

هالة حجار

Enhance student performance .Marya%5ELLLHala 22 AUG %283%29.docx

 Midocean University

Document Details

Submission ID

trn:oid:::2945:409008263

Submission Date

Aug 22, 2026, 10:25 AM GMT+3

Download Date

Aug 22, 2026, 11:19 AM GMT+3

File Name

Enhance student performance .Marya^LLHala 22 AUG (3).docx

File Size

6.5 MB

44 Pages

11,710 Words

68,973 Characters

23% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Match Groups

-  **231 Not Cited or Quoted 23%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 17%  Internet sources
- 14%  Publications
- 16%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 231 Not Cited or Quoted 23%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 17% Internet sources
- 14% Publications
- 16% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	mdpi-res.com	2%
2	Internet	www.mdpi.com	1%
3	Internet	louisdl.louislibraries.org	<1%
4	Internet	ojs.ninetyjournal.com	<1%
5	Submitted works	University of Wales Institute, Cardiff on 2026-08-01	<1%
6	Internet	rchubpublisher.com	<1%
7	Internet	arxiv.org	<1%
8	Internet	export.arxiv.org	<1%
9	Submitted works	Midocean University on 2025-09-05	<1%
10	Submitted works	National Economics University on 2026-05-30	<1%

11	Internet	qspace.qu.edu.qa	<1%
12	Internet	link.springer.com	<1%
13	Internet	pmc.ncbi.nlm.nih.gov	<1%
14	Publication	Mustapha Skittou, Mohamed Merrouchi, Taoufiq Gadi. "Development of an Early ...	<1%
15	Internet	mental.jmir.org	<1%
16	Internet	livrepository.liverpool.ac.uk	<1%
17	Internet	pdffox.com	<1%
18	Publication	Narayanan, Usha. "Patient Behavior Prediction Using Big Data Analytics and Mac...	<1%
19	Internet	www.researchsquare.com	<1%
20	Publication	Atolagbe, Babatunde. "Sensitivity Analyses and Meta-Modeling in Complex Decisi...	<1%
21	Publication	Wei Cong, Shaohua Zhang, Kai Yang, Pan Li. "Blended Learning Behaviors Analysi...	<1%
22	Publication	Pawan Singh, Sudesh Kumar, Sachin Kumar Gupta, Abhay Kumar Rai, Abdu Saif. "...	<1%
23	Internet	hdprediction.commonsc.gc.cuny.edu	<1%
24	Internet	spectrum.library.concordia.ca	<1%

25	Internet	www.kaggie.com	<1%
26	Submitted works	Karunya University on 2025-07-07	<1%
27	Submitted works	University of Birmingham on 2025-06-27	<1%
28	Submitted works	University of Hertfordshire on 2025-01-06	<1%
29	Internet	ijece.iaescore.com	<1%
30	Internet	ijrpr.com	<1%
31	Internet	pubs.thesciencein.org	<1%
32	Internet	slejournal.springeropen.com	<1%
33	Submitted works	Southampton Solent University on 2024-09-02	<1%
34	Submitted works	Universiti Tun Hussein Onn Malaysia on 2026-07-20	<1%
35	Submitted works	University of Essex on 2026-08-21	<1%
36	Submitted works	University of Newcastle on 2023-08-28	<1%
37	Internet	dirros.openscience.si	<1%
38	Internet	josre.shmpublisher.com	<1%

39	Internet	dokumen.pub	<1%
40	Submitted works	Udayana University on 2026-07-24	<1%
41	Internet	par.nsf.gov	<1%
42	Internet	www.medrxiv.org	<1%
43	Submitted works	Malayan Colleges Mindanao, A Mapua School on 2025-03-09	<1%
44	Submitted works	South Dakota Board of Regents on 2024-12-24	<1%
45	Internet	ijds.mtu.edu.iq	<1%
46	Internet	www.tandfonline.com	<1%
47	Submitted works	University of Hong Kong on 2026-07-28	<1%
48	Publication	Yueming Hu, Abhnil Prasad, Merlinde Kay. "Smart integration of aerosols for DNI ...	<1%
49	Internet	elar.urfu.ru	<1%
50	Internet	mfuir.mfu.ac.th	<1%
51	Internet	pure.rug.nl	<1%
52	Submitted works	University of Sheffield on 2019-09-11	<1%

53	Submitted works	University of Ulster on 2026-08-14	<1%
54	Internet	hellanicus.lib.aegean.gr	<1%
55	Internet	impactfactor.org	<1%
56	Internet	sjuo2.uoz.edu.krd	<1%
57	Internet	www.preprints.org	<1%
58	Submitted works	Liverpool John Moores University on 2022-09-13	<1%
59	Publication	Manzoor Ansari, Syed Arshad Ali, Mansaf Alam. "Next-Generation Artificial Intelli...	<1%
60	Publication	Poonam Nandal, Tapas Kumar, Meeta Singh, Mamta Dahiya. "Progressive Compu...	<1%
61	Submitted works	Thakur College of Engineering and Technology (TCET) on 2026-08-01	<1%
62	Submitted works	University of Huddersfield on 2024-06-09	<1%
63	Internet	123dok.org	<1%
64	Submitted works	Gisma University of Applied Sciences GmbH on 2026-06-26	<1%
65	Submitted works	University of Sydney on 2025-10-25	<1%
66	Submitted works	University of Witwatersrand on 2026-08-17	<1%

67	Internet	shdl.mmu.edu.my	<1%
68	Submitted works	Addis Ababa University on 2026-06-15	<1%
69	Submitted works	College of Public Health on 2026-03-09	<1%
70	Submitted works	College of Public Health on 2026-03-17	<1%
71	Submitted works	Malta College of Arts,Science and Technology on 2026-06-05	<1%
72	Submitted works	Moodle2025 on 2026-08-18	<1%
73	Publication	Salma Boujmiraz, Hassane Darhmaoui, Ahmed Drissi el maliani. "Predicting stude...	<1%
74	Publication	Sukhpreet Kaur, Amanpreet Kaur, Manish Kumar. "Recent Advances in Computati...	<1%
75	Submitted works	The University of the West of Scotland on 2026-08-16	<1%
76	Submitted works	Universiti Malaysia Perlis on 2015-06-23	<1%
77	Submitted works	Universiti Sains Malaysia on 2025-07-29	<1%
78	Submitted works	University of Northumbria at Newcastle on 2023-05-18	<1%
79	Submitted works	University of Queensland on 2025-11-10	<1%
80	Submitted works	University of Warwick on 2026-02-09	<1%

