly, Dhanalakshmi et al. [65] used fuzzy genetic algorithms and decision trees to predict student academic performance on master- and bachelor-level courses. To avoid failures in the final assessment, they focused on identifying the students requiring special attention [65]. The outcomes demonstrated that decision trees could identify students at risk of failing class, while genetic algorithms more effectively identified students who were more likely to pass [65]. The researchers followed a pessimistic approach using decision trees and an optimistic approach using genetic algorithms for classifying students into safe and at-risk states [65].

Mesaric and Šebalj [66] designed a model for classifying student performance into two groups and identifying different success factors in the first academic year. To develop the system, they focused on student information collected from high school courses after the completion of the first year [66]. The current machine learning-based approaches demonstrated a positive relationship between student engagement and academic performance [67, 68]. However, these techniques were found to be limited in the context of online learning platforms due to the complexities caused by the massive data and different features. Consequently, the existing research approaches were not validated for large-scale online learning environments.

Several researchers have applied deep learning methods for predicting student performance and dropout rates in combination with shallow machine learning methods using handcrafted features [69]. For example, Fei et al. [70] used the LSTM model to extract relevant features based on students' interaction with questionnaires, video lectures, and problems. Whitehill et al. [71] applied a fully connected neural network to predict students' dropout rates. Feng et al. [72] extracted context-aware features and used this information for learning student behaviour based upon an attention-based mechanism.

Graph-based techniques have been found to be promising solutions for effectively analyzing the massive amount of data available within online learning platforms in terms of nodes (data set variables) and edges (similarity between different nodes) [73]. The researchers integrated the knowledge graph with the deep learning method in two different ways. First, they integrated the extracted semantic information with deep learning and machine learning-based approaches. Second, they used a knowledge graph with a deep learning method to optimize constrained objectives for the purpose of training deep learning models [74]. Gaur et al. [75] suggested the use of a graph convolution network (GCN) within a semi-supervised learning framework to classify large-scale data from online learning platforms. This study attempted to classify the student learning style using a heterogeneous knowledge graph and predict the performance of students enrolled in online courses [75]. Their model eliminated the gradient vanishing problem by using direct mapping between different layers [75]. This study also leveraged a heterogeneous knowledge graph to determine the complex association between students and variables within the learning environment [75]. The identified engagement levels were classified into normal, high, potential at-risk, and at-risk [75]. This study also developed a labelled dataset corresponding to different engagement levels that can be used as a benchmark dataset for future research [75].

Knowledge graphs have also been successfully used in MOOCs [76, 77], in different education-related domains, teaching and classroom resources, education management, and educational technologies. The K12EduKG system was developed based on knowledge graphs using K-12 educational subjects [78]. The developers of K12EduKG identified knowledge concepts and associations based on the probabilistic association rule and Conditional Random Field (CRF) model. Su and Zhang [79] suggested a knowledge graph-based method for accommodating large-scale educational data.

1.3.3 Enhancing Student Performance Using Remedial Actions

Most educational institutions have developed technology-assisted platforms to track and analyze student performance. Technology-assisted platforms enable the formulation of mechanisms for evaluating student performance and identifying their future requirements. However, very few research efforts have focused on the identification of remedial actions based on historical data for improving the performance of poorly performing students, reducing failures, and reducing dropout rates. This section will examine some of the research efforts that have been observed in this area.

Alghamdi et al. [80] proposed a remedial approach for supporting at-risk students. They conducted semi-structured interviews for three screening issues associated with identifying underperforming students: remedial session, procedure, and student attitudes [80]. The study demonstrated that lack of time with the teachers and issues within the logistical structure affected the effectiveness of remedial sessions [80]. The lack of case studies was deemed to be one major cause of poor-performing students [80]. They concluded that remedial support benefits from the existence of a supportive environment and simplified instructions for the students that can be easily addressed in the classroom context [80].

The sub-sections presented above show that there is a general lack of studies that provide appropriate remediation strategies targeting at-risk students. There are no existing mechanisms in place that can provide an objective assessment of the proposed remediation strategies. It is, for this reason, Objective Two in this dissertation involves the development of effective remediation strategies that can provide useful and effective guidance to at-risk students with the underlying intention of preventing failure and drop-out

1.3.4 Enhancing Students Performance Using Personalized Learning

The concept of personalized learning can be traced back to the works of Fred Keller in the 1960s with the introduction of the Personalized System of Instruction (PSI); an individualized learning method that focused on enhancing students' proficiency [81]. This mode of instruction was further refined within the Adaptive Learning Environment Model proposed by Margaret Wang in the 1990s. This framework promoted the adaptability of instruction mode to target the individual needs and requirements of students in diverse classrooms [81]. In the wake of these foundational frameworks for modifying instructional methods to enhance learning in students, personalized learning has been defined as a mode of learning that leverages individualization (adjusting the pace of instruction to match the diverging learning rate of students) and differentiation (modifying approach to instruction delivery) to allow students and teachers a degree of freedom with respect to the time, place, subject matter and the strategies, tools and technologies used to enable the learning process [82]. Ultimately, the goal of personalized learning is to effectively mold the educational experiences of individual students with respect to their diverse needs, talents and interests [83]. Some studies have provided personalized education to students in line with their academic capabilities and shortcomings [84, 85]. Technology-assisted learning platforms provide considerable opportunities to analyze and leverage students' data for the purpose of offering personalized learning to prevent failure. Educators can plan personalized countermeasures for different students based on their performance and shortcoming [86, 87]. Personalized learning methods include social promotions and interactive learning using game-based activities [88, 89].

Rodriguez-Segura [90] reviewed prominent experimental work on the provision of personalized learning within the context of low- and middle-income countries. They concluded that appropriate interventions could improve student performance and the learning process, leading to effective learning outcomes.

Similarly, Escueta et al. [91] assessed the effectiveness of computer-assisted learning in low and middle-income countries. They concluded that computer-assisted learning could be adapted according to each individual student's academic needs and requirements [91]. Xie et al. [92] also reviewed personalized learning on a holistic level and analyzed global developments and trends in technology-assisted learning platforms in contrast to conventional learning methods. Zhang et al. [93] summarized 71 research studies specific to technology-assisted personalized learning. The authors highlighted that the use of digital platforms led to positive results in terms of academic outcomes, students' attitudes, and engagement [93].

As the above discussion highlights, the potential benefits of using computer-assisted personalized learning platforms to improve student performance are substantial. Despite this, few studies have specifically explored their use as a means of targeting at-risk students. This emphasizes the need for more research in this field to determine the effectiveness of personalized learning as a solution to address the unique needs of at-risk students and, thereby, decrease dropout rates.

1.3.5 Literature Summary

The various studies discussed above indicate that the use of machine learning and deep learning methods combined with the availability of a large amount of student data from technology-assisted educational platforms has proven to be an accurate means of predicting students' academic performance and identifying at-risk students. This suggests that the incorporation of these techniques and data into educational systems has the potential to improve the identification and support of at-risk students, ultimately reducing dropout rates.

Table 1.1 summarizes the ML-based student performance prediction studies described above. As can be observed in Table 1.1, DT, LR, NB, MLP, and SVM are the most commonly used techniques for predicting student performance. Most researchers used a single technique, and only a few studies considered ensemble or hybrid approaches, even though such methods have achieved promising results in predicting student performance.

Table 1.1: Literature review summary

Ref	Year	Approach	Technique	RBM	BS	ML	XML	KG	RA
[43]	2010	Early prediction of student success	CART	✓		✓			
[37]	2010	Predicting students performance	NB, NN, WINNOW Algo. based ensemble			✓			
[94]	2010	Recommender System	Matrix Factorization,Logistic Reg.,CF		✓				
[44]	2011	Predicting School Failure	NN, BN		✓	✓			
[26]	2012	Predicting Students Performance	MLP, NB, J48		✓	✓			
[27]	2012	Assessing Students Performance	DT	✓		✓			
[45]	2012	Predicting student performance	Ensemble, LR, SVM, MLP, J48, RF, AdaBoost, Bagging, Voting			✓			
[32]	2013	Forecasting Students Grades	BP, LR, LWR, SMOReg, M5rules	✓		✓			
[95]	2013	Assessing Students Performance	WATWIN ALGORITHM, Jadud's algorithm, LR	✓					
[28]	2013	Predict Student Failure	JRip, NNge, OneR, Prism, Ridor, J48, C4.5, SimpleCart, AD tree, Random Tree, REPTree	✓		✓			
[42]	2015	Identifying At-risk Students	JLR-SEQ, LR-SIM, LR			✓			
[96]	2015	Predicting Students Performance	DT, NB, Rule Based	✓		✓			
[97]	2016	Predicting Student final GPA	DT	✓		✓			
[98]	2017	Student performance prediction	SVM, C4.5, NB, BN		✓	✓			
[59]	2017	Student Performance Prediction	LR		✓				
[34]	2017	Student Grade Prediction	CF, MF, RBM	✓	✓				
[30]	2017	Student performance Prediction	KNN, SVM			✓			
[99]	2018	Early Detection of Students at Risk	ANN, LR, DT, AdaBoost			✓			
[100]	2018	Modelling student performance	LR		✓				
[101]	2019	Decision support systems	ANN, SVM			✓			
[102]	2019	Dropout Prediction	kNN, DT, SVM, DL			✓			
[50]	2019	Student difficulties Prediction	ANN, SVM, LR, NB, DT		✓	✓			
[103]	2020	Predicting At-risk Students	SVM			✓			
[104]	2020	Student Performance Prediction	DT, NB, kNN			✓			
[105]	2020	Student Performance Prediction	NB, DT, RF, JRip		✓	✓			
[106]	2020	Student Performance Prediction	RF, KNN			✓			
[36]	2021	Enhancing prediction of student success	NB,SVM,RF, ANN		✓	✓			
[107]	2021	Predicting Student Success	NB, SVM, RF, ANN			✓			
[51]	2022	Predicting Student dropout	RF, XGBoost, GB, FNN		✓	✓			
[52]	2022	Predicting Student performance and their influencing factors	CNN, LSTM		✓	✓	✓		
[53]	2022	Identifying factors helping career placements	NB, LR, DT, SVM, KNN, Ensemble models		✓		✓		

RBM- Rule-based model, ML- Machine learning, XML- Explainable machine learning, BS- Basic statistics, KG- Knowledge graph, RA- Remedial Actions

As the information in Table 1.1 highlights, prior research has predominantly focused on predicting student performance using basic features and machine learning techniques. However, some recent studies have begun to shift their focus toward developing early warning systems for at-risk students, identifying the factors contributing to poor performance, developing effective remediation strategies to prevent failure, improving system performance through efficient feature engineering, and developing effective ranking systems. This shift in focus highlights the need for a more comprehensive approach to understanding and addressing the challenges facing at-risk students in educational systems.

1.4 Research Gaps

The primary findings of the literature review are as follows:

- Only a few studies have focused on the early detection of at-risk students [46–49].
- Many studies have used machine learning techniques to predict at-risk students with reliable performance [21, 26–35]. However, the different factors influencing student academic underachievement were not sufficiently explored in these studies.
- Lack of transparency is preventing the different educational institutions from adopting traditional black-box ML-based applications to generate meaningful insights into the factors that will be beneficial within remediation efforts.
- The current machine learning studies do not identify relationships or similarities between students; they treat students independently.
- The potential of combining knowledge graphs with different machine

learning and deep learning techniques is an underexplored research area within the educational domain.

- In order to provide robust and reliable performance, most ML-based algorithms require a considerable amount of quality data. Machine learning algorithms trained on small datasets might not be able to provide results that can be generalizable in real-world applications. Most studies do not examine the effects of data quality and quantity on the overall performance of the frameworks proposed for at-risk student classification.
- Very few studies have focused on class balancing or data balancing [108]. Class balancing is considered important in obtaining high classification performance [108].
- The use of temporal features for the classification of at-risk students has not been leveraged effectively in existing studies [46, 70, 109]. The values of these features change over time due to their dynamic nature. Incorporating temporal features within classification methods can enhance their performance [46,70,109].
- Most of the research studies address student performance assessment as binary classification [21, 26–35]. Very few studies have focused on unsupervised learning-based approaches. The introduction of other classes could facilitate the development of more effective intervention plans.
- Further attention should be invested in feature engineering tasks, as features can influence the predictor's performance. The existing studies used students' demographics, academic performance, and e-learning interaction session logs.

- Most of the studies used traditional machine learning algorithms, such as SVM, DT, NB, and KNN [21, 26–35]. Only a handful of studies investigated the potential of deep learning algorithms [70–72, 110].
- Most of the studies utilized basic-level ML methods to predict whether students are at-risk. Very little attention was invested in the development of prediction models that combine supervised and unsupervised learning-based approaches.
- No existing study has leveraged cluster-based knowledge graphs (KGs) in combination with ML to classify students into high-risk and low-risk categories. KGs can provide information regarding student performance throughout the course, and customized remedial actions can be developed for at-risk students in a timely manner.

1.5 Summary

This chapter presented the problem statement and the main objectives of the thesis. The literature review presented an exhaustive overview of prior studies and identified gaps in knowledge and understanding. The information presented in this literature review also provided a solid overview of the current developments in the field of interest. The insights generated from this review will be discussed in more depth in the next section. In summary, the literature review has highlighted that there is a need for more research on the identification of at-risk students, the factors that contribute to their poor performance, and the development of effective remediation strategies to prevent failure. Additionally, there is a need for more transparent and interpretable machine learning methods, the incorporation of temporal features, and the use of unsupervised learning approaches. Furthermore, more attention should be invested in feature engineering and the use of deep learning algorithms. Overall, there is a requirement for further research to develop personalized and effective interventions for at-risk students in higher educational institutions.

Chapter 2: Methods and Results

This section describes the approach by which a solution was developed to achieve the four objectives that underpinned this research. It will also present the specifics of the dataset, various pre-processing operations, and suggested models associated with the proposed solution. This discussion will be subdivided into four sections, one for each of the four objectives. The techniques and frameworks used to analyze the data varied from objective to objective. As such, a distinct subsection will be included to describe the proposed models for each of the objectives, the subsequent results and outcomes, and a discussion of the results. The various advantages and disadvantages of each of the frameworks will then be described to help inform future research efforts.

2.1 Methodology

This research aimed to address the challenges identified in the literature review through the development and application of four objectives that cumulatively created an effective framework for identifying and assisting students at risk of failing or dropping out from their courses. Figure 2.1 illustrates the methods used to achieve the overall goal of the research, which was to develop a solution that facilitates the early detection of students at risk of failure or dropping out. The four main objectives of the research are also presented in Figure 2.1. The educational dataset was gathered for this study before pre-processing techniques, tools, and methods for data analysis were applied. This section presents a method for identifying and predicting at-risk students using a rule-based model, ML techniques, and knowledge graphs. We developed a customized rule-based warning model that accurately identifies the performance of at-risk students. The model takes into account several factors that may contribute to a student's poor performance and identifies individuals who are at risk of failure. We also constructed an explainable ML-based framework capable of predicting the performance of at-risk students in the early stages of their course and identifying factors that influence student performance. Knowledge of these factors can subsequently be leveraged to inform the planning of appropriate remedial actions for at-risk students.

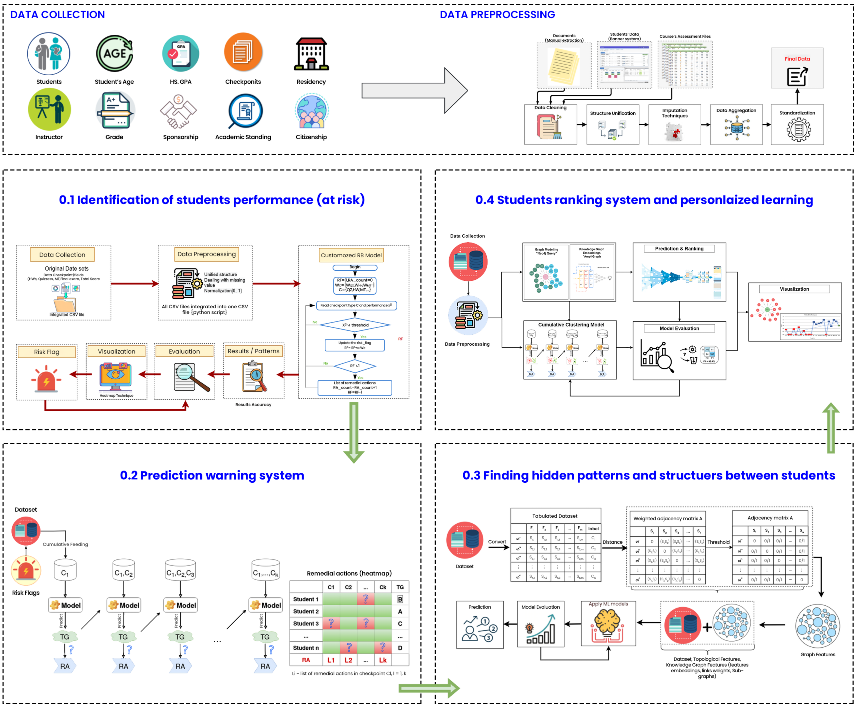


Figure 2.1: Overview methodology of the study

The model takes into account several factors that may contribute to a student's poor performance and uses machine learning techniques to predict which students are at risk of failure. We then combine knowledge graph topological features with original data features to improve the prediction accuracy of the proposed student performance. This involves the use of graph-based techniques to extract features from the data. These features can be used to improve the accuracy of the prediction solution and generate personalized learning methods based on knowledge graphs per student's performance and capabilities. In the current study, this involved adding to the solution developed as part of the previous objectives to rank students based on their performance and then generating personalized learning methods for each student based on their individual performance and capabilities.

2.2 Data Collection

The dataset used within the research was collected from the College of Information Technology (CIT) at the United Arab Emirates University (UAEU) and incorporated data spanning Undergraduate courses were taught separately to gender-segregated classes between 2016 and 2021. Three different datasets were used in this research study (See Table 2.1), depending on the course from which the data was collected.

Table 2.1: An overview of the datasets used

	D1	D2	D3
Course name	Algorithms & Problem Solving	Object-oriented Programming	Algorithms & Problem Solving
Course code	CSBP119	CSBP219	CSBP119
Course period	Fall 2016-Spring 2019	Fall 2017-Spring 2021	Fall 2017-Spring 2021
Dataset size	218	230	201
Attributes	Checkpoints features	Checkpoints & historical features	Checkpoints features

Dataset 1 contains data from 218 students enrolled on the the “Algorithms & Problem Solving" course. Dataset 2 contains course data from 230 students enrolled in the “Object-oriented Programming” course, along with information regarding historical features (e.g., past performance in different courses, age, gender, year of enrollment). Dataset 3 spanned 201 students enrolled on the “Algorithms & Problem Solving" course, along with information regarding learning objectives and topics covered at each checkpoint. To classify the students into high-risk and low-risk categories, data regarding the performance of the students was collected from homework assignments, quizzes, midterms, and finals. Prior to pre-processing, raw data from 730 students enrolled in these courses were collected. After pre-processing functions were employed to remove redundant, inconsistent and incomplete entries, the final data contained entries spanning 648 students (where each course contributed 218, 230 and 201 student data) with 38 features remaining (See Table 2.2). The aforementioned performance indicators and their relevant statistics (e.g., mean, median, and mode values for each indicator) were used for training the proposed models to achieve the four objectives. The manner in which the data was used from the three datasets varied to some extent. For example, the achievement of Objective One was possible using data from Dataset 1 only. Whereas Objectives Two and Three required data from all three datasets.

2.3 Data Pre-Processing

A variety of pre-processing techniques were used to ensure that the raw data was efficiently cleaned, polished, and prepared before it was used for training and validation of the ML and rule-based models associated with each of the objectives. This section describes the pre-processing tasks carried

Table 2.2: Dataset features

Internal Assessment Features (Checkpoints)		**Historical Features**	
Feature	Description	Feature	Description
ID	Student ID	AGE_ADMITTED	Age of the student when he/she was admitted to university
Quiz1Norm	Score on Quiz 1	AgeCourseStart	Age of the student when he/she took course CSBP219
Quiz2Norm	Score on Quiz 2	Reg. Hrs	Registered Hours
HW1Norm	Score of HW 1	HS_GPA	High School GPA
HW2Norm	Score of HW 2	MATH	Math Grade in High School
HW3Norm	Score of HW 3	PHYS	Physics Grade in High School
HW4Norm	Score of HW 4	Repeated Grade (CSBP119)	How many times the student repeated course CSBP119
MTNorm	Midterm Exam Score	CSBP119-max	Grade of CSBP119 course
Quiz3Norm	Score on Quiz 3	Repeated Grade (CSBP219)	How many times the student repeated course CSBP219
Quiz4Norm	Score on Quiz 4	ESPN	Introduction to Academic English
Quiz5Norm	Score on Quiz 5	CSBP121	Programming Lab I
Quiz6Norm	Score on Quiz 6	MATH105	Calculus I
Qzs	Average Quiz Score	CENG205	Digital Design & Computer Organization
HWs	Average Homework Score	PHYS105	General Physics I
PRJ1Norm	Project Score	CENG202	Discrete Mathematics
FENorm	Final Exam Score	CIT?	Student in College of IT (Yes/No)
TGNorm	Total Grade in CSBP219	AcademicStanding	1 = Good standing, 0 = Probation
		Citizenship	1 = Citizen, 0 = Non-Citizen
		Gender	F/M
		Sponsor	Yes/No
		AlAinResident	0 = Resident outside Al-Ain, 1 = Resident of Al-Ain

out for all four of the objectives mentioned in this dissertation. The first step involved gathering and consolidating course and student data from various sources. The second phase involved removing any entries that contained omissions or inaccuracies. Since different portions were taught by various instructors, data from some sections had to be changed until it was uniform across all sections to achieve homogeneity. For normalization purposes, the values were all changed from categorical to numerical (binary) during the last stage of the data processing stage. A new column was introduced to the dataset in relation to various rules and checkpoints.

Table 2.3: Example of an input file for the model

Student ID	C_1	C_2	C_3	**...**	C_n
$Student_1$	$g_{1,1}$	$g_{1,2}$	$g_{1,3}$	$\cdots$	$g_{1,n}$
$Student_2$	$g_{2,1}$	$g_{2,2}$	$g_{2,3}$	$\cdots$	$g_{2,n}$
$Student_3$	$g_{3,1}$	$g_{3,2}$	$g_{3,3}$	$\cdots$	$g_{3,n}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$Student_m$	$g_{m,1}$	$g_{m,2}$	$g_{m,3}$	$\cdots$	$g_{m,n}$
Max grade	$max(g_{C_1})$	$max(g_{C_2})$	$max(g_{C_3})$	$\cdots$	$max(g_{C_n})$

An example data file structure (see Table 2.3) was employed following that of [111]. This structure is shown below:

C_i	- name of the predefined checkpoint
$g_{i,j}$	- grade of the j^{th} student at checkpoint C_i
$max(g_{C_i})$	- maximum possible grade for checkpoint C_i
m	- number of students
n	- number of checkpoints in the course
i,j	- indices, $i=\overline{1,n}, j=\overline{1,m}$

All three datasets included homework components ($HW_i, i=\overline{1,h^D}$, $HW_{mean} = \frac{1}{h^D}\sum_{i=1}^{h^D} HW_i$,$h^D$ is a number of homework assignments), quiz scores ($Qz_i, i=\overline{1,q^D}$, $Qz_{mean} = \frac{1}{q^D}\sum_{i=1}^{q^D} Qz_i$), mid-term grades MT, final exam grades FE, and the total grade TG, where $\cdot^D$ denotes the dataset used, $h^{D1}=4, h^{D2}=1, h^{D3}=2$, $q^{D1}=6, q^{D2}=4, q^{D3}=5.$ All checkpoints were applied cumulatively up to the final exam as input variables to the model.

2.4 Evaluation Measures

The set of methods we proposed in this research work are known to be effective techniques for improving the current prediction approaches. In this case, we will consider the performance of the proposed models are satisfactory if the specificity and sensitivity of the classification models are higher than 80% and the fraction of the MAE over the range of the predicted feature is less than 10% based on the regression models.

Assessment indicators of the regression task were used to assess the difference between the predicted and actual values. In this work, the mean-squared error (MSE), mean absolute error (MAE), and determination coefficient (R^2) were chosen as evaluation indicators for the prediction of students' performance. The MAE is used to assess the quality of a regression model. This is a measure of absolute difference between two continuous variables; it gives a clear understanding of the error between the actual and predicted values and is widely used for decision-making in different communities. Suppose that the size of the dataset is n, the actual value is y, and the predicted value is $\hat{y}$. The formulae for calculating these indicators are:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2, \tag{2.1}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|, \tag{2.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2}. \tag{2.3}$$

The range of values of R^2 is $[0,1]$; the closer the value is to 1, the better the fit of the model. To investigate the quality of the evaluation indicators, significance and correlation analyses were conducted. The p-value is a measure of significance; in general, if the p-value is less than 0.01, the correlation between two groups of data is deemed to be statistically significant, and the smaller the p-value, the greater the significance level. The Pearson correlation coefficient r, which is a statistical metric that describes the degree of linear correlation between two variables, was used to measure correlations between the models and the actual results.

To assess the quality of the classification models we use sensitivity, specificity, ROC AUC, accuracy and balanced accuracy metrics. The confusion or error matrix is built for each predictive model to show how it can distinguish between classes. Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC) are used for performance evaluation of the classifiers and memorization of the trade-off between true positive rate (TPR) and false positive rate (FPR) using different probability thresholds. Sensitivity is true positive rate, while Specificity is true negative rate. Sensitivity and specificity are used in ML to evaluate the accuracy of classification models. Sensitivity measures how often the model correctly identifies positive outcomes, while specificity measures how often it correctly identifies negative outcomes. High sensitivity means good detection of positive cases, while high specificity means good avoidance of false positives. Both metrics are important for evaluating the overall performance of a classification model and are often used in combination with other metrics.

$$TPR(sensitivity) = \frac{TP}{TP+FN} \tag{2.4}$$

$$TNR(specificity) = \frac{TN}{TN+FP} \tag{2.5}$$

$$BAC(Balanced\ Accuracy) = \frac{Sensitivity + Specificity}{2} \tag{2.6}$$

The overall accuracy of the model is defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.7}$$

where TP, TN, FP, FN are the true positive, true negative, false positive and false negative values representing the confusion matrix of classification model respectively.

Models are trained using 10-fold cross-validation technique. The metrics are calculated for each fold separately and then averaged values are used as final measure.

2.5 Objective One: Identify At-risk Students Using a Customized Rule-Based Model

A rule-based system is typically composed of an AI-based approach to provide scores for the current observations based on previously acquired and recorded knowledge and experiences. This facilitates the generation of meaningful insights regarding the research problem. In the case of Objective One, this was the identification of at-risk students. The sets of rules defined in the model are often chosen by a human subject matter expert based on their past experiences of identifying at-risk students at an early stage in the course progression. In this regard, a customized rule-based model was used in this study to identify at-risk students. Figure 2.2 highlights the proposed model for Objective One. The next sub-section will explore the technical aspects of the proposed rule-based model for the identification of at-risk students.

2.5.1 Objective One: Method

A customized rule-based model was developed that could facilitate the achievement of Objective One (see Figure 2.2). The input features relate to the performance of the students in various quizzes, assignments, midterms, and final exams for individual courses. The raw data from the different students were pre-processed, so any raw data inconsistencies related to the different instructors and their teaching practices were reduced. The detailed work for Objective One can be found in Appendix (Article one [111]).

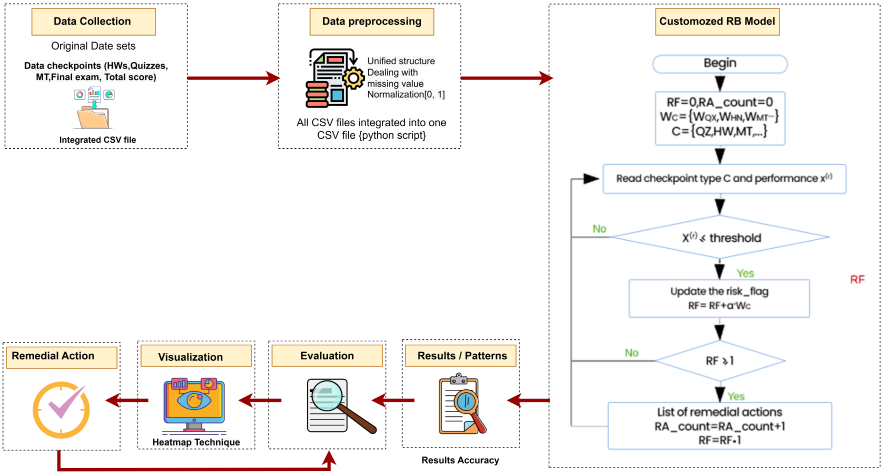


Figure 2.2: Proposed model for identifying and classifying at-risk students

A sequential rule-based model was created that determines a student's risk flag depending on how well they are currently performing on the given indications (checkpoints) in the data. The threshold value for the model can be manually set by the instructor to specify the minimum amount of risk tolerated before corrective action is taken. This will help to ensure that the right remedial actions are implemented at the right time to lower the rate of student failure. The output of the model is a rule-based outcome that delivers a weighted value that adequately determines whether the student falls into the category of at-risk or not-at-risk students.