81	Publication	"Innovations in Sustainable Sports Infrastructure for Eco-Friendly Design and Op...	<1%
82	Publication	"Intelligent Data Engineering and Automated Learning — IDEAL 2000. Data Minin...	<1%
83	Publication	Baghbanzadeh, Mahdi. "Machine Learning Approaches for Genomic Sequence An...	<1%
84	Publication	Domingues, Gonalo. "REASONING - Protein Secondary Structure Prediction Usin...	<1%
85	Publication	Hamad Almaghrabi, Ben Soh, Alice Li, Idrees Alsolbi. "SoK: The Impact of Educatio...	<1%
86	Submitted works	Kingston University on 2024-08-20	<1%
87	Submitted works	Liverpool John Moores University on 2025-02-28	<1%
88	Publication	Romero-Zaldivar, V.A.. "Monitoring student progress using virtual appliances: A c...	<1%
89	Publication	Sameer Kumar Singh. "ChemSafeAI+: A Machine Learning Driven Dynamic Safety ...	<1%
90	Publication	Sondius Matogu Budiman Silalahi, Richardus Indrajit, Teddy Mantoro. "Developm...	<1%
91	Submitted works	University of Bolton on 2026-01-12	<1%
92	Submitted works	University of Moratuwa on 2026-04-09	<1%
93	Submitted works	University of Surrey on 2025-09-04	<1%
94	Submitted works	University of Ulster on 2026-08-12	<1%

95	Submitted works	University of Westminster on 2026-07-02	<1%
96	Submitted works	VIT University on 2025-03-19	<1%
97	Publication	Yonglin Zhang, Tianle Sun, Li Wang, Bo Huang et al. "Portraying on-road CO2 con...	<1%
98	Internet	achieverpapers.com	<1%
99	Internet	d-nb.info	<1%
100	Internet	ejournal.techcart-press.com	<1%
101	Internet	elibrary-dev.nusamandiri.ac.id	<1%
102	Internet	files.core.ac.uk	<1%
103	Internet	spatial-data-mining-winter2025.readthedocs.io	<1%
104	Publication	"Recent Advances on Soft Computing and Data Mining", Springer Science and Bus...	<1%
105	Publication	Amr Assad. "AI Educational Support Framework for At-Risk Media Students", 2025...	<1%
106	Submitted works	Colorado Technical University Online on 2026-08-02	<1%
107	Publication	Ghareebaih, Yaman Abu. "Car Price Prediction Using Machine Learning: Analyzin...	<1%
108	Publication	Jerina Eda, Tamador Mohaidat, Samiul Islam Niloy, Zhiqi Niu, Md rahat Kader Kha...	<1%

109	Publication	Kaur, Davinder. "Trustworthy AI: Ensuring Explainability and Acceptance", Purdu...	<1%
110	Publication	Lilibeth Azuela, Herbert Azuela, Anna Lyza Felipe, Armin Aragoncillo, Mia Eleazar. ...	<1%
111	Submitted works	Liverpool John Moores University on 2022-02-14	<1%
112	Publication	Nida Aslam. "Explainable Artificial Intelligence Approach for the Early Prediction ...	<1%
113	Publication	Pramod Singh Rathore, Sunil Kumar. "Smart and Distributed Computing", CRC Pr...	<1%
114	Publication	Praveer Dubey, Mohit kumar. "Personalized Diabetes Decision Support Using an E...	<1%
115	Publication	Ranjini Surendran, Anitha J, Jude D. Hemanth. "Recognition of human action for s...	<1%
116	Publication	Reda A. Zayed, Lamiaa Fattouh Ibrahim, Hesham A. Hefny, Hesham A. Salman, Ab...	<1%
117	Submitted works	Taylor's Education Group on 2026-07-28	<1%
118	Submitted works	University of Bedfordshire on 2022-05-20	<1%
119	Submitted works	University of East London on 2026-05-07	<1%
120	Submitted works	University of Hull on 2025-04-24	<1%
121	Submitted works	University of Malaya on 2026-07-02	<1%
122	Publication	V. Sekhar, P. Ajay Kumar Reddy, K. Logesh, T. Raghunadha Reddy, Dara Raju. "Ne...	<1%

123	Internet	dal.scholaris.ca	<1%
124	Internet	dspace.daffodilvarsity.edu.bd:8080	<1%
125	Internet	jurnal.polbeng.ac.id	<1%
126	Internet	kth.diva-portal.org	<1%
127	Internet	prism.ucalgary.ca	<1%
128	Internet	repository.aaup.edu	<1%
129	Internet	repository.cmu.edu	<1%
130	Internet	web.realinfo.tv	<1%
131	Internet	www.gepsoft.com	<1%
132	Internet	www.ijraset.com	<1%
133	Publication	Ángel Hernández-García, Carlos Salcedo-Valverde, Carlos Cuenca-Enrique, Laura ...	<1%
134	Publication	Abdulaziz Almohimeed, Redhwan M. A. Saad, Sherif Mostafa, Nora El-Rashidy et a...	<1%
135	Publication	Balqis Albreiki, Zahiriddin Rustamov, Jaloliddin Rustamov, Nazar Zaki. "Chapter 1...	<1%
136	Publication	Bo Xie. "Statistical Feature Selection for Mandarin Speech Emotion Recognition", ...	<1%

137	Publication	C. Guedes Soares, F.C. Salvado. "Innovations in Maritime Technology and Enginee...	<1%
138	Publication	Dothang Truong. "Demystifying AI - Data Science and Machine Learning Using IB...	<1%
139	Publication	Hassiba, Khaled Jamal Eldin. "Ann Framework for Bearing Fault Classification.", H...	<1%
140	Submitted works	Heriot-Watt University on 2025-08-17	<1%
141	Submitted works	Higher Education Commission Pakistan on 2025-10-15	<1%
142	Submitted works	King's College on 2023-08-08	<1%
143	Publication	Lecture Notes in Computer Science, 2015.	<1%
144	Publication	Nikhila Kathirisetty, Rajendrasinh Jadeja, Hiren Kumar Thakkar, Deepak Garg et a...	<1%
145	Publication	Ossama H. Embarak, Shatha Hawarna. "Automated AI-driven System for Early De...	<1%
146	Publication	Oumaima Saidani, Muhammad Umer, Amal Alshardan, Nazik Alturki, Michele Na...	<1%
147	Publication	Phauk Sökkhey, Takeo Okazaki. "Developing Web-based Support Systems for Pre...	<1%
148	Publication	Salha Al-Ahmari, Farrukh Nadeem. "Improving Surgical Site Infection Prediction ...	<1%
149	Publication	Samuel-Soma M. Ajibade, Ivy Batican, Lisle Valle, Edralin General et al. "Early Pre...	<1%
150	Publication	Saurav Mallik, Soumita Seth, Ben Othman Soufiene. "HealthTech Horizons - Chart...	<1%