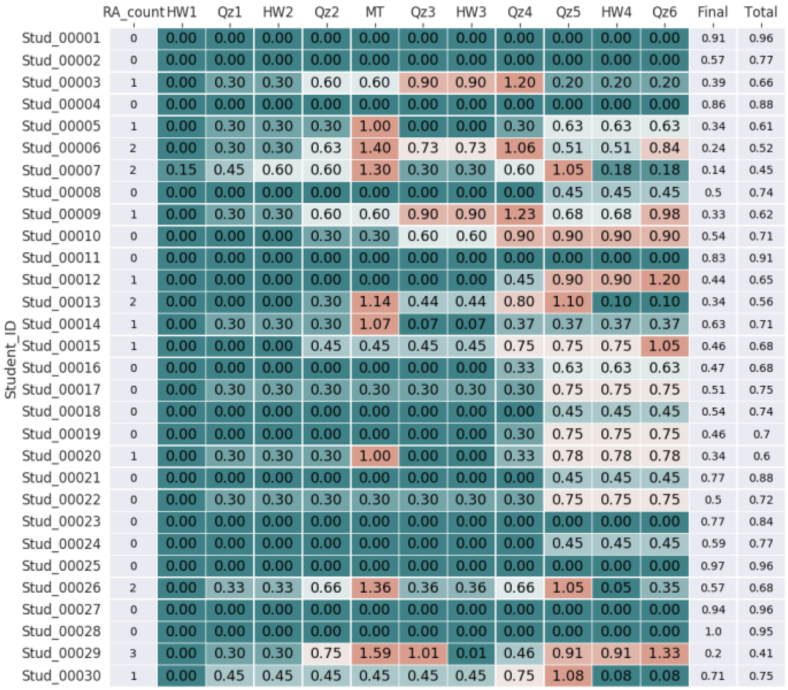
The student is labeled as being at risk if the value of threshold falls below 0.7. The risk flag (RF) indicator, which determines whether the student is at risk, is computed using Equation 2.8.

$$RF_i = RF_{i-1} + a_i * W_i \tag{2.8}$$

W_i is a weighted value for a specific checkpoint (such as a quiz, assignment, midterm, or final), a_i is a weighted coefficient to enhance the risk factor for a certain checkpoint, and RF_{i-1} is the value of risk flag determined for the prior checkpoint in Equation 2.8. If the value of RF_i exceeds 1, the remedial action counter, RA_count, is incremented by one, and 1 is deducted from the value of RF_i. The suggested approach achieves this by emphasizing the frequency of corrective measures needed for each student as well as the overall risk score to categorize whether or not a student is at risk at any particular checkpoint. Each of the values within the model's various variables can be readily updated, allowing the instructor to modify them (e.g., weighted scores and weighted values) for each checkpoint in line with the requirements of each unique course.

2.5.2 Objective One: Results and Discussion

Using the proposed framework developed for Objective One, instructors will be able to identify at-risk students early in the semester in order to implement timely interventions. In addition to this model, a heatmap-based visualization method was developed that can effectively visualize the at-risk landscape and required remedial actions. Figure 2.3 shows the visualization that the instructors can use to view student performance. At-risk students are separately highlighted. In addition to providing instructors with insights regarding localized (checkpoint-by-checkpoint) based performance and risk calculation, the heatmap also offers a globalized overview of the final grades of the students and the frequency of remedial actions required by each student over time. Using the proposed model and heatmap-based visualization can allow the instructor to highlight the different students who are at risk. This can be



Student_ID	RA_count	HW1	Qz1	HW2	Qz2	MT	Qz3	HW3	Qz4	Qz5	HW4	Qz6	Final	Total
Stud_00001	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.96
Stud_00002	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.57	0.77
Stud_00003	1	0.00	0.30	0.30	0.60	0.60	0.90	0.90	1.20	0.20	0.20	0.20	0.39	0.66
Stud_00004	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.86	0.88
Stud_00005	1	0.00	0.30	0.30	0.30	1.00	0.00	0.00	0.30	0.63	0.63	0.63	0.34	0.61
Stud_00006	2	0.00	0.30	0.30	0.63	1.40	0.73	0.73	1.06	0.51	0.51	0.84	0.24	0.52
Stud_00007	2	0.15	0.45	0.60	0.60	1.30	0.30	0.30	0.60	1.05	0.18	0.18	0.14	0.45
Stud_00008	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.45	0.45	0.45	0.5	0.74
Stud_00009	1	0.00	0.30	0.30	0.60	0.60	0.90	0.90	1.23	0.68	0.68	0.98	0.33	0.62
Stud_00010	0	0.00	0.00	0.00	0.30	0.30	0.60	0.60	0.90	0.90	0.90	0.90	0.54	0.71
Stud_00011	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.83	0.91
Stud_00012	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.45	0.90	0.90	1.20	0.44	0.65
Stud_00013	2	0.00	0.00	0.00	0.30	1.14	0.44	0.44	0.80	1.10	0.10	0.10	0.34	0.56
Stud_00014	1	0.00	0.30	0.30	0.30	1.07	0.07	0.07	0.37	0.37	0.37	0.37	0.63	0.71
Stud_00015	1	0.00	0.00	0.00	0.45	0.45	0.45	0.45	0.75	0.75	0.75	1.05	0.46	0.68
Stud_00016	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.63	0.63	0.63	0.47	0.68
Stud_00017	0	0.00	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.75	0.75	0.75	0.51	0.75
Stud_00018	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.45	0.45	0.45	0.54	0.74
Stud_00019	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.75	0.75	0.75	0.46	0.7
Stud_00020	1	0.00	0.30	0.30	0.30	1.00	0.00	0.00	0.33	0.78	0.78	0.78	0.34	0.6
Stud_00021	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.45	0.45	0.45	0.77	0.88
Stud_00022	0	0.00	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.75	0.75	0.75	0.5	0.72
Stud_00023	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.77	0.84
Stud_00024	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.45	0.45	0.45	0.59	0.77
Stud_00025	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.96
Stud_00026	2	0.00	0.33	0.33	0.66	1.36	0.36	0.36	0.66	1.05	0.05	0.35	0.57	0.68
Stud_00027	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.96
Stud_00028	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.0	0.95
Stud_00029	3	0.00	0.30	0.30	0.75	1.59	1.01	0.01	0.46	0.91	0.91	1.33	0.2	0.41
Stud_00030	1	0.00	0.45	0.45	0.45	0.45	0.45	0.45	0.75	1.08	0.08	0.08	0.71	0.75

Figure 2.3: Visualization of the level of risk faced by different students and the overall classification of at-risk students

used as a starting point for developing effective remedial actions that can address students’ weak points in a timely manner.

The study also revealed that the number of remedial actions required is inversely proportional to the predicted performance at the end of the semester. For example, a student requiring a greater level of remedial actions will achieve lower grades than a student who requires a lower number of remedial interventions a throughout the course of the semester. In this respect, Figure 2.4 highlights the relationship between overall final performance at the end of the semester and the need for remedial actions. This demonstrates that the overall performance of the student will be substantially affected as the need for remedial actions increases.

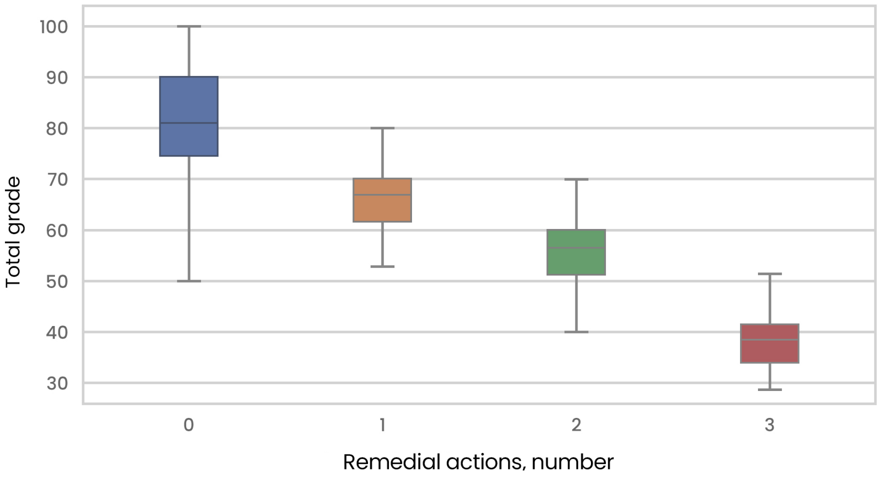


Figure 2.4: Barplot showing the relationship between the overall performance of the students and the need for remedial actions throughout the semester

Consequently, the instructors should focus on the development of effective remediation strategies and interventions that can minimize the need for further action.

2.6 Objective Two: Hybrid Framework Using Rule-Based Model and Explainable Machine Learning for Classification of At-Risk Students

Objective Two involved the development of a hybrid framework using a rule-based model and explainable machine learning for the classification of at-risk students. Further details about the work of Objective Two can be found in the Appendix, (Article Two, [112]). Although the rule-based approach is successful in identifying at-risk students, as highlighted in the prior section, it does have some limitations in terms of its capabilities to effectively tackle the underlying research problem of identifying existing and future at-risk students based on the existing data related to student performance. To mitigate some of the shortcomings of the rule-based approach, a slightly different course of action was leveraged to achieve Objective Two. Specifically, an explainable ML-based approach was used to identify the different underlying factors and elements that contribute to students being classified as either "at risk" or "not at risk". Objective Two was to enhance the functionalities of the developed solution into the following: (i) to provide an improved system for the classification of at-risk students, and (ii) to provide effective and timely remedial actions for at-risk students so that their overall course performance can be improved and failure can be prevented. The next sub-section will shed light on the technical aspects and different experiments performed to achieve Objective Two. It will also present an overview of the overall performance of the developed system.

2.6.1 Objective Two: Method

Objective Two was to develop an effective explainable ML and rule-based framework. A systematic approach was adopted to combine learnings from three different experiments as a means of enhancing the overall framework. Table 2.4 summarizes the three experiments that were performed to achieve Objective Two.

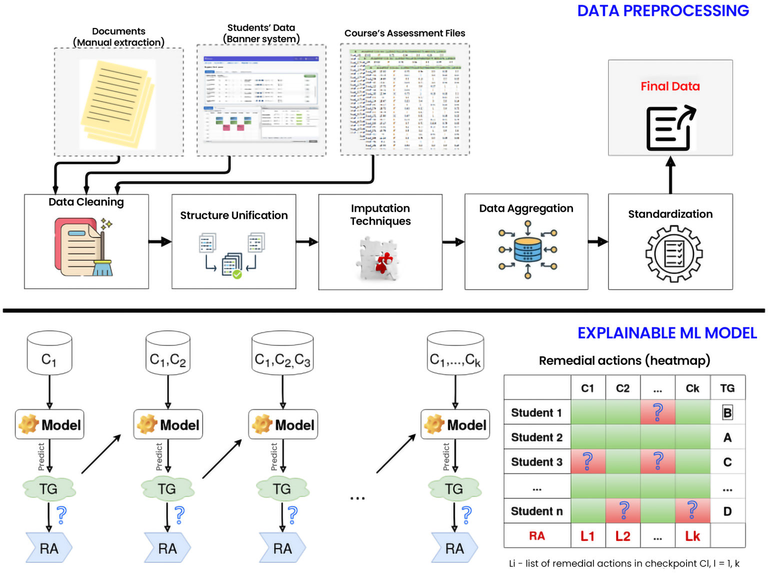
Table 2.4: Summary of experiments

	Experiments 1	**Experiments 2**	**Experiments 3**
Dataset (size)	D1 (218), D2 (230)	D1 (218), D2 (230)	D3 (201)
Attributes	Checkpoints	Checkpoints & Historical Features	Checkpoints
Target performance in	TG	MT, TG	MT, TG
Classification	Multiclassification: Good / AtRisk / Failed	Multiclassification: Good / AtRisk / Failed	Binary classification: Good / AtRisk
Remedial Actions	No	No	Yes

MT - midterm exam grade, TG - total grade.

The proposed framework developed for the classification of at-risk students and the identification of appropriate remedial actions is outlined in Figure 2.5. The top section of Figure 2.5 highlights the different functions for pre-processing and preparing the raw data from different students so that it can be used for training and validating the proposed ML-based models. In the first experiment, course checkpoints were used with dataset 1 and 2 to train different ML-based models (e.g., Support Vector Machines (SVM), XG Boosting classifier, Light Gradient Boosting Machine (LightGBM), Naive Bayes (NB), ExtraTrees (ET), Random Forests (RF), Multilayer Perceptron (MLP), and Bagging models). The final output classified the different students into one of three categories: good, at-risk or failed categories. Using valuable insights gained from the first experiment, the second experiment was designed such that different features (e.g., historical features such as age, gender, past performance, and course-level features, such as midterm, final, quiz and assignment grades) were utilized.

The diverse features were categorized in terms of importance using different feature selection methods (e.g., mean absolute difference [MAD], information gain, Chi-square test, and correlation coefficient). Different ML-based models were trained on the final set of features to classify students into good, at-risk, or failed categories. Using valuable lessons gained from the first two experiments, experiment 3 was designed to use different features, course-level features, and risk flags features. The data from different checkpoints were added cumulatively to train a hybrid model that was based on rule-based and explainable ML-based approaches for the classification of at-risk students. After classification, the system developed for experiment 3 could propose remedial action at each checkpoint to



	C1	C2	...	Ck	TG
Student 1	🟩	🟩	?	🟩	B
Student 2	🟩	🟩	🟩	🟩	A
Student 3	?	🟩	?	🟩	C
...	🟩	🟩	🟩	🟩	...
Student n	🟩	?	🟩	?	D
RA	L1	L2	...	Lk	

Li - list of remedial actions in checkpoint Cl, l = 1, k

Figure 2.5: Proposed model for identifying and classifying at-risk students

provide at-risk students with timely interventions that could potentially improve their final grades.

2.6.2 *Objective Two: Results and Discussion*

This section will explore the outcomes of experiments described in the preceding subsections. The three experiments covered in the previous section were used to evaluate the performance of the various models using a variety of different performance measures, including accuracy, true positive rate (TPR), and true negative rate (TNR). Equations were used to compute the performance metrics (see Section 2.4).

The training and validation of the different ML-based algorithms provide four (or more, depending on the number of classes) indicators that are used for calculating the performance of the models; namely, True

Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values. These variables were applied to equations 2.4, 2.5 and 2.7 to calculate the values of TPR, TNR and Accuracy. Of the different ML-based models examined via experiment 1, the best results were generated by the Extra Trees algorithm, which achieved accuracy results of 87% and an AUC score of 0.95. Table 2.5 summarizes the results.

Figure 2.6 highlights the relationship between the different features and their ability to predict the final classification of at-risk students using the proposed ML-based model. Some of the different features that are representative when predicting at-risk students include homework score, assignment score, overall grade point average, and other historical features, such as prior course performance and student's age, to name a few.

Table 2.5: Performance of classification models in predicting students' groups from checkpoints before the midterm exam (datasets D1 and D2)

Dataset	D1	D2
Predictors	Qz1, HW1, Qz2, HW2	Qz1, HW1, Qz2
Correlation	Positive 0.66%	Positive 0.54%
Best ML classifier	ExtraTrees	ExtraTrees
Accuracy	0.86	0.87
ROC AUC	0.96	0.95

To use the most appropriate and representative features for training the aforementioned classification models used in the first approach, various feature selection methods based on the characteristics used in the original dataset were applied. These feature selection methods ensure that the most appropriate features are used for training and validating the ML-based models by reducing feature redundancy and limiting the model size. The

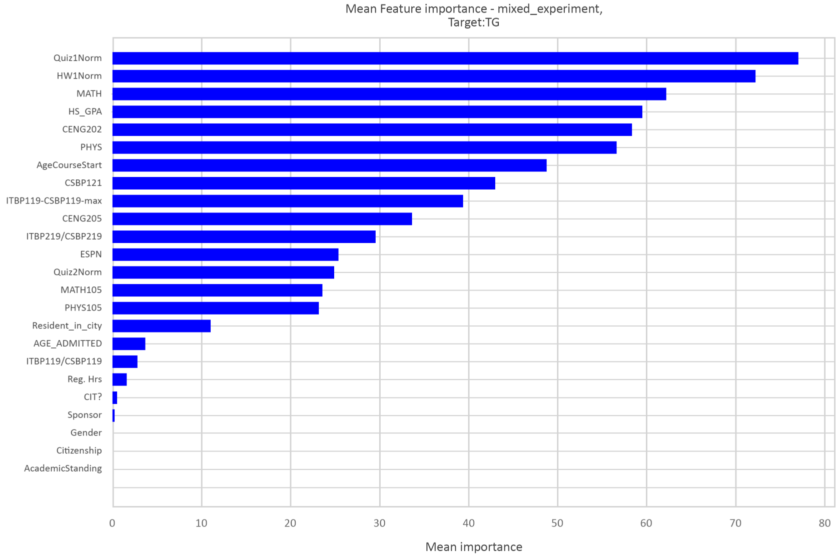


Figure 2.6: Visualization of the level of risk faced by different students and the overall classification of at-risk students

ten best features were chosen from among the many features and used to train and validate the model for classifying students at risk. For the second experiment, the use of different ML-based models was evaluated to identify the most effective means of using different sets of features to improve the final grade and midterm grade-based target performance. The ML-based model that used historical data, such as grades from previous courses, GPA, and age, as well as course-based checkpoints, such as quizzes and homework assignments, delivered the best performance, with the Extra Trees classifier achieving 84% accuracy and an AUC score of 0.96. An innovative ML-based and rule-based framework was created for the third and final experiment. In addition to generating the at-risk classification, this approach identified appropriate remediation and intervention strategies at different stages of the course progression, specifically addressing weak points and recommending pertinent intervention strategies for at-risk students. The mean AUC values of the best classifier for both sets of input features are shown in Table 2.6. As the features were cumulatively added, the prediction results clearly improved. Table 2.6 further demonstrates that the performance was enhanced by 2.05 percent by incorporating risk flags from the rule-based model. As a result, we are able to reasonably predict student performance at the course's first checkpoint, which is beneficial for both students and instructors.

Table 2.6: AUC performance of ExtraTrees model classifying students into not at risk and at risk groups

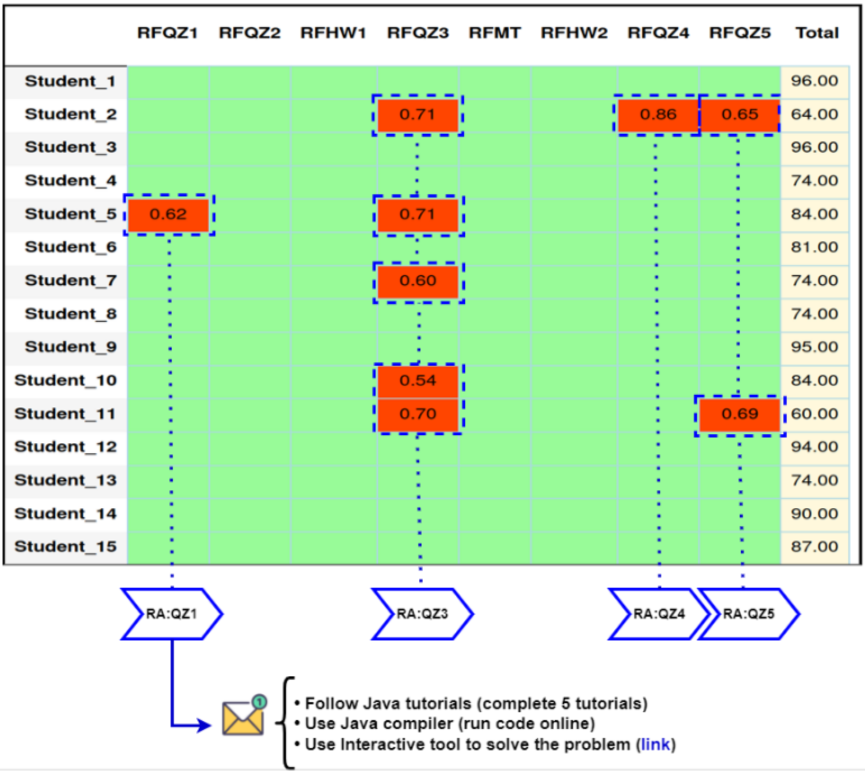
Features	C1	C1,C2	C1,C2,C3	C1, C2, C3, C4
C	0.664	0.930	0.946	1.000
C + RF	0.691	0.958	0.959	1.000
Gain	+3.9%	+2.92%	+1.36%	-

C_i - checkpoints added cumulatively

RF - risk flag added to each checkpoint

The heatmap-based visualization-based output is shown in Figure 2.7, along with corrective measures for each checkpoint at which the students acquire a favorable result for the risk flag calculation shown in Equation 2.8. Figure 2.7 demonstrates how several checkpoints along the course's progression might provide different recommendations for effective remediation, thereby ensuring that the performance of the at-risk student improves at subsequent checkpoints during the course.

Objective Two was achieved by combining ML-based and rule-based approaches. Recent research studies have reported the creation of several models for identifying and categorizing at-risk students



	RFQZ1	RFQZ2	RFHW1	RFQZ3	RFMT	RFHW2	RFQZ4	RFQZ5	Total
Student_1									96.00
Student_2				0.71			0.86	0.65	64.00
Student_3									96.00
Student_4									74.00
Student_5	0.62			0.71					84.00
Student_6									81.00
Student_7				0.60					74.00
Student_8									74.00
Student_9									95.00
Student_10				0.54					84.00
Student_11				0.70				0.69	60.00
Student_12									94.00
Student_13									74.00
Student_14									90.00
Student_15									87.00

- Follow Java tutorials (complete 5 tutorials)
- Use Java compiler (run code online)
- Use Interactive tool to solve the problem (link)

Figure 2.7: Heatmap-based visualization of the level of risk and the overall classification of at-risk students, along with remediation and intervention recommendations at each step. The risk flag is activated for different students at different checkpoints during the course progression

[38, 113–119]. When compared to the proposed model, many previous efforts (e.g., [117–119]) did not evaluate the importance of features or use feature selection methods to obtain a final set of relevant, representative features. The proposed method was based on the development of an explainable ML-based model, which, unlike previous studies [113, 115, 117–119], can clearly highlight feature selection, model development, model training, and the explainable outputs of the ML-based and rule-based models. When leveraged in combination, these features can help instructors and end users who may not be technically proficient in their understanding. The suggested approach could identify at-risk students in a timely manner and offer remedial options that are specifically designed to address issues related to the various course milestones.

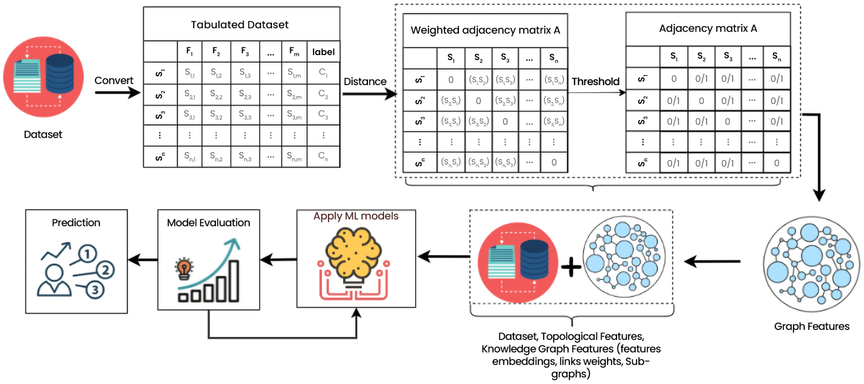
2.7 Objective Three: Graph-Based Deep Neural Networks with Graph Topological Features

Objective Three involved the development of an explainable ML-based approach for at-risk student classification. Several historical (e.g. historical features such as age, gender, and past performance) and course-level features (e.g. midterm, final, quiz and assignment grades) were utilized in the models that were examined. In Appendix, Article Three, you can find further information about the work for Objective Three. Objective Three sought to improve the performance of the at-risk student classification system with the help of features that have not yet been fully explored in the existing state-of-the-art frameworks.

Based on the existing gaps in the literature, the use of topological features with knowledge graphs and graph-based Convolutional Neural Networks (GCN) was proposed. It was anticipated that the use of topological features with knowledge graphs would provide an effective visualization tool that the instructors could leverage to better understand the inter-relationship between the different features as a means of enhancing the feature-selection process and facilitating the development of models that can accurately classify at-risk students in the future. At this stage of the research, different novel approaches were developed, validated and compared with state-of-the-art frameworks to develop robust, reliable, and practical systems for the classification and onward management of at-risk students. The following sub-sections will provide an overview of the technical details of the proposed method and the analysis of its performance.

2.7.1 Objective Three: Proposed Method

Knowledge graphs can be automatically built using ML and graph mining approaches to offer different insights into a given topic. Knowledge graphs make it easier to understand a distinct topic by revealing information in structures and removing data abstraction. They offer a straightforward, yet highly accurate method, for transforming tabular data into graphs, enabling major advancements in ML classifications. By creating an adjacency matrix with the students as nodes, calculating the distance norm (using distance metrics like Euclidean, Cosine, and Chebyshev) between the data points, and generating graph edge weights, the original tabular data from the students was transformed into a graph. Topological features were extracted from the created graphs and combined with those from the original tabular dataset. To determine the most significant features in this modified dataset, the features were rated using several feature selection techniques. To categorize the students into good, at-risk, or failed, the final dataset characteristics were utilized to train and validate several ML-based models. A graph-based convolutional network (GCN) was used to compare the performance of conventional ML-based models. A knowledge graph based on the features of the suggested dataset was utilized to look at the relationships between various features. The proposed model created for categorizing at-risk students is highlighted in Figure 2.8.


Dataset
Convert
Distance
Threshold
Graph Features
Dataset, Topological Features, Knowledge Graph Features (features embeddings, links weights, Sub-graphs)
Apply ML models
Model Evaluation
Prediction

Tabulated Dataset

	F_1	F_2	F_3	...	F_m	label
s_1	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	...	$S_{1,m}$	C_1
s_2	$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	...	$S_{2,m}$	C_2
s_3	$S_{3,1}$	$S_{3,2}$	$S_{3,3}$	...	$S_{3,m}$	C_3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
s_n	$S_{n,1}$	$S_{n,2}$	$S_{n,3}$	...	$S_{n,m}$	C_n

Weighted adjacency matrix A

	s_1	s_2	s_3	...	s_n
s_1	0	$(s_1 s_2)$	$(s_1 s_3)$	...	$(s_1 s_n)$
s_2	$(s_2 s_1)$	0	$(s_2 s_3)$	...	$(s_2 s_n)$
s_3	$(s_3 s_1)$	$(s_3 s_2)$	0	...	$(s_3 s_n)$
⋮	⋮	⋮	⋮	⋮	⋮
s_n	$(s_n s_1)$	$(s_n s_2)$	$(s_n s_3)$	...	0

Adjacency matrix A

	s_1	s_2	s_3	...	s_n
s_1	0	0/1	0/1	...	0/1
s_2	0/1	0	0/1	...	0/1
s_3	0/1	0/1	0	...	0/1
⋮	⋮	⋮	⋮	...	⋮
s_n	0/1	0/1	0/1	...	0

Figure 2.8: Illustration of the methodological steps (pipeline) for Objective Three

2.7.2 *Objective Three: Results and Discussion*

Objective Three involved the development of a four-layered graph-based deep learning model that can reliably solve complex problems using graph-based representations of original features. The GCN model required a normalized adjacency matrix and feature matrix as inputs to develop the at-risk students' classification model. The performance of the different ML-based approaches and GCN was examined and compared using metrics such as accuracy, AUC, true positive rate (TPR) and true negative rate (TNR). The formula for these three performance metrics was discussed in Section 2.4. Different datasets and diverse sets of features were used to train different ML-based models (e.g., SVM, XG Boosting classifier, LightGBM, Naive Bayes, ExtraTrees, RF, MLP, and Bagging models). The best performance of the ML-based model was provided by the Extra Trees classifier with an accuracy of 95% and AUC of 99.5%. When the original dataset was combined with topological features using Euclidean and Cosine metrics, the best performance was exhibited by Random Forest classifier with an accuracy of 89.1% and AUC of 98.3%. In the final stage of the proposed model, the combination of original dataset features and topological features was trained and validated using GCN. Figure 2.9 presents a comparison of the performance of the different approaches outlined in this paper.