151	Publication	Shasanka Sekhar Rout, Akash Saxena, Suman Lata Tripathi. "Smart Electronics De...	<1%
152	Submitted works	Teesside University (Blackboard Ultra) on 2026-05-01	<1%
153	Submitted works	The African Institute for Mathematical Sciences Cameroon (AIMS Cameroon) on 2...	<1%
154	Submitted works	Universiti Tenaga Nasional on 2022-08-15	<1%
155	Submitted works	University of Essex on 2026-08-21	<1%
156	Submitted works	University of Essex on 2026-08-21	<1%
157	Submitted works	University of Hertfordshire on 2025-01-06	<1%
158	Submitted works	University of Portsmouth on 2025-09-17	<1%
159	Submitted works	University of Portsmouth on 2026-05-15	<1%
160	Submitted works	University of Ruhuna Matara on 2024-06-03	<1%
161	Submitted works	University of Ulster on 2026-08-17	<1%
162	Submitted works	University of Wales, Lampeter on 2026-05-13	<1%
163	Publication	Xianhuang Li, Weixiang Liu, Ruoyun Hu, Yiyi Liu, Tong Wang, Haojun Chen, Ziyi Li,...	<1%
164	Publication	Yiwei Lin, Jingyu Li, Kaiqiang Fan, Tianying Qiu, Zhichao Li, Shujie Wang, Chuan G...	<1%

165	Internet	ecdpm.org	<1%
166	Internet	ejournal.pnc.ac.id	<1%
167	Internet	github.com	<1%
168	Internet	iieta.org	<1%
169	Internet	journal.perbanas.id	<1%
170	Internet	mail.mattioli1885journals.com	<1%
171	Internet	nur.nu.edu.kz	<1%
172	Internet	www.eriesjournal.com	<1%
173	Internet	www.frontiersin.org	<1%
174	Internet	www.iariajournals.org	<1%
175	Internet	www.inderscience.com	<1%



Enhance student performance using ensemble learning and feature selection in Saudi Arabia

Prepared By

Student 1: Marya Mohammed Bawazir

ID: 253000692

Student 2: Hala Haytham Hajjar

ID: 253001456

Submitted in fulfillment of the requirements for the degree Master of Science in
Artificial Intelligence

At

College of Informatics
Midocean University
UAE

Supervisor Name: **Hager Saleh.**

College of Informatics, Midocean University, Comoros

Abstract:

The widespread adoption of Learning Management Systems (LMSs) has generated large volumes of educational data that can be utilized to identify academically at-risk students at an early stage. This study proposes an AI-based Early Warning framework for predicting students' academic performance using a private educational dataset collected from an LMS in Saudi Arabia. The main objectives are to investigate the effects of feature selection and class balancing on predictive performance, compare traditional and ensemble ML models, and identify the most effective model for early academic risk prediction. The proposed framework includes data preprocessing, stratified train-test splitting, feature selection using RF feature importance, class balancing using SMOTE, hyperparameter optimization using Grid Search with Stratified Cross-Validation, and model evaluation. The dataset was split into 80% training and 20% testing sets using stratified sampling. Feature selection and SMOTE were applied only to the training data to prevent data leakage. The dataset contains 1,395 student records described by 17 academic and behavioral features. The original prediction task was transformed into a four-class classification problem comprising Excellent, Good, Pass, and Fail. Eight supervised ML models were evaluated under four experimental configurations using full and selected feature sets, with and without SMOTE. The results showed that XGBoost consistently achieved the best predictive performance, attaining an accuracy of 82.80%, precision of 83%, recall of 82.80%, and an F1-score of 82.46%. Feature selection reduced model complexity while maintaining competitive performance, whereas SMOTE improved class balance and supported the recognition of minority classes. These findings demonstrate the potential of the proposed framework to support early identification and timely intervention for academically at-risk students within the Saudi educational context.

Keywords: Educational Data Mining, SMOTE, Feature Selection, Student Performance Prediction, Machine Learning.

Table of Contents

List of Abbreviations..... xiv

List of Figures..... xvi

List of Tables..... xviii

Table of Contents

Chapter 1 Introduction	6
1.1 Problem Definition	6
1.2 Thesis Objectives.....	6
1.3 Thesis Contributions	7
1.4 Thesis Organization	8
Chapter 2: Background and Related Work.....	9
2.1 Background	9
2.2 Related Work:	16
Chapter 3	23
Methodology and Proposed Approach.....	23
Introduction.....	23
3.1 Dataset Description.....	24
3.2 Data Preprocessing	25
3.3 Data Splitting	25
3.4 Handling imbalance dataset using SMOTE:	25
3.5 Feature Selection	25
3.6 Optimization Models.....	26
3.7 Machine Learning Models.....	27
3.8 Ensemble Learning Models	28
3.9 Model Evaluation Metrics	29
Chapter 4 Results.	31
4.1 Experimental Setup	31

4.2 Class Distribution	31
4.3 Feature Importance Analysis	31
4.4 Experimental Configurations	32
4.5 Results without Applying SMOTE	32
4.6 Results with Applying SMOTE.....	35
Chapter 5 Conclusion and Future Work	40
5.1 Conclusion	40
5.2 Future Work	40
References:.....	42

List of Abbreviations :

- **Acc.:** Accuracy.
- **ANN:** Artificial Neural Network.
- **ML:** Machine Learning.
- **EL:** Ensemble Learning.
- **XAI:** Explainable Artificial Intelligence.
- **SVM:** Support Vector Machine.
- **RF:** Random Forest.
- **DT:** Decision Tree.
- **KNN:** K-Nearest Neighbors.
- **DL:** Deep Learning.
- **NV:** Naïve Bayes.
- **CNN:** Convolutional Neural Network.
- **LMS:** Learning Management System.
- **LR:** Logistic regression.
- **XGB:** Extreme Gradient Boosting.
- **SMOTE:** Synthetic Minority Oversampling Technique.

List of Figures :

Figure 1. AI [2]	9
Figure 2. Types of ML [4]	10
Figure 3. DT structure[7]	11

Figure 4 .RF algorithm structure.[9]	11
Figure 5 .KNN classification[6]	12
Figure 6.SVM [3].	12
Figure 7. The proposed framework.....	24
Figure 8. Distribution of Student Performance Classes with SMOTE.	32
Figure 9. Random Forest Feature Importance Ranking.	32
Figure 10 Confusion Matrices of the Models without SMOTE-full features.	34
Figure 11 Confusion Matrices of the Models without SMOTE-selected features.	34
Figure 12 Confusion Matrices of the Models with SMOTE-full feature.....	36
Figure 13 Confusion Matrices of the Models with SMOTE-selected feature.	37
Figure 14 Accuracy comparison of Optimized ML models using full and selected feature sets with and without	38
Figure 15. Class-wise Performance of the XGBoost Model with and without SMOTE	39

List of Tables:

Table 1 Of The Reviewed Papers	21
Table 2. Distribution of Student Performance Classes.	24
Table 3 Description of the 17 Educational Features	24
Table 4. Hyperparameter Search Space Used for Grid Search Optimization.....	27
Table 5 Results of applying models using full and selected features without SMOTE.....	33
Table 6 Results of applying models using full and selected features with SMOTE.	35

Chapter 1 Introduction

1.1 Problem Definition

Despite the increasing availability of student data in educational sectors, many systems still face challenges in identifying students at academic risk at an early stage to allow them to intervene effectively. Traditional evaluation methods often rely on certain and usual results. This reduces the opportunity to provide timely support or implement appropriate personalized intervention plans. The lack of this diagnostic plan is an additional risk, as each student's skills, goals, and needs vary, necessitating early diagnosis to provide individualized support.

Also, educational data is very complex and diverse. It includes demographic information, behavioral data, attendance records, and interactions within LMS. Such complexity makes it difficult to figure out which factors have the biggest effect on predicting academic performance. This data is also often enormous and challenging to understand, which makes it even harder to analyze and make decisions.

Besides, the presence of extraneous or less important features may lead to various difficulties in constructing prediction models, including overfitting, higher computation costs, and non-explainability of the models. Consequently, the use of feature selection techniques to extract only the most relevant features will have great potential for constructing accurate and explainable models.

In contrast, artificial intelligence (AI), such as machine learning (ML) and deep learning (DL), can analyze information from the education field, recognize patterns, detect students' behaviors, and predict student performance. With the help of a proper dataset, it will be possible to determine whether a student will probably perform well or not and what factors may cause their failure. However, the effectiveness of AI algorithms depends on data quality, feature selection, and model interpretability.

Therefore, the application of a data-driven early warning system based on AI is necessary for predicting and identifying at-risk students at an early stage. In particular, a data-driven model can help classify at-risk students (Excellent, Good, Pass, and Fail).

1.2 Thesis Objectives

The major goal of the research is the development of an early warning framework that can be applied to classify students' risk levels in terms of their academic performance by using AI and ML methods. The project will rely on LMS data, including students' academic performance indicators, engagement, and demographics, to identify those who have risks of poor performance as early as possible. For this purpose, the target (Final Mark) is transformed into a classification problem (Excellent, Good, Pass, and Fail). The research

uses the technique of data pre-processing and feature engineering to increase the quality of the data used for predictions. Several ML classifiers are implemented in a stacking ensemble model to make more accurate predictions. Accuracy, precision, and recall measures are used to evaluate the algorithm, where special attention will be paid to recall in order to increase the chances of detecting at-risk students.

1.3 Thesis Contributions

This thesis contributes in the following steps:

1. Developed an AI-Based Early Warning Framework for predicting students' academic performance using a private Learning Management System (LMS) dataset collected in Saudi Arabia. The proposed framework enables the early identification of academically at-risk students and supports timely educational interventions.
2. Conducted a thorough comparative analysis of eight supervised machine learning algorithms, including six traditional machine learning algorithms (DT, RF, LR, SVM, KNN, and NB) and two ensemble learning algorithms (AB and XGB) to select the most effective predictive algorithm.
3. Design and evaluate four different experimental configurations to examine the effect of both Feature Selection and SMOTE to explore their individual and combined influence on prediction accuracy.
4. Enhance the quality of data and improve prediction performance via a preprocessing pipeline including data transformation, stratified train-test splitting, Feature Selection based on the Random Forest importance of features, and class balancing using SMOTE. Preprocessing techniques allowed decreasing the dimensionality of data, handling class imbalance problems, and improving overall prediction performance.
5. Improve classifier performance via hyperparameter tuning with the Grid Search method with cross-validation. The technique was proven to be effective in improving the performance of several classifiers, especially DT and RF. It was also shown that the high predictiveness of the model was stable for XGB.
6. Select the most reliable prediction model by obtaining the best classification results using the XGB model for all experimental configurations. Despite the fact that the Decision Tree performed better than XGB for one experimental configuration regarding classification accuracy after hyperparameter tuning, it is a more stable and robust model. Therefore, the XGB model was chosen as the final one.
7. Proposed a practical framework that will help educational institutions detect academically at-risk students to conduct effective educational interventions.

1.4 Thesis Organization

The remainder of this thesis consists of as follows:

- **Chapter 2** includes a literature review and previous work associated with student performance prediction, educational data mining, machine learning algorithms, feature selection, class imbalance, and AI-based early warning frameworks.
- **Chapter 3** is devoted to the proposed methodology that involves the educational dataset, data pre-processing steps, feature selection procedure, experimental settings, machine learning models, hyperparameter tuning, and the proposed AI-Based Early Warning Framework approach.
- **Chapter 4** covers the experimental results and gives a detailed performance analysis of the proposed machine learning models within the four experimental settings. It compares the efficiency of the analyzed models with classification metrics and finds the most efficient prediction model.
- **Chapter 5** provides the major findings and contributions of the conducted research, the conclusions of the study, the limitations of the research, and the further research recommendations.

Chapter 2: Background and Related Work

2.1 Background

2.1.1 AI

The ability of computer systems to do tasks like learning, reasoning, decision-making, and problem-solving that require human intelligence is known as AI. Systems analyze massive data using computational methods to produce predictions. AI is used more and more in the field of education to evaluate student data and assist with decision-making processes like identifying students who are at risk and enhancing learning outcomes[1].

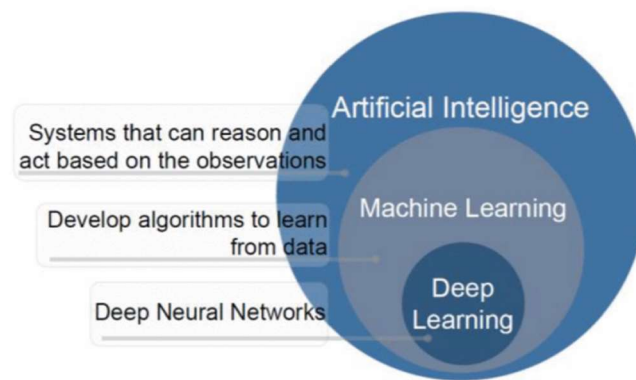


Figure 1. AI [2]

2.1.2 ML

It is a branch of AI that focuses on creating algorithms that, without explicit programming, enable computers to recognize patterns in data and perform better. It can be used to predict or classify new data after being trained on historical data. Because it allows for examining massive amounts of data of students and finding their performance, ML is crucial to educational data mining and learning analytics. (Pattern Recognition and ML)[3].