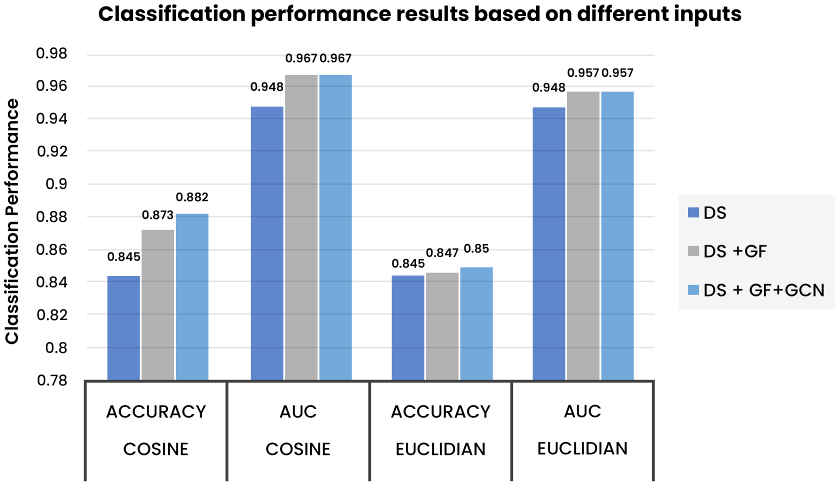


Figure 2.9: Classification performance of the ML-based classifiers in different set of features

As can be observed in Figure 2.9, the highest level of accuracy was achieved by combining the original dataset features with topological features trained using four-layered GCN to provide an accuracy of 88.2% and AUC of 96.7%. The knowledge graph is generally composed of entities along with their semantic properties and relationship. The final knowledge graph, which was based on a total of 14 features in the given dataset, is presented in Figure 2.10. In addition to showing the different features and entities, it also highlights the inter-relationship between the entities of interest and the primary node; i.e., the students.

We compared baseline approaches from the literature to the results of the current study. The outcomes reveal that the framework proposed for Objective Three outperformed the state-of-the-art approaches. As such, it represents a novel method for estimating student performance using a framework that incorporates the original dataset features, graph representation features, and a GCN. We graphed student data using distance measurements. To enrich the data, we extracted topological features from the graph to capture structural relationships and obtain deeper insights into past feature points. Combining graph features with extracted dataset features improved the prediction performance of the ML algorithms. The GCN graph topological features improved prediction performance even further. An ensemble of graph embedding features with several ML models yielded better results.

2.8 Objective Four: Students' Ranking Improvement Using Cluster-based Knowledge Graphs and Entity-Relation Representation with Machine Learning

Objective Three involved the use of topological features in combination with GCN for the classification of at-risk students. Achieving Objective Four involved extending the development of the at-risk student classification by incorporating additional beneficial features that can assist instructors in a number of different ways. Fore more details on the work towards Objective Four see Appendix, Article Four. The use of the knowledge graph was expanded for the development of student ranking and at-risk student classification to provide a personalized learning-based approach for at-risk students. The use of knowledge graphs and an

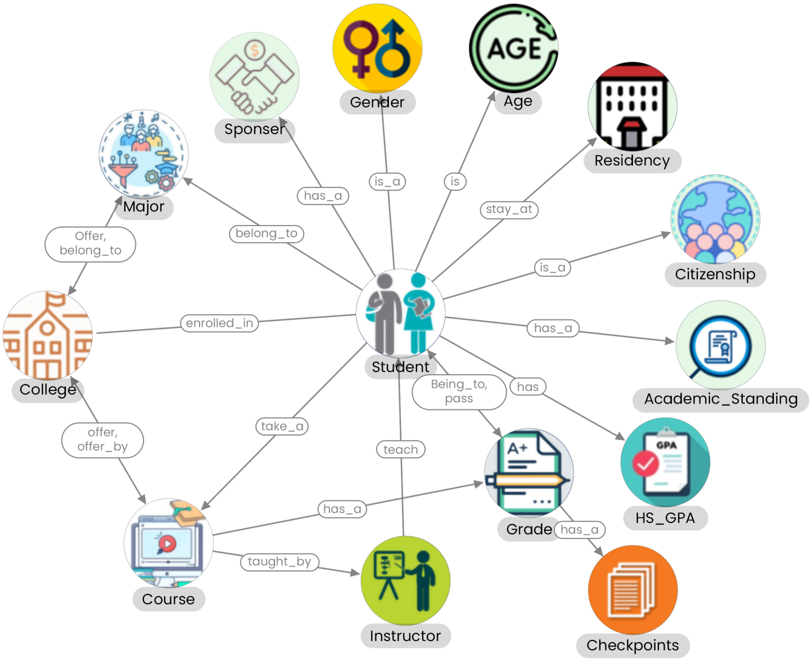


Figure 2.10: Knowledge graph of the inter-relationship between the different topological features used for the classification of at-risk students

unsupervised learning-based approach was leveraged to provide an efficient student ranking system supplemented with an effective visualization-based approach for the instructors to examine the different factors leading to the low performance of students. Personalized learning was used in this study to provide individualized and tailored recommendations that can enable at-risk students to improve their course-level performance.

2.8.1 Objective Four: Proposed Method

Unlike the previous three objectives, Objective Four involved the use of a novel method of student ranking and at-risk classification using a combination of Cluster-based KGs and ML-based methods. Using this method, an improved student ranking and at-risk student assessment approach was developed that was supplemented with the development of personalized remedial actions to improve student performance and prevent course failures. The proposed method for Objective Four was based on the following processes: (i) KGs generation using entity representations and graph embedding, (ii) K-Means Clustering to develop clusters of students based on their academic performance, (iii) Development of ML models to predict and rank student performance, and (iv) Identification of at-risk students and associated remedial recommendations using cluster-wise KGs. Figure 2.11 presents a visual representation of the different steps involved in achieving Objective Four.

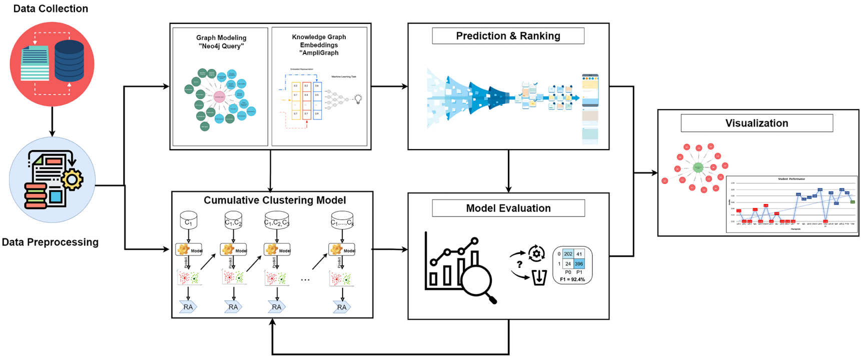


Figure 2.11: Illustration of the methodological steps (pipeline) for Objective Four

A knowledge graph was created to assess student performance in relation to a variety of variables (e.g., overall grade, high school grade, major, instructors, age, and other variables). The assessment of the students was based on the utilization of relevant numerical data, such as subject scores, checkpoints, and historical features. In addition to this numerical data, categorical data about the instructor, residency, citizenship, sponsor, college, and majors were utilized. Once the graph was generated, each student was ranked according to their scores. Based on their past performance or current checkpoints, a personalized learning and recommendation plan based on their overall grade was also generated.

Knowledge graph $G = (E, R, T)$ consisted of entities E, relations R between entities, and triples $T : (p, r, q)$, where r is a relation between p and q. A triple (p, r, q) represents knowledge between subject p and object q with a relation r. We may denote each possible triple in G as true (with a positive value, 1) or false (with a negative value, -1) in equation 2.9.

A scoring function $f(p, r, q) \in \mathbb{R}$ which represents a value of triple (p, r, q) to be true. The main goal of KG is to learn the scoring function such that the higher the value of $f(p, r, q)$, the bigger the probability that (p, r, q) is true. Different knowledge graph embedding (KGE) methods have been used for mapping relations, or entities into low dimensional space (e.g.TransE [120], HolE [121], and ComplEx [122]).

$$y(p, r, q) = \begin{cases} +1 \quad \text{if} \quad (p, r, q) \quad \text{is true,} \\ -1 \quad \text{if} \quad (p, r, q) \quad \text{is false.} \end{cases} \tag{2.9}$$

The KGE methods were used in the current study to identify students at risk and suggest appropriate remedial actions. The training process was introduced to generate the embedding vectors. First, false triple values were generated by computing the p or q, along with its scoring function $f(p, r, q)$, and training dataset $\mathbb{D}$ of given true and false values. Finally, the proposed method learned the representations of entities and relations described within the scoring function by optimizing a loss function $\mathbb{L}(\Theta) \rightarrow \min$, Θ is a set of all possible embedding vectors, $\Theta = e_i \cup r_j$. Figure 2.12 illustrates the pipeline development of embedded vectors. This study visually explored the data by reducing the embeddings to 2 or 3 dimensions using t-distributed Stochastic Neighbor Embedding (t-SNE) and Principle Component Analysis (PCA). Furthermore, K-means clustering, an unsupervised machine learning algorithm, was also employed based on unlabelled input vectors, which are used to find k centroids and assign each data point to the closest cluster to ensure minimal centroid size. Additionally, KG features were combined with insights from k-means clustering to recognize and monitor underperforming and at-risk students and develop personalized intervention strategies.

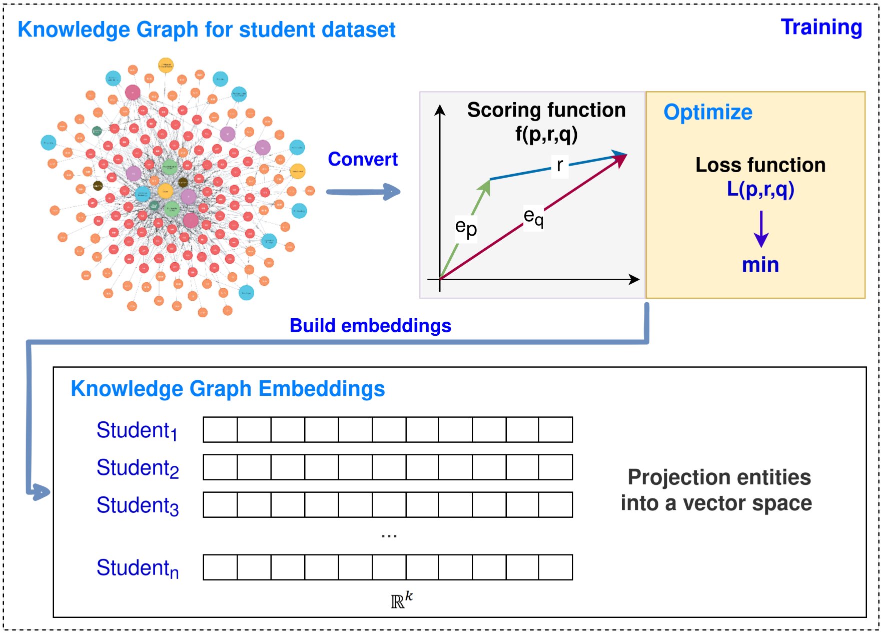


Figure 2.12: Student node representation with knowledge graph embedding vectors

2.8.2 Objective Four: Results and Discussion

This section presents an overview of the performance of the student ranking system developed for Objective Four. Different loss functions were used to assess the performance of the KGs. Each of the different KGEs requires the use of different loss functions (e.g.TransE [120], HolE [121], and ComplEx [122]). The loss function of the TransE method [120] can be described as follows:

$$\mathbb{L}_{TransE} = \sum_{(p,r,q)\in G} \sum_{(p',r,q')\in G'_{(p,r,q)}} [\gamma d(p+r,q) - d(p'+r,q')]_{+} \tag{2.10}$$

In the above equation $[\gamma d(p+r,q) - d(p'+r,q')]_{+}$ shows the absolute value. The γ is a hyper-parameter; its value is always greater than zero. $G'_{(p,r,q)}$ shows the negative sample. $d(p+r,q)$ shows the energy values, which implies that the relationship is built. The margin loss function of HolE [121] and log-likelihood loss of CompleEx [122] are as per the following equations:

$$\mathbb{L}_{HolE} = \sum_{(p,r,q)\in\mathbb{D}} \left[\gamma + \sigma(f_{HolE}(p',r,q')) - \sigma(f_{HolE}(p,r,q))\right] \tag{2.11}$$

$$\mathbb{L}_{ComplEx} = \sum_{(p,r,q)\in\mathbb{D}} log(1+e^{-y(p,r,q)f_{ComplEx}(p,r,q)}) + \lambda \left\|\Theta\right\|_2^2 \tag{2.12}$$

The following performance metrics were used to assess the performance of K-means clustering: Recall (R), Precision (P), F-measure (F1), and Accuracy (Acc). These metrics were defined and discussed in Section 2.4. Mean Square Error (MSE), Mean Absolute Error (MAE), and

Determination Coefficient (R2) were employed to assess the performance of the ML-based models used in this study (e.g., Linear Regression [LR], Support Vector Machine [SVM], LightGBM Regression [LGBM], and Adaptive Boosting [Adaboost]).

The range of values of the determination coefficient R^2 is between [0, 1], such that the closer the value is to 1, the better the ML model performs. The final performance of the proposed method leveraging a combination of ML-based models and clustering-based KGEs was assessed using True Positive Rate, which was described in Section 2.4. For the case of K-means clustering algorithm, adding the features of the KGE model (e.g., TransE, HolE, and ComplEx) produced more accurate results than clustering results based on student performance using the original features.

With respect to the results for at-risk student prediction using different ML-based approaches, the best results were obtained (i.e., minimization of MSE and MAE and obtaining higher values for R^2) when the features of the knowledge graph and original dataset were combined. When using historical features and combining checkpoint and historical features, Adaboost achieved the highest R^2 values and the lowest MAE and MSE values. After examining the individual results for the different steps and processes, it is essential to discuss the final results that are generated for student performance ranking using a prediction method that integrates cluster-based KGs with ML-based methods. Using Adaboost as the best example for ML-based student classification, the final performance of the proposed method is given in Table 2.7. Table 2.7 outlines the results of TPR of the different ML models integrated with cluster-based KGs for predicting and ranking the high-achieving and under-achieving students. The findings reveal that the inclusion of embedded features based on cluster-based KGs increased the ranking in the TransE model from 0.5 to 0.8, while the ranking of the top 20 students increased by 0.8 in ComplEx and 0.98 for all students who were identified as being at risk of failing a course. This further shows that combining different features instead of relying on original features resulted in considerably improved capabilities for predicting under-achieving and over-achieving students.

Table 2.7: Students' ranking results based on TPR

Features	**Top 10**	**Top 20**	**Good**	**At-risk**	**Lower 10**	**Lower 20**
DS (HS +CP)	0.50	0.70	0.89	0.86	0.65	0.60
DS + EF (TransE)	0.80	0.75	0.91	0.89	0.65	0.70
DS + EF (HolE)	0.70	0.75	0.91	0.92	0.70	0.70
DS + EF (ComplEx)	0.60	0.80	0.90	0.98	0.75	0.70

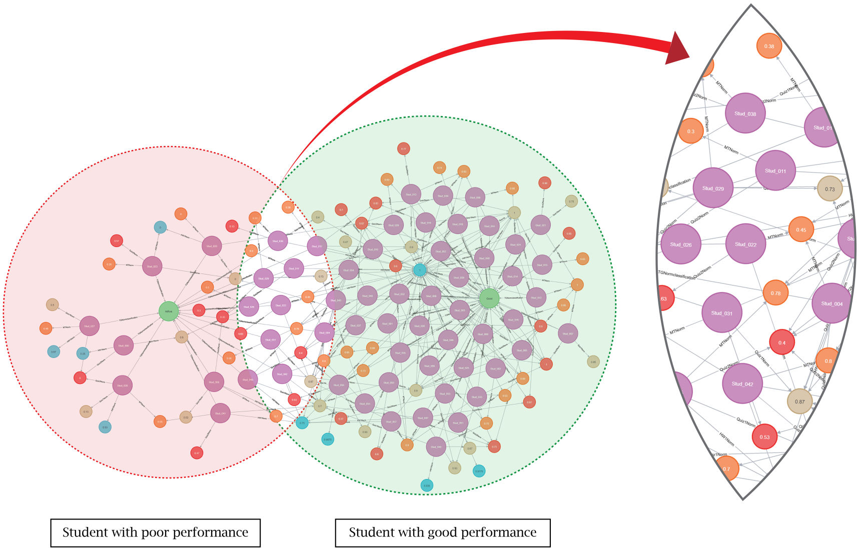


Figure 2.13: Knowledge graph representation of two clusters

In terms of visualization of the student clusters, the clear graphic representation of KG can assist instructors in performing an early assessment of student performance and propose relevant and necessary remedial strategies to prevent failure and improve student performance. The portion of the knowledge graph that is relevant to the student performance in the first four checkpoints is displayed in Figure 2.13. The purple node in the graph represents the student entity, and the green nodes, in turn, represent the clusters to which students belong—either the "good" cluster or the "at risk" cluster. The other color nodes represent the various checkpoint scores. The data indicates that students located at the intersection (Stud_004, Stud_011, Stud_022, Stud_026, Stud_029, Stud_031, Stud_038, Stud_042) are more likely to be in the middle of these two clusters. Depending on how well they perform in the upcoming checkpoints, the students may be allocated to any of these categories. These students require more attention in areas where it is possible to intervene and provide additional remedial actions in order to improve their performance. Using this visualization strategy can enhance learning and decrease the chance that students might drop out or fail their courses in the long run.

Chapter 3: Research Findings and Discussions

Chapter 3 will examine how the achievement of the four objectives of this dissertation can translate to tangible benefits for educators and students. This dissertation proposes the development of an early-warning solution that can detect and inform instructors and students whether a student is at risk of failing or dropping out of a learning program at various checkpoints throughout the course progression. Instructors can subsequently implement appropriate interventions to effectively enhance learning and reduce the likelihood that students will fail their courses in the future. The first objective of the research was to develop a customized rule-based approach for the classification of at-risk students. This involved using an AI-based approach to score current observations based on previously recorded knowledge and experiences. The rule sets, chosen by human experts, are based on their past experiences with at-risk students and the factors that help in identifying such students early on. The potential of this approach has been explored and shown to be successful in identifying at-risk students. However, the existing approaches that use various techniques to identify at-risk students still have limitations in terms of the accuracy of corrective measures. To achieve Objective One, a customized rule-based approach was proposed to quickly detect at-risk students. This strategy reacts promptly to a student's performance and provides a visual representation to help instructors identify students at risk. However, there are limitations to using a rule-based model for predicting student performance.

The second objective built on the first to create an improved solution that combines a rule-based model and an explainable ML model to better identify and predict at-risk students. Additionally, the second objective requires customized remedial actions to enhance course performance and reduce course failures. This improved solution can provide valuable information to instructors and helpful recommendations to at-risk students to improve their overall course performance. In the second objective of our research, we sought to improve upon the previously established rule-based approach for identifying at-risk students. To do so, we introduced an explainable ML-based approach that leveraged a combination of historical and course-level features to inform our predictions. The result of this combined approach was a significant improvement in our ability to accurately predict students' final grades, especially in the early stages of the course, such as before the midterm exam. Previous frameworks designed to perform at-risk student classification have relied on ML- and statistics-based methods. They have frequently fallen short in terms of providing a complete understanding of the risk factors or key indicators of student success. Our proposed method, however, offered a more transparent and interpretable understanding of these factors, allowing non-experts to gain a clearer understanding of what influences student performance. This not only enhances the accuracy of our predictions but also provides a more practical and actionable approach for both instructors and students to identify and improve performance. Additionally, the objective aimed to improve student outcomes and organizational efficiency by promptly informing students of any potential shortcomings in their performance. The mapping of checkpoint performance to Course Learning Outcomes (CLOs) and subjects helped to give targeted and specific advice to students who are at risk of failing a particular checkpoint.

In the third objective, our aim was to enhance the at-risk student prediction using novel approaches. Building on the knowledge gained from the second objective, we introduced the use of topological features in conjunction with historical and course-level features. The use of topological features with knowledge graphs in a graph-based Convolutional Neural Network (GCN) provides a more effective visual representation for instructors to better understand the relationship between features and the feature selection process for training the classification model. Additionally, by mapping students' checkpoint performance to CLOs and course subjects, the solution can provide instant corrective actions, increasing both student achievement and organizational effectiveness. The main goal of the third research work was to provide a more improved solution for at-risk student prediction using unique techniques not described in previous work. This solution offers substantial advantages in terms of performance enhancements compared to prior methods, as it leverages a variety of features to give useful information on the contributing factors to students' at-risk status. Recent advancements in representation learning on graphs have boosted node classification and connection prediction for graph-structured data [123]. Research suggests that topology can enhance classification performance [124–126], and our study supports this, demonstrating that incorporating topological features and graph embeddings significantly improves classification model performance (as shown in Article 3). Additionally, other authors have demonstrated that converting tabular data to graph data can also improve classification model performance [127]. The authors employed a scalar product to extract node relations and tested various distance norms. As educational datasets contain student performance data, a distance metric can uncover hidden relationships between students. The proposed method demonstrated superior performance compared to traditional ML models across all datasets, with the Cosine and Euclidean norms producing the best results for classifying students as Good, At-Risk, or Failed. The addition of graph topological features further improved prediction accuracy, allowing for early identification of at-risk students. The outcomes of this study can benefit universities by enhancing their performance predictions and reputation. The model can monitor student performance, providing decision-makers and instructors with information about at-risk students and allowing them to take actions that may improve course outcomes.

Objective Four aimed to develop a more effective framework for ranking students, providing instructors with the tools they need to identify the factors affecting at-risk student performance and make recommendations for improvement. This objective builds upon the frameworks established as an outcome of the first three objectives to create a more precise solution for classifying at-risk students and offering personalized recommendations. The model was improved by incorporating elements, such as knowledge graphs and unsupervised learning techniques, that can further help instructors. This results in an efficient student ranking approach and an effective visualization approach for instructors to better understand the factors affecting student performance. Personalized learning is also utilized to provide customized recommendations for at-risk students to improve their course performance. In the fourth objective, we utilized knowledge graphs, clustering, and machine learning to analyze student performance data and demonstrate how the combination of these methods improves prediction accuracy and provides valuable insights that can help educators identify at-risk students and prevent course failure. The results of the k-means clustering indicate that incorporating KGE into historical data and checkpoints significantly increases clustering accuracy, highlighting the importance of KGE features in accurately clustering student performance.

Additionally, the clear visualization of the KG helps instructors quickly assess student performance and implement corrective measures as needed. The proposed model provides numerous benefits for instructors and educational institutions, including improved at-risk student classification and personalized recommendations for improved course-level performance. First, the at-risk student classification in Objective Four provides improved performance compared to state-of-the-art methods in ranking students and identifying at-risk students at different stages of course progression. Second, the use of the visualization tools developed as part of this objective gives instructors valuable insights for creating personalized learning experiences and making tailored recommendations to help at-risk students improve their course-level performance. This approach enhances learning and reduces the likelihood of students dropping out or failing their courses. As a result, educational institutions can maintain their reputation for academic excellence through improved student performance.

3.1 Research Strengths

The four objectives described in this dissertation were designed to facilitate the development of a model that could identify and predict at-risk students in a comprehensive manner. One of their key strengths is the creation of an early-warning solution that informs instructors and students about a student's risk status at different stages of the course. This allows instructors to take proactive measures to enhance learning and reduce the chances of course failure. Each of the four studies employs different approaches to create an early-warning solution, with Objective One using a rule-based approach and risk factor flag. The findings from Objective One provide a foundation for an effective at-risk student classification model using a rule-based approach and offer valuable insights that inform the development of more efficient and effective solutions in subsequent studies. Therefore, Objective One successfully achieved its goal of developing an at-risk student classification solution. Objective Two builds upon the foundations of Objective One by combining various ML-based and rule-based approaches for classifying at-risk students and implementing remediation strategies aimed at addressing and mitigating their weaknesses at each checkpoint throughout the course. The goals of Objective Two were successfully achieved, enhancing the functionality and performance of the solution developed in Objective One. In addition, objective Two achieved its goal of identifying at-risk students early on in the course and providing personalized remediation strategies to help them improve. One of its key strengths is the use of explainable machine learning-based models for classification, which enhances the solution's transparency and facilitates a better understanding of the decision-making process. In this study, the explainable ML models provide instructors with a transparent tool for determining the at-risk status of students. This is in contrast to traditional black-box-based ML models. The combination of explainable ML models and effective visualization approaches allowed the solution to identify the specific weaknesses in features, such as course-level and historical features, that led to a student being considered at-risk. In Objective Three, the use of ML-based models in conjunction with GCN-based and knowledge graph-based concepts enhances the feature acquisition for training the GCN model, resulting in improved performance compared to the use of ML-based models alone. The utilization of knowledge graph-based concepts allows for a more comprehensive understanding of the student's performance and risk factors, leading to a more accurate prediction. The primary goal of Objective Three was to improve the at-risk student prediction solutions developed in objectives one and two. This was achieved. The model improved its ability to correctly identify underperforming students in different courses by using topological features with GCN outperforming existing state-of-the-art frameworks. This advancement is a positive step towards developing a practical and usable solution. In Objective Four, a combination of ML-based models and clustering-based KGEs was used to predict student ranking. This objective effectively showcases the benefits and performance of different approaches to classifying at-risk students. The primary goal of Objective Four was to gain valuable insights from the first three objectives, which were used to develop a student ranking approach that could accurately distinguish top-performing and low-performing students. In Objective Four, the prediction of student ranking is performed using a combination of ML-based models and clustering-based knowledge graph embeddings (KGEs). The main objective of this objective was to leverage the insights gained from the previous three studies to develop a student ranking approach that accurately identifies top-performing and low-performing students. The approach uses unsupervised learning (K-means clustering) with KGEs and ML-based models, combining both historical and course-based features to provide improved predictions of at-risk students' final performance. The combination of clustering-based KGE features and ML-based features leads to an improvement in the accuracy of the student ranking approach, resulting in more accurate identification of at-risk students throughout the course progression. The use of K-means clustering provides effective visualization tools for instructors to view and inspect groups of at-risk students and develop personalized remedial strategies to prevent course failure, improve student performance, and promote student retention.

3.2 Research Limitations

It is crucial to note that the four objectives in this dissertation form a progressive improvement in the creation of an at-risk student identification solution. The approach follows a step-by-step process, where initial experiments are conducted to achieve the first objective and the limitations of the previous methods are addressed in the subsequent objectives to yield substantial advancements in terms of performance, effectiveness, and practical applicability.

In this dissertation, the findings from objectives One and Two were used to develop a more advanced solution in objectives three and four. This step-by-step approach allowed for a systematic improvement of the at-risk student identification solution. In this section, some of the limitations of the four studies reported in this dissertation will be discussed collectively. The first objective developed a rule-based model for at-risk students' classification; however, due to its preliminary nature, it did not provide information on model performance or remedial actions for at-risk students. In the case of Objective Two, a hybrid rule-based and ML-based approach was used with different ML models for the classification of at-risk students and the development of remediation strategies to prevent course failure. However, the second objective did not effectively highlight the impact these remediation strategies had on students' final grades, which is an important factor to consider. In Objective Three, knowledge graph-based topological features and GCN were utilized to develop an at-risk student classification solution. Lastly, Objective Four utilized clustering-based KGEs and ML-based models for the classification of at-risk students.