2.1.3 Types of ML

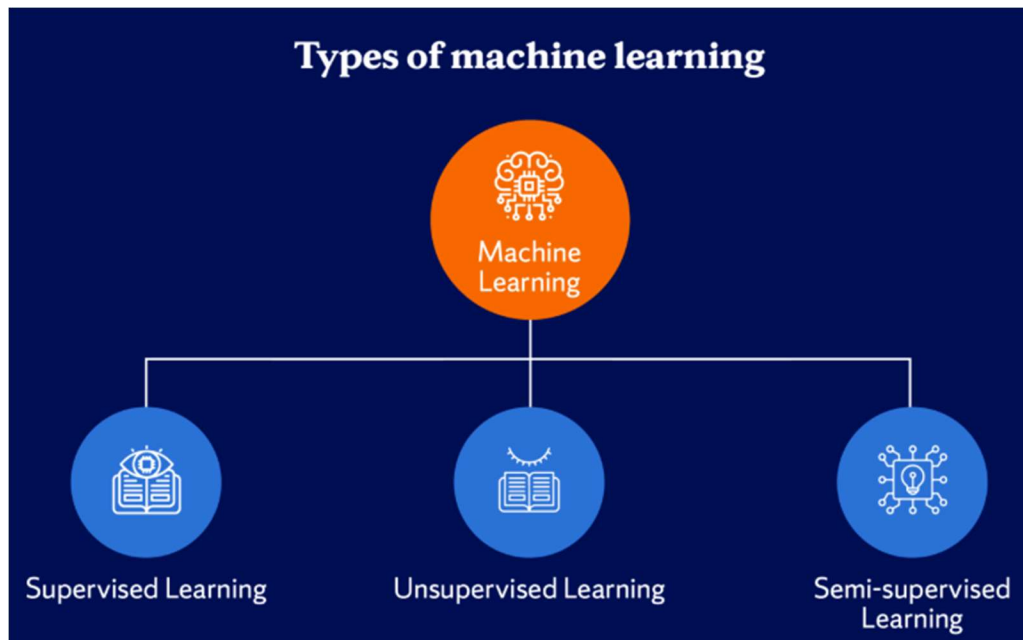


Figure 2. Types of ML [4]

1. Supervised Learning

The model is trained using labeled data in this ML method. There is an input and an output label for every training example. Learning a mapping between inputs and outputs is the aim of supervised learning, which enables the model to correctly predict the output for fresh, unseen data[4].

2. Unsupervised Learning

It uses unlabeled data to train models. In this kind of learning, the algorithm looks for unidentified patterns or structures in the data. In unsupervised learning, clustering and association rule learning are popular methods[3] .

3. Semi-Supervised Learning

It includes both labeled and unlabeled data. This method works well when there is a lot of unlabeled data but little labeled data. By utilizing both kinds of data, it can enhance model performance[4].

2.1.4 Classification

Predicting a categorical label for a given input is the goal of this kind of supervised learning task. For instance, classification algorithms can be used in educational systems to assess a student's risk level based on behavioral and academic data[5].

2.1.5 ML Algorithms

1. DT

It is a supervised learning algorithm that uses a tree structure to represent decisions. A feature is represented by each internal node, a decision rule by each branch, and a class label by each leaf node. DT are frequently used in classification problems because they are simple to understand [6].

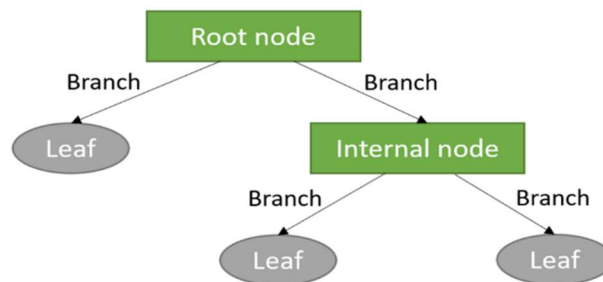


Figure 3. DT structure[7]

2. RF

Many DT are combined in the RF group learning technique to increase prediction accuracy and decrease overfitting. A subset of the dataset is used to train each tree in the forest, and the final prediction is produced by adding up all of the trees' predictions [6].

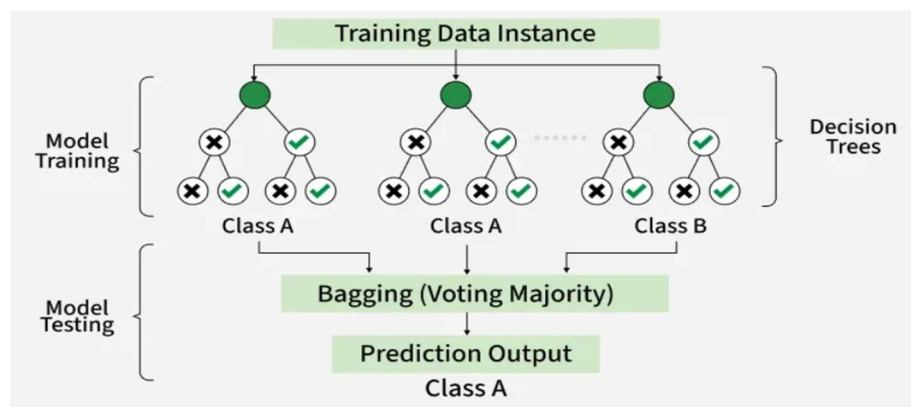


Figure 4 .RF algorithm structure.[9]

3. KNN

KNN uses the classes of a data point's closest neighbors in the feature space to predict the data point's class. It uses Euclidean distance, and other distance metrics are used by the algorithm to determine similarity [3].

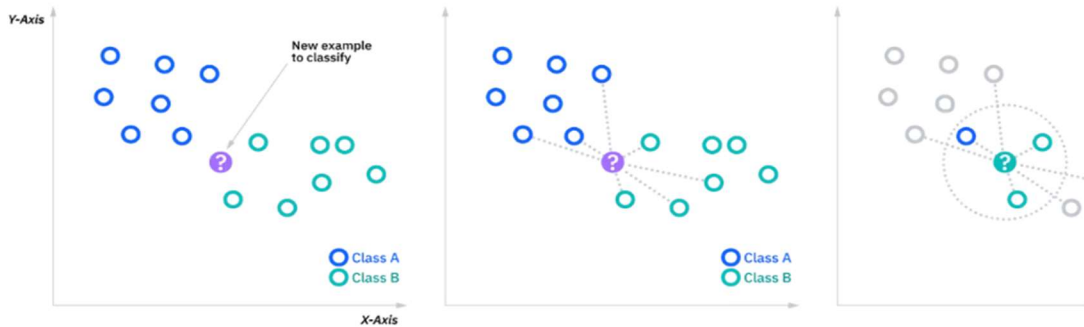


Figure 5 .KNN classification[6]

4. SVM

It is a supervised learning algorithm for tasks involving regression and classification. To find the best hyperplane that divides data points from various classes, with the greatest margin, is how it operates [3].

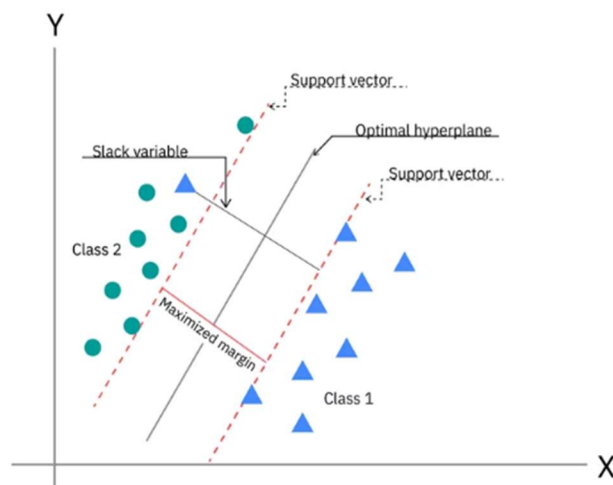


Figure 6.SVM [3].

5. NV

It is an algorithm of classification. It makes the assumption that features are conditionally independent; it makes computation easier and enables the model to work well even with massive datasets[4].

2.1.6 Ensemble Learning

Ensemble learning is a type of machine learning approach wherein a combination of several base learners is employed in order to enhance the accuracy of predictions as opposed to employing only one learner model. Instead of depending on a single classifier, an ensemble approach involves using the output of several learners in order to achieve better accuracy and efficiency of the process [7, 9]. The principle behind the approach is that a set of different learners can compensate for each other's weaknesses. The two main classes of ensemble learning approaches include Bagging and Boosting. Both techniques involve the use of several learner models but differ in their approach.

- Boosting

Another type of ensemble algorithm that increases predictive ability through the sequential training of many weak learners is called boosting. In contrast to bagging, the observations misclassified in the preceding rounds are given more weight, and thus further learners can concentrate on the hard cases [9]. The prediction result is calculated as the sum of all weak learners, weighted:

$$F(x) = \sum_{m=1}^M \alpha_m h_m(x)$$

in which: $h_m(x)$ refers to the prediction by the m^{th} weak classifier, α_m is its corresponding weight, M refers to the total number of boosting iterations.

As opposed to bagging, the boosting method tends to have better predictive accuracy due to the fact that each learner tries to fix the mistakes made by previous learners; however, these algorithms usually cost much more computing power and may be sensitive to noisy observations. The widely used boosting algorithms include **AdaBoost** and **XGBoost**, which were considered in this paper.

1- AdaBoost

One of the oldest and highly successful boosting algorithms is called Adaptive Boosting (AdaBoost). Introduced by Freund & Schapire (1997), this method builds a composite strong classifier by combining weak classifiers iteratively and by modifying sample weights at each iteration.

$$H(x) = \text{sign} \left(\sum_{m=1}^M \alpha_m h_m(x) \right)$$

Which $H(x)$ refers to the final prediction, $h_m(x)$ is the output of the m^{th} weak classifier, and α_m shows the importance applied to that classifier.

2- Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is one of the modern implementations of the Gradient Boosting algorithm, which was invented to enhance efficiency and increase accuracy in predictions. The algorithm builds an ensemble of decision trees sequentially in such a way that the next tree minimizes the residual errors made by the preceding ones. Different from other boosting algorithms, XGBoost uses the regularized approach to avoid overfitting and preserve predictive performance [9].

The prediction of XGBoost is computed as:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i)$$

where:

- $\hat{y}_i$ is the predicted output for the i -th sample.
- x_i is the input feature vector of the i -th sample.
- $f_k(x_i)$ is the prediction produced by the k -th decision tree.
- K is the total number of decision trees in the ensemble.

The objective function of XGBoost combines the training loss with a regularization term to improve generalization and reduce overfitting:

$$\text{Obj} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

where:

- **Obj** is the overall objective function.
- $l(y_i, \hat{y}_i)$ is the training loss that measures the difference between the true value y_i and the predicted value $\hat{y}_i$.
- $\Omega(f_k)$ is the regularization term that controls the complexity of the k -th decision tree.
- n is the total number of training samples.
- K is the total number of trees in the ensemble.

2.1.7 Model Evaluation Metrics

We used it to check how well classification models work and make sure they make accurate predictions, which is especially important in serious situations like finding students who are at risk[5].

1. Accuracy

It is the percentage of correctly predicted cases among all of the model's predictions.

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

Where: **TP** = True Positives, **TN** = True Negatives, **FP** = False Positives, **FN** = False Negatives. It is simple and intuitive; however, not deal with imbalanced datasets[6].

2. Precision

It is the percentage of correctly predicted positive cases among all predicted positive cases.

Precision = TP / (FP + TP), where it is important when the cost of false positives is high[6].

3. Recall

It is sometimes referred to as sensitivity, which gauges how well the model can identify real positive examples.

$$\text{Recall} = \frac{TP}{FN + TP}$$

When there is a positive case (like a student who is at risk), recall is very important[6].

4. F1-Score

When working with imbalanced datasets, the F1-score—which is the harmonic mean of precision and recall—offers a fair assessment of model performance.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This is especially helpful when you have datasets that aren't balanced, and you need both precision and recall [6].

2.1.8 Feature Selection

Feature selection is a type of preprocessing technique that allows identifying the most significant features for model training and discarding redundant or irrelevant features. This can reduce the number of features, simplify the training process, lower the cost, and enhance the quality of the trained model by excluding noise from the input data [9].

2.1.9 Synthetic Minority Oversampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) is the data balancing technique proposed as an improvement of the class imbalance problem. As opposed to replicating the samples that belong to the minority class, SMOTE creates synthetic samples by interpolating between the neighbors from the minority class[10].

A synthetic sample is generated according to the following formula:

$$x_{\text{new}} = x_i + \lambda(x_{\text{nn}} - x_i),$$

Where : x_i refers to its nearest neighbor, x_{nn} refers to its nearest neighbor, and λ is refers to a random number between 0 and 1.

2.1.10 Grid Search with Cross-Validation

Grid Search is a hyperparameter tuning technique that allows finding the optimal set of parameter values. Grid Search tests all possible combinations of parameter values and chooses the best one, which gives the best result. For more reliable performance estimation, Grid Search is often employed together with k-fold cross-validation. In this case, the training dataset is split into k equal parts. Each time, one part serves as a validation set, while the other parts form the training set. This procedure is performed until each part has served as the validation set[3]. The average cross-validation score is calculated as

$$CV_{score} = \frac{1}{k} \sum_{i=1}^k Score_i$$

where

- k is the fold number,
- $Score_i$ is the evaluation score obtained from the i^{th} fold.

2.1.11 Impact of Artificial Intelligence in Education

The utilization of AI in the education sector makes it possible for teachers to recognize potential issues that students may have in their studies in a timely manner and, therefore, make necessary adjustments. Academic risks can be predicted through analysis of the information collected from the performance of students using various Machine Learning techniques. By employing predictive models powered by artificial intelligence, educational institutions can keep students[8, 9].

2.2 Related Work:

Several studies have developed and applied ML models to improve students' academic performance through early prediction.

2.2.1 Prediction of Students' Academic Performance Using ML.

Macfadyen [10] described how LMS data might be used to forecast the academic performance of students and determine which ones need extra help. The data sample for this research was obtained from the Blackboard Vista LMS at the University of British Columbia in Canada. Data cleaning was conducted, which entailed the elimination of observations that had missing values and structuring for predictive modeling. The regression model technique was adopted for developing the predictive model for academic success using activity data from the LMS. According to the findings, activity data of students from the LMS can be used to make accurate predictions regarding their performance. The regression model was able to explain more than 30% of the variance in the final grade. Additionally, the logistic regression model accurately predicted 81% of the failures.