Chapter 4: Conclusion

4.1 Conclusion and Future Recommendations

Educational institutions are currently facing significant challenges with regard to student failure rates, which results in low retention rates. Early prediction of students' academic performance can play a crucial role in addressing these issues through appropriate interventions such as student counseling, development of intelligent tutoring systems, continuous performance monitoring, and policy-making to improve the academics of low-performing students and reduce the number of dropouts and late graduates. Despite advancements in technology, most online and offline learning platforms still lack reliable performance prediction modules. It is imperative for educational institutions to make accurate and timely predictions about student performance to improve their average grades and lower dropout rates. Furthermore, early assessment of student performance can aid in implementing strategies to address the causes of low performance at both the student and instructor levels. Effective visualization tools for instructors can also help identify various groups of at-risk students in different courses. Personalized learning can be used to develop targeted remediation strategies for these students at key points, allowing instructors to provide relevant and effective guidance to improve performance and decrease course failure rates, leading to higher long-term student retention rates. The use of personalized learning within technological tools, along with an efficient student ranking and at-risk student identification solution, allows instructors to customize course instruction and objectives to meet the unique learning needs of each at-risk student. In this dissertation, four separate studies were conducted to develop novel at-risk student classification solutions. These studies should not be viewed as isolated solutions with distinct goals but rather as progressive steps towards a common objective: To provide a solution for early detection and classification of at-risk students and provide instructors with tools to develop effective recommendations for improving student performance. The goal of these studies is to create an early warning solution for instructors and students in educational institutions to reduce course-level failure rates. In this dissertation, four studies were conducted to develop novel at-risk classification solutions. These studies should not be seen as separate solutions with separate objectives but as steps towards a unified goal of providing an early-warning solution for instructors and students in educational institutions to prevent course-level failure rates. The purpose of these objectives is to provide critical information at an early stage during the course progression so instructors can develop effective remediation strategies to help at-risk students succeed. The four studies use different methods and approaches, ranging from rule-based, hybrid (ML and rule-based) to deep learning frameworks, such as GCN, to achieve the same goal of early detection and classification of at-risk students. The achievement of the four objectives described in this dissertation involved experimenting with using different academic and non-academic features to train and validate models. The final feature set was optimized using various feature selection methods to eliminate redundant features. The feasibility of different machine learning-based approaches, such as SVM, XG Boosting classifier, LightGBM, Naive Bayes, ExtraTrees, Random Forest, MLP, and

Bagging models, was also tested for training and validation of the at-risk student classification solution. The frameworks developed during this research can potentially be improved in the future. For instance, future studies can explore using a knowledge graph-based approach to uncover patterns between students' online behaviors and activities on social media platforms. These insights could serve as additional features for classifying at-risk students alongside student performance. Additionally, students' academic history combined with current course-level performance can be effectively utilized to train and validate machine learning models for at-risk student classification. Once a functional and highly performing automated solution for at-risk student classification has been developed, it can be integrated into the official university grading system. This will allow students and instructors to be informed about at-risk students and proposed remediation strategies after each checkpoint. Researchers can also evaluate the practical feasibility of the system by interviewing instructors on its use and gathering their feedback on potential areas for improvement. Moreover, additional data can be collected for future studies aimed at developing robust and reliable systems for at-risk student classification. Future studies may aim to gather a large amount of student data and features from various courses, departments, and disciplines to better comprehend inter-disciplinary and inter-departmental differences, needs, and requirements for at-risk student assessment and remedial action development. Another area of future exploration is collecting feedback from instructors, either qualitative or quantitative, on the practical effectiveness and usefulness of various approaches for ranking students and developing effective remediation strategies. Improving performance towards developing accurate and intelligent student ranking approach and effective remediation strategies to enhance student performance is also a priority.

References

[1] C. F. Rodríguez-Hernández, M. Musso, E. Kyndt, and E. Cascallar, "Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100018, 2021.

[2] X. Chen, D. Zou, and H. Xie, "Fifty years of british journal of educational technology: A topic modeling based bibliometric perspective," *British Journal of Educational Technology*, vol. 51, no. 3, pp. 692–708, 2020.

[3] X. Chen, D. Zou, G. Cheng, and H. Xie, "Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of computers & education," *Computers & Education*, vol. 151, p. 103855, 2020.

[4] G.-J. Hwang, H. Xie, B. W. Wah, and D. Gašević, "Vision, challenges, roles and research issues of artificial intelligence in education," p. 100001, 2020.

[5] L. M. Barbosa Manhães, S. M. S. da Cruz, and G. Zimbrão, "Towards automatic prediction of student performance in stem undergraduate degree programs," in *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, 2015, pp. 247–253.

[6] ThinkImpact, "College dropout rates," 2022. [Online]. Available: https://www.thinkimpact.com/college-dropout-rates/

[7] A. Ettorre, F. Michel, and C. Faron, "Prediction of students' performance in e-learning environments based on link prediction in a knowledge graph," in *The 23rd International Conference on Artificial Intelligence in Education (AIED 2022)*, 2022.

[8] S. Kouki and M. Al Shemaili, "Science education in the united arab emirates," in *Science Education in Countries Along the Belt & Road*. Springer, 2022, pp. 225–241.

[9] R. Hasan, S. Palaniappan, S. Mahmood, A. Abbas, K. U. Sarker, and M. U. Sattar, "Predicting student performance in higher educational institutions using video learning analytics and data mining techniques," *Applied Sciences*, vol. 10, no. 11, p. 3894, 2020.

[10] A. Alshanqiti and A. Namoun, "Predicting student performance and its influential factors using hybrid regression and multi-label classification," *IEEE Access*, vol. 8, pp. 203 827–203 844, 2020.

[11] A. Ahadi, R. Lister, H. Haapala, and A. Vihavainen, "Exploring machine learning methods to automatically identify students in need of assistance," in *Proceedings of the eleventh annual international conference on international computing education research*, 2015, pp. 121–130.

[12] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S. N. Liao, "Predicting academic performance: a systematic literature review," in *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education*, 2018, pp. 175–199.

[13] H. Golino and C. M. A. Gomes, "Four machine learning methods to predict academic achievement of college students: a comparison study," *Revista E-Psi*, vol. 1, pp. 68–101, 2014.

[14] F. Widyahastuti and V. U. Tjhin, "Predicting students performance in final examination using linear regression and multilayer perceptron," in *2017 10th International Conference on Human System Interactions (HSI)*. IEEE, 2017, pp. 188–192.

[15] M. Kavitha, D. Raj *et al.*, "Educational data mining and learning analytics-educational assistance for teaching and learning," *arXiv preprint arXiv:1706.03327*, 2017.

[16] M. H. Abdi, G. Okeyo, and R. W. Mwangi, "Matrix factorization techniques for context-aware collaborative filtering recommender systems: A survey," 2018.

[17] A. Zohair and L. Mahmoud, “Prediction of student¡¦s performance by modelling small dataset size,” *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, pp. 1–18, 2019.

[18] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, “Analyzing and predicting students’ performance by means of machine learning: A review,” *Applied sciences*, vol. 10, no. 3, p. 1042, 2020.

[19] Z. Ahmad and E. Shahzadi, “Prediction of students’ academic performance using artificial neural network.” *Bulletin of Education and Research*, vol. 40, no. 3, pp. 157–164, 2018.

[20] E. Lau, L. Sun, and Q. Yang, “Modelling, prediction and classification of student academic performance using artificial neural networks,” *SN Applied Sciences*, vol. 1, no. 9, pp. 1–10, 2019.

[21] N. Zaki, E. A. Mohamed, and T. Habuza, “From tabulated data to knowledge graph: A novel way of improving the performance of the classification models in the healthcare data,” *medRxiv*, 2021.

[22] A. Namoun and A. Alshanqiti, “Predicting student performance using data mining and learning analytics techniques: A systematic literature review,” *Applied Sciences*, vol. 11, no. 1, p. 237, 2020.

[23] A. M. Shahiri, W. Husain *et al.*, “A review on predicting student’s performance using data mining techniques,” *Procedia Computer Science*, vol. 72, pp. 414–422, 2015.

[24] T. Anderson and R. Anderson, “Applications of machine learning to student grade prediction in quantitative business courses,” *Global Journal of Business Pedagogy*, vol. 1, no. 3, pp. 13–22, 2017.

[25] P. G. de Barba, G. E. Kennedy, and M. D. Ainley, “The role of students’ motivation and participation in predicting performance in a mooc,” *Journal of Computer Assisted Learning*, vol. 32, no. 3, pp. 218–231, 2016.

[26] E. Osmanbegovic and M. Suljic, “Data mining approach for predicting student performance,” *Economic Review: Journal of Economics and Business*, vol. 10, no. 1, pp. 3–12, 2012.

[27] B. K. Baradwaj and S. Pal, “Mining educational data to analyze students’ performance,” *arXiv preprint arXiv:1201.3417*, 2012.

[28] C. Márquez-Vera, A. Cano, C. Romero, and S. Ventura, “Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data,” *Applied intelligence*, vol. 38, no. 3, pp. 315–330, 2013.

[29] Y.-H. Hu, C.-L. Lo, and S.-P. Shih, “Developing early warning systems to predict students¡¦ online learning performance,” *Computers in Human Behavior*, vol. 36, pp. 469–478, 2014.

[30] H. Al-Shehri, A. Al-Qarni, L. Al-Saati, A. Batoaq, H. Badukhen, S. Alrashed, J. Alhiyafi, and S. O. Olatunji, “Student performance prediction using support vector machine and k-nearest neighbor,” in *2017 IEEE 30th canadian conference on electrical and computer engineering (CCECE).* IEEE, 2017, pp. 1–4.

[31] A. Acharya and D. Sinha, “Early prediction of students performance using machine learning techniques,” *International Journal of Computer Applications*, vol. 107, no. 1, 2014.

[32] S. B. Kotsiantis, “Use of machine learning techniques for educational proposes: a decision support system for forecasting students¡¦ grades,” *Artificial Intelligence Review*, vol. 37, no. 4, pp. 331–344, 2012.

[33] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison, “A machine learning framework to identify students at risk of adverse academic outcomes,” in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1909–1918.

[34] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran, “Machine learning based student grade prediction: A case study,” *arXiv preprint arXiv:1708.08744*, 2017.

[35] H. Almarabeh, “Analysis of students’ performance by using different data mining classifiers,” *International Journal of Modern Education and Computer Science*, vol. 9, no. 8, p. 9, 2017.

[36] H. Zeineddine, U. Braendle, and A. Farah, “Enhancing prediction of student success: Automated machine learning approach,” *Computers & Electrical Engineering*, vol. 89, p. 106903, 2021.

[37] S. Kotsiantis, K. Patriarcheas, and M. Xenos, “A combinational incremental ensemble of classifiers as a technique for predicting students¡¦ performance in distance education,” *Knowledge-Based Systems*, vol. 23, no. 6, pp. 529–535, 2010.

[38] D. Buenaño-Fernández, D. Gil, and S. Luján-Mora, “Application of machine learning in predicting performance for computer engineering students: A case study,” *Sustainability*, vol. 11, no. 10, p. 2833, 2019.

[39] F. Al-Obeidat, A. Tubaishat, A. Dillon, and B. Shah, “Analyzing students¡¦ performance using multi-criteria classification,” *Cluster Computing*, vol. 21, no. 1, pp. 623–632, 2018.

[40] B. Albreiki, N. Zaki, and H. Alashwal, “A systematic literature review of student¡¦performance prediction using machine learning techniques,” *Education Sciences*, vol. 11, no. 9, p. 552, 2021.

[41] J. Kuzilek, M. Hlosta, D. Herrmannova, Z. Zdrahal, J. Vaclavek, and A. Wolff, “Ou analyse: analysing at-risk students at the open university,” *Learning Analytics Review*, pp. 1–16, 2015.

[42] J. He, J. Bailey, B. Rubinstein, and R. Zhang, “Identifying at-risk students in massive open online courses,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[43] Z. Kovacic, “Early prediction of student success: Mining students’ enrolment data.” 2010.

[44] C. S. Galbraith, G. B. Merrill, and D. M. Kline, “Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? a neural network and bayesian analyses,” *Research in Higher Education*, vol. 53, no. 3, pp. 353–374, 2012.

[45] D. Aggarwal, S. Mittal, and V. Bali, “Significance of non-academic parameters for predicting student performance using ensemble

learning techniques," *International Journal of System Dynamics Applications (IJSDA)*, vol. 10, no. 3, pp. 38–49, 2021.

[46] W. Wang, H. Yu, and C. Miao, "Deep model for dropout prediction in moocs," in *Proceedings of the 2nd international conference on crowd science and engineering*, 2017, pp. 26–32.

[47] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," *arXiv preprint arXiv:1606.06364*, 2016.

[48] S. Nagrecha, J. Z. Dillon, and N. V. Chawla, "Mooc dropout prediction: lessons learned from making pipelines interpretable," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 351–359.

[49] J. Liang, C. Li, and L. Zheng, "Machine learning application in moocs: Dropout prediction," in *2016 11th International Conference on Computer Science & Education (ICCSE)*. IEEE, 2016, pp. 52–57.

[50] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, "Using machine learning to predict student difficulties from learning session data," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 381–407, 2019.

[51] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100066, 2022.

[52] H.-C. Chen, E. Prasetyo, S.-S. Tseng, K. T. Putra, S. S. Kusumawardani, and C.-E. Weng, "Week-wise student performance early prediction in virtual learning environment using a deep explainable artificial intelligence," *Applied Sciences*, vol. 12, no. 4, p. 1885, 2022.

[53] P. Guleria and M. Sood, "Explainable ai and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling," *Education and Information Technologies*, pp. 1–36, 2022.

[54] B. Albreiki, N. Zaki, and H. Alashwal, “A systematic literature review of studentperformance prediction using machine learning techniques,” *Education Sciences*, vol. 11, no. 9, p. 552, 2021.

[55] A. A. Mubarak, H. Cao, I. M. Hezam, and F. Hao, “Modeling students¡¦ performance using graph convolutional networks,” *Complex & Intelligent Systems*, pp. 1–19, 2022.

[56] R. Ferguson, “Learning analytics: drivers, developments and challenges,” *International Journal of Technology Enhanced Learning*, vol. 4, no. 5/6, pp. 304–317, 2012.

[57] M. Liu, S. Zha, and W. He, “Digital transformation challenges: A case study regarding the mooc development and operations at higher education institutions in china,” *TechTrends*, vol. 63, no. 5, pp. 621–630, 2019.

[58] A. A. Mubarak, H. Cao, W. Zhang, and W. Zhang, “Visual analytics of video-clickstream data and prediction of learners’ performance using deep learning models in moocs’ courses,” *Computer Applications in Engineering Education*, vol. 29, no. 4, pp. 710–732, 2021.

[59] H. Zhang, M. Sun, X. Wang, Z. Song, J. Tang, and J. Sun, “Smart jump: Automated navigation suggestion for videos in moocs,” in *Proceedings of the 26th international conference on world wide web companion*, 2017, pp. 331–339.

[60] 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*, pp. 1–20, 2020.