Akçapınar et al.[11] studied 76 students enrolled in the Computer Hardware course in the Department of Computer Education and Instructional Technology at a public university in Turkey. The participants were between 21 and 23 years old, including 41 women and 35 men.

The dataset included students' academic achievement records together with their online learning interaction data collected at the end of the semester, including 3,803 sessions, 119,921 page visits, 4,566 written texts, 62 discussion forum questions, 297 responses, 8,601 evaluations, and 3,937 tags used to label written messages. They used two different preprocessing methods. These included techniques for data transformation and feature selection. The CN2 (86%) and KNN (86%) algorithms achieved the highest classification accuracy, and the Gini index was used to identify the most informative features. According to the results of the study, at the end of the term, the KNN algorithm had an accuracy of 89% in predicting students at risk of academic failure.

With various ML algorithms, Khan, Ijaz, et al. [12] created a set of prediction models. The dataset was collected from a course delivered at Buraimi University College, Sultanate of Oman. It consisted of 151 student records described by ten predictive features and one target class used for model development. The noise is either eliminated or minimized using a variety of methods. The training dataset contains 151 instances after the noisy instances are eliminated. WEKA, an open-source program with a variety of algorithms for data pre-processing, classification, clustering, regression, and association rules, was created at the University of Waikato in New Zealand. Four popular ML algorithms were chosen. They selected KNN, which is implemented as IBk in WEKA, from lazy algorithms. In WEKA, RepTree is a DT implementation. In a similar vein, they selected NV as the fourth algorithm and Multilayer Perceptron, a class of Artificial Neural Networks. The models were evaluated using 10-fold cross-validation. The proposed model successfully identified academically at-risk students by the sixth week of the semester, providing timely intervention and supporting both instructors and students. Four out of five students were able to produce an end result that was acceptable.

The employment of AI techniques to predict student academic performance within the BL environment of SEU was first done by Hamadneh [13]. The study sought to identify the major determinants of student academic achievement and develop a PD for student academic performance. Data for this study were extracted from Blackboard LMS. Data preprocessing techniques were employed. These techniques included descriptive statistics, correlation, multicollinearity test, and regression analysis, which used ordinary least squares, fixed-effects, and random-effects models. To determine the quality of the data and the key factors that determine students' academic success, the following procedures were followed. The researchers subsequently built a PD using the artificial neural network technique. The ANN in this case is a multilayer perceptron, whose input nodes are four, hidden nodes are ten, and the output node is one, which represents the final grade. The results indicated that the proposed hybrid neural network model achieved strong predictive performance with low prediction error.

Queiroga et al.[14] presented a research paper that aimed to predict academically at-risk secondary school students using ML and learning analytics. The study also emphasized the importance of using previous academic performance as a predictor of future academic performance. The database for this study was sourced from the National Administration of Public Education of Uruguay that covered the educational life journey of students from primary school to secondary school spanning 2015-2020. This included information on 258,440 students. Data preprocessing was also undertaken prior to modeling. Data preprocessing entailed the cleansing of data, integration of databases for multiple education systems,

conversion of the database into a time series database, and derivation of additional attributes. Data mining methodology. ML algorithms such as RF, DT, AdaBoost, LR, NB, and MLP were investigated. However, the latter model was preferred since it had better performance and was more interpretable. From the analysis, it is clear that the ML models have done remarkably well by achieving a very high accuracy rate when predicting at-risk students. The AUROC and F1 scores produced results above 0.90 and 0.88, respectively.

Christou et al. [15] have proposed an ML approach for predicting students' future academic performance and study time based on historical educational records. Their study introduced a feature selection and feature construction technique, known as FSC4RBF, implemented using Radial Basis Function Neural Networks combined with grammatical evolution. The dataset was collected from multiple departments at the University of Ioannina. The preprocessing stage involved organizing, cleaning, and transforming the academic records into suitable input features for ML models. The proposed method was evaluated against several classification techniques, including DT, Bayesian Network, SVM, and different Neural Network models. Experimental results demonstrated that FSC4RBF achieved the highest classification accuracy while producing the lowest mean squared error across regression tasks, indicating its effectiveness in predicting students' academic outcomes.

Realinho et al. [16] introduced a study in which a public higher education dataset containing 4424 student records with demographic, socioeconomic, macroeconomic, and academic features was developed to predict at-risk students and academic success. The data were collected from multiple institutional sources, merged, cleaned, and preprocessed to support ML classification tasks. Several ML techniques were applied to analyze student performance and identify at-risk students. The study used a public dataset from the Polytechnic Institute of Port Alegre, Portugal, containing 4424 student records with 35 attributes. Data collected from multiple sources, such as academic systems and national databases. The data were then combined into one single dataset. Data cleaning was done by eliminating duplicates, managing missing data, managing outliers, renaming variable names, and selecting suitable features. The study applied RF, XGBoost, Light GBM, and CatBoost after performing data integration, preprocessing, and feature importance analysis using permutation feature importance. The study did not include results about specific accuracy since the main goal was presentation and analysis of the dataset, but not the development of a single ML model. Nonetheless, the presented dataset allows predicting at-risk students and student achievement and can be used for benchmarking ML models and building learning analytics systems.

2.2.3 Ensemble, Deep Learning, and Hybrid Models.

Johora [17] developed an XML approach to predict the academic performance of undergraduate students in Bangladesh. In this research, a dataset consisting of 872 samples was created. Preprocessing steps were followed, including handling of the missing data, elimination of unimportant columns, and conversion of the categorical variables to numeric values using one-hot encoding. To overcome the class imbalance problem, SMOTE was used, and feature scaling was done. After that, several ML approaches were implemented. They included RF, GB, XGBoost, SVM, KNN, DT, and LR. As can be seen, it was proved that a stacking ensemble model is better in terms of its performance. In order to interpret the model and see what features play a key role in achieving success among students, SHAP and LIME methods were

implemented in this research project. It should be noted that the highest accuracy reached 86.38% in terms of the stacking ensemble model.

CR-NODE was proposed by Bettahi et al[18], where continuous time was combined with completion-specific features to predict early dropout behavior. The dataset consists of longitudinal interactions recorded in the Canvas Learning Management System at a large public university spanning several terms. In the dataset, there are 100,878 students' enrollments that have final grades, out of which we filtered 89,734 temporal interactions meeting the minimum observations requirement. Data preprocessing was performed to ensure temporal consistency and data quality. They used six different baseline methods (Logistic Regression, RF, XGBoost, Gradient Boosting, LSTM, and Basic Neural ODE) to compare CR-NODE and set comprehensive performance benchmarks. The results showed that CR-NODE had the best balance between the rates of true positives and true negatives. XAI was used to understand model predictions.