[61] S. Yang, M. Zhu, J. Hou, and X. Lu, “Deep knowledge tracing with convolutions,” *arXiv preprint arXiv:2008.01169*, 2020.

[62] A. Aleem and M. M. Gore, “Educational data mining methods: A survey,” in *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*. IEEE, 2020, pp. 182–188.

[63] S. K. Yadav, B. Bharadwaj, and S. Pal, "Mining education data to predict student's retention: a comparative study," *arXiv preprint arXiv:1203.2987*, 2012.

[64] D. K. Kolo and S. A. Adepoju, "A decision tree approach for predicting students academic performance," 2015.

[65] V. Dhanalakshmi, D. Bino, and A. M. Saravanan, "Opinion mining from student feedback data using supervised learning algorithms," in *2016 3rd MEC international conference on big data and smart city (ICBDSC)*. IEEE, 2016, pp. 1–5.

[66] J. Mesarić and D. Šebalj, "Decision trees for predicting the academic success of students," *Croatian Operational Research Review*, vol. 7, no. 2, pp. 367–388, 2016.

[67] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student engagement level in an e-learning environment: Clustering using k-means," *American Journal of Distance Education*, vol. 34, no. 2, pp. 137–156, 2020.

[68] Y. Jung and J. Lee, "Learning engagement and persistence in massive open online courses (moocs)," *Computers & Education*, vol. 122, pp. 9–22, 2018.

[69] H. Karimi, T. Derr, J. Huang, and J. Tang, "Online academic course performance prediction using relational graph convolutional neural network." *International Educational Data Mining Society*, 2020.

[70] M. Fei and D.-Y. Yeung, "Temporal models for predicting student dropout in massive open online courses," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2015, pp. 256–263.

[71] J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley, "Delving deeper into mooc student dropout prediction," *arXiv preprint arXiv:1702.06404*, 2017.

[72] W. Feng, J. Tang, and T. X. Liu, "Understanding dropouts in moocs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 517–524.

[73] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua, “Graph-based semi-supervised learning with multiple labels,” *Journal of Visual Communication and Image Representation*, vol. 20, no. 2, pp. 97–103, 2009.

[74] Z. Hu, X. Ma, Z. Liu, E. Hovy, and E. Xing, “Harnessing deep neural networks with logic rules,” *arXiv preprint arXiv:1603.06318*, 2016.

[75] M. Gaur, K. Faldu, and A. Sheth, “Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable?” *IEEE Internet Computing*, vol. 25, no. 1, pp. 51–59, 2021.

[76] Y. Zheng, R. Liu, and J. Hou, “The construction of high educational knowledge graph based on mooc,” in *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. IEEE, 2017, pp. 260–263.

[77] P. Chen, Y. Lu, V. W. Zheng, X. Chen, and B. Yang, “Knowedu: A system to construct knowledge graph for education,” *Ieee Access*, vol. 6, pp. 31 553–31 563, 2018.

[78] P. Chen, Y. Lu, V. W. Zheng, X. Chen, and X. Li, “An automatic knowledge graph construction system for k-12 education,” in *Proceedings of the fifth annual ACM conference on learning at scale*, 2018, pp. 1–4.

[79] Y. Su and Y. Zhang, “Automatic construction of subject knowledge graph based on educational big data,” in *Proceedings of the 2020 The 3rd International Conference on Big Data and Education*, 2020, pp. 30–36.

[80] F. Alghamdi and O. Siddiqui, “Supporting low-achieving efl learners: Expectations, procedure and significance of remedial sessions at a saudi university.” *Journal of Education and Training Studies*, vol. 4, no. 12, pp. 204–212, 2016.

[81] S. R. Marilyn Murphy and J. Twyman, “Handbook on personalized learning for states, districts and school,” *Center on Innovations in Learning, Temple University: Philadelphia, PA*, 2016.

[82] O. o. E. T. U.S. Department of Education, “Transforming american education: Learning powered by technology (national educational technology plan, 2010),” *Washington, DC: US Department of Education*, 2010.

[83] J. F. Pane, “Strategies for implementing personalized learning while evidence and resources are underdeveloped,” *Santa Monica, CA: RAND Corporation*, 2018.

[84] G. Natriello, “The adaptive learning landscape,” *Teachers College Record*, vol. 119, no. 3, pp. 1–46, 2017.

[85] L. Major, G. A. Francis, and M. Tsapali, “The effectiveness of technology-supported personalised learning in low-and middle-income countries: A meta-analysis,” *British Journal of Educational Technology*, vol. 52, no. 5, pp. 1935–1964, 2021.

[86] A. Maseleno, N. Sabani, M. Huda, R. B. Ahmad, K. A. Jasmi, and B. Basiron, “Demystifying learning analytics in personalised learning,” *International Journal of Engineering and Technology (UAE)*, 2018.

[87] J. F. Pane, E. D. Steiner, M. D. Baird, and L. S. Hamilton, “Continued progress: Promising evidence on personalized learning.” *Rand Corporation*, 2015.

[88] K. Hirsh-Pasek, J. M. Zosh, R. M. Golinkoff, J. H. Gray, M. B. Robb, and J. Kaufman, “Putting education in educational apps: Lessons from the science of learning,” *Psychological Science in the Public Interest*, vol. 16, no. 1, pp. 3–34, 2015.

[89] A. Pardo, J. Jovanovic, S. Dawson, D. Gašević, and N. Mirriahi, “Using learning analytics to scale the provision of personalised feedback,” *British Journal of Educational Technology*, vol. 50, no. 1, pp. 128–138, 2019.

[90] D. Rodriguez-Segura, “Educational technology in developing countries: A systematic review,” *University of Virginia EdPolicy Works Working Papers. https://curry. virginia.*

edu/sites/default/files/uploads/epw/72_Edtech_in_Developing_Countries.pdf, 2020.

[91] M. Escueta, V. Quan, A. J. Nickow, and P. Oreopoulos, “Education technology: An evidence-based review,” 2017.

[92] H. Xie, H.-C. Chu, G.-J. Hwang, and C.-C. Wang, “Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017,” *Computers & Education*, vol. 140, p. 103599, 2019.

[93] L. Zhang, J. D. Basham, and S. Yang, “Understanding the implementation of personalized learning: A research synthesis,” *Educational Research Review*, vol. 31, p. 100339, 2020.

[94] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme, “Recommender system for predicting student performance,” *Procedia Computer Science*, vol. 1, no. 2, pp. 2811–2819, 2010.

[95] C. Watson, F. W. Li, and J. L. Godwin, “Predicting performance in an introductory programming course by logging and analyzing student programming behavior,” in *2013 IEEE 13th international conference on advanced learning technologies*. IEEE, 2013, pp. 319–323.

[96] F. Ahmad, N. H. Ismail, and A. A. Aziz, “The prediction of students¡¦ academic performance using classification data mining techniques,” *Applied Mathematical Sciences*, vol. 9, no. 129, pp. 6415–6426, 2015.

[97] M. A. Al-Barrak and M. Al-Razgan, “Predicting students final gpa using decision trees: a case study,” *International journal of information and education technology*, vol. 6, no. 7, p. 528, 2016.

[98] A. Daud, N. R. Aljohani, R. A. Abbasi, M. D. Lytras, F. Abbas, and J. S. Alowibdi, “Predicting student performance using advanced learning analytics,” in *Proceedings of the 26th international conference on world wide web companion*, 2017, pp. 415–421.

[99] J. Berens, K. Schneider, S. Görtz, S. Oster, and J. Burghoff, “Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods,” *Available at SSRN 3275433*, 2018.

[100] C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, “Data mining for modeling students¡¦ performance: A tutoring action plan to prevent academic dropout,” *Computers & Electrical Engineering*, vol. 66, pp. 541–556, 2018.

[101] Y. Nieto, V. García-Díaz, C. Montenegro, and R. G. Crespo, “Supporting academic decision making at higher educational institutions using machine learning-based algorithms,” *Soft Computing*, vol. 23, no. 12, pp. 4145–4153, 2019.

[102] W. Xing and D. Du, “Dropout prediction in moocs: Using deep learning for personalized intervention,” *Journal of Educational Computing Research*, vol. 57, no. 3, pp. 547–570, 2019.

[103] K. T. Chui, D. C. L. Fung, M. D. Lytras, and T. M. Lam, “Predicting at-risk university students in a virtual learning environment via a machine learning algorithm,” *Computers in Human Behavior*, vol. 107, p. 105584, 2020.

[104] S. Alhusban, M. Shatnawi, M. B. Yasin, and I. Hmeidi, “Measuring and enhancing the performance of undergraduate student using machine learning tools,” in *2020 11th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2020, pp. 261–265.

[105] N. Walia, M. Kumar, N. Nayar, and G. Mehta, “Student¡¦s academic performance prediction in academic using data mining techniques,” in *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*, 2020.

[106] E. Wakelam, A. Jefferies, N. Davey, and Y. Sun, “The potential for student performance prediction in small cohorts with minimal available attributes,” *British Journal of Educational Technology*, vol. 51, no. 2, pp. 347–370, 2020.

[107] S. K. OuahiMariame *et al.*, “Feature engineering, mining for predicting student success based on interaction with the virtual learning environment using artificial neural network,” *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 6, pp. 12 734–12 746, 2021.

[108] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, "Early dropout prediction using data mining: a case study with high school students," *Expert Systems*, vol. 33, no. 1, pp. 107–124, 2016.

[109] S. U. R. Khan, M. A. ISLAM, M. Aleem, and M. A. Iqbal, "Temporal specificity-based text classification for information retrieval," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 26, no. 6, pp. 2915–2926, 2018.

[110] Z. Lei, Y. Sun, Y. A. Nanehkaran, S. Yang, M. S. Islam, H. Lei, and D. Zhang, "A novel data-driven robust framework based on machine learning and knowledge graph for disease classification," *Future Generation Computer Systems*, vol. 102, pp. 534–548, 2020.

[111] B. Albreiki, T. Habuza, Z. Shuqfa, M. A. Serhani, N. Zaki, and S. Harous, "Customized rule-based model to identify at-risk students and propose rational remedial actions," *Big Data and Cognitive Computing*, vol. 5, no. 4, p. 71, 2021.

[112] B. Albreiki, "Framework for automatically suggesting remedial actions to help students at risk based on explainable ml and rule-based models," *International Journal of Educational Technology in Higher Education*, vol. 19, no. 1, pp. 1–26, 2022.

[113] L.-M. S. Buenaño-Fernández D, Gil D, "Application of machine learning in predicting performance for computer engineering students: A case study," *Sustainability*, vol. 11, no. 10, p. 2833, 2019.

[114] I. Borrella, S. Caballero-Caballero, and E. Ponce-Cueto, "Taking action to reduce dropout in moocs: Tested interventions," *Computers Education*, vol. 179, no. 104, p. 412, 2022.

[115] K. Fahd, S. Venkatraman, S. J. Miah, and K. Ahmed, "Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature," *Education and Information Technologies*, pp. 1–33, 2021.

[116] E. Evangelista, "A hybrid machine learning framework for predicting students' performance in virtual learning environment," *International*

Journal of Emerging Technologies in Learning (iJET), vol. 16, no. 24, pp. 1–10, 2021.

[117] O. Iatrellis, I. K. Savvas, P. F. P, and V. C. Gerogiannis, "A two-phase machine learning approach for predicting student outcomes," *Education and Information Technologies*, vol. 26, no. 1, p. 69–88, 2021.

[118] D. T. Ha, P. T. T. Loan, C. N. Giap, and N. T. L. Huong, "An empirical study for student academic performance prediction using machine learning techniques," *International Journal of Computer Science and Information Security (IJCSIS*, vol. 18, no. 3, pp. 1–11, 2020.

[119] N. Tomasevic, N. G. N, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Computers education*, vol. 143:103, no. 103, p. 676, 2020.

[120] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," *Advances in neural information processing systems*, vol. 26, 2013.

[121] M. Nickel, L. Rosasco, and T. Poggio, "Holographic embeddings of knowledge graphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[122] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *International conference on machine learning*. PMLR, 2016, pp. 2071–2080.

[123] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *arXiv preprint arXiv:1709.05584*, 2017.

[124] N. Bhatti, A. Hanbury, and J. Stottinger, "Contextual local primitives for binary patent image retrieval," *Multimedia Tools and Applications*, vol. 77, no. 7, pp. 9111–9151, 2018.

[125] T. Dey, S. Mandal, and W. Varcho, “Improved image classification using topological persistence,” in *Proceedings of the conference on Vision, Modeling and Visualization*, 2017, pp. 161–168.

[126] C. Hofer, R. Kwitt, M. Niethammer, and A. Uhl, “Deep learning with topological signatures,” *Advances in neural information processing systems*, vol. 30, 2017.

[127] N. Zaki, E. A. Mohamed, and T. Habuza, “From tabulated data to knowledge graph: A novel way of improving the performance of the classification models in the healthcare data,” *medRxiv*, 2021.

List of Other Publications

Journal Publications

1. Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student’performance prediction using machine learning techniques. Education Sciences, 11(9), 552. https://doi.org/10.3390/educsci11090552
2. Al Braiki, B., Harous, S., Zaki, N., & Alnajjar, F. (2020). Artificial intelligence in education and assessment methods. Bulletin of Electrical Engineering and Informatics, 9(5), 1998-2007.https://www.beei.org/index.php/EEI/article/view/1984

Conference Proceedings Papers

3. Albreiki, B.; Habuza, T.; Zaki, N. Identify Students at Risk Using Graph Representation and Conventional ML Methods. Proceedings of The 5th World Conference on Future of Education, 15th August, 2022, Stockholm - Sweden. urlhttps://www.dpublication.com/abstract-of-5th-wcfeducation/66-4666/
4. Al Breiki, B., Zaki, N., & Mohamed, E. A. (2019, November). Using educational data mining techniques to predict student performance. In 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA) (pp. 1-5). IEEE. https://ieeexplore.ieee.org/abstract/document/8959676

UAE UNIVERSITY DOCTORATE DISSERTATION NO. 2023: 2

Academic underachievement, graduation delays, and student dropouts remain common problems in educational institutions. This dissertation addresses the challenges and factors associated with predicting student performance during the early stages of their courses. It presents a solution to help instructors to track and monitor student performance, identify at-risk students early in the course progression, improve student performance prediction, provide personalized interventions.

Balqis Albreiki received her PhD from the Department of Computer Science and Software Engineering, College of Information Technology at UAE University, UAE. She received her MS in Engineering Management from the College Engineering, UAE University, UAE.