Ujkani et al. [9] proposed a hybrid model based on CNN, LSTM, and XGBoost to predict students' academic success. The study used the Open University Learning Analytics Dataset (OULAD), which includes 32,593 students. Data preprocessing involved missing value handling, feature engineering, data integration, and data splitting. The proposed model was compared with several ML and DL models, and XAI was employed to interpret the predictions. The results demonstrated that the proposed model achieved the highest accuracy of 94%.

Bagunaaid [19] introduced E-FedCloud, an AI-based e-learning system. The method was intended to predict student performance and identify potential difficulties that students may face. To enhance performance, the students used Shannon entropy, which allowed secured data access. The application of deep reinforcement learning was achieved by using a method called ID2QN, which was used to predict the performance of students and categorize them according to their level of risk, such as students that might drop out of school, poor performers, and students that may fail the class were used to create personal recommendations for each student. The OULAD dataset comprising 32,593 students taking 22 courses was used as a training dataset. Preprocessing techniques were performed before using the modeling procedure. These include data cleansing, feature selection, multiple sources' data integration, and preparing behavioral, academic, and demographic variables. In addition, various ML algorithms and deep learning algorithms were used. As a result, ID2QN Deep Reinforcement Learning algorithm was chosen for predicting academic performance to classify students based on the level of their involvement and the customization recommendation system. In terms of securing the collected data, federated learning and Cycle-GAN were applied. It is worth mentioning that the E-FedCloud framework exhibited a high prediction capability of identifying students who might experience difficulties. The accuracy of the model equaled almost 97%, and the F1-score equaled 95%.

2.2.3 Early Warning Systems.

David Bañeres et al. [8] created a new system with multiple contributions. The Gradual At-Risk (GAR) model was the first contribution, a predictive model. Such a model predicted students' chances of failing a course based solely on their grades. Using the dataset in a data mart, a thorough analysis of the model was carried out on the entire collection of university courses.

2
151
KNN, DT, NB, SVM, LR, Hierarchical Mixed models, and Neural Network models. The second is a technique to establish a cutoff point for classifying a course's trained GAR model as high- or low-quality. In an actual classroom, the GAR model gave students feedback according to their risk level. As a result, a teacher cannot afford to give a student incorrect advice or a message. EWS uses GAR to modify the kind of message that is sent to the student. The third contribution involved using EWS in two courses in an actual classroom. In terms of dashboards and feedback, the GAR model was utilized to give teachers and students useful information. They examined how well at-risk students are identified in such a real-world learning environment and how this affects the courses' final performance. Two undergraduate courses were used to test the system, and the results demonstrated a high level of accuracy in identifying students who were at risk.

2
81
Albreiki, Balqis, et al. [20] stress the use of a customized rule-based system with the Risk Flag to identify and visualize at-risk students early in the course delivery. The dataset was gathered from an undergraduate programming course at the United Arab Emirates University's College of Information Technology. The information shows how well students performed in the introductory programming course during the academic years 2016–2017 (fall and spring), 2017–2018 (fall and spring), and 2018–2019 (fall and spring). The data was pre-processed to make it readable for the ML methods. A Python script was used to combine the files into a single CSV file, and min-max normalization was used to normalize the marks. The averaged value of the same coursework components was used to replace the missing values, which were handled using imputation techniques. Timewise, the results are still inefficient, and there are a lot of false predictions. They suggest a tailored rule-based model that provides at-risk students with prompt identification. It also helps the teacher identify the at-risk students and provide prompt intervention by visualizing them

120
29
14
30
The project of Skittou et al [21] developed an early warning system to identify students at risk of dropping out by considering a range of socio-cultural, structural, and educational factors that influence students' decisions to leave school. The study used an original dataset created by integrating data from multiple operational systems, comprising 125,354 students; it included 59,384 female students. The researchers identified a substantial number of significant features for distinguishing students who were likely to drop out from those who were not, reflecting a comprehensive approach to selecting dropout indicators. Several ML classification algorithms, including SVM, RF, Stochastic Gradient Descent, and KNN, were evaluated. Among these, the KNN model recorded the best performance, attaining an accuracy of over 99.5% on the training set and more than 99.3% on the testing set. To support educational planning and decision-making, the prediction results were presented through a Django-based web application developed specifically for the study.

145
The proposed RADAR system by Embarak et al[22] used statistical analysis and ML algorithms to identify patterns and trends in the data that indicate a student was at risk of falling behind or dropping out. A dataset of students from a particular school district was taken to build the system; the information was gathered and processed properly through feature integration, data cleansing, and feature selection. In this case, PAP can help in evaluating the ability of the student to continue studying in the future, as well as predicting the level of his/her success. With the help of such a comparison, Embarak et al. can emphasize the limitations of these

methods. The ML system incorporating XAI can process multiple kinds of data and consider other factors associated with the risks posed to students.

Table 1 Of The Reviewed Papers

Reference	Year	The best model	Dataset	Type	Result
[10]	2010	LR model	LMS dataset from Blackboard (student activity, messages, logins, assessments)	Private	Predicted at-risk students with 81% accuracy and explained 30% of the variation in final grades
[11]	2019	(KNN)	The study involved 76 undergraduate students from a public university in Turkey.	Private -Need request-	The kNN model achieved an accuracy of 89% in identifying students at risk of academic failure by the end of the semester.
[12]	2021	Developed a set of prediction models	The study utilized students' academic records from a course offered at Buraimi University College (BUC).	Private - Need request-.	The prediction model proved to be an effective and valuable tool for supporting both instructors and students.
[13]	2022	Neural Network trained by Firefly Algorithm	Dataset from the Blackboard Learning Management System at the Saudi Electronic University	Private	Reported strong predictive performance of 0.39 and highlighted midterm exams, assignments, attendance, and study mode as key factors
[14]	2022	RF	Nationwide K-12 dataset (Uruguay), 258,440 students; demographic data, grades, absences, social welfare, school data (2015–2020)	Private	Achieved high performance (AUROC > 0.90, F1 > 0.88) and successfully predicted at-risk and dropout students early
[15]	2023	Feature selection and construction method for RBF network	Student academic records from several departments, University of Ioannina	Private	Achieved the highest accuracy and the lowest error compared with other models in predicting grades and dropout
[16]	2022	Not specified	Student Dropout and Academic Success Dataset – Polytechnic Institute of port Alegre	Public No Access	Identified the most important factors for dropout prediction, such as grades, approved courses, tuition status, and course type

[17]	2025	Stacking Ensemble (RF + GB + SVC)	872 undergraduate students from multiple universities; demographic, academic, and behavioral features (31 features)	Private	Achieved 86.38% accuracy; SHAP and LIME identified key factors such as attendance affecting student performance
[18]	2025	CR-NODE	Large public university Dataset.	Private	CR-NODE achieves the best balance
[9]	2024	Custom Neural Network Architecture	Open University Learning Analytics Dataset	Public	Achieved up to 94% accuracy in student success prediction.
[19]	2024	ID2QN (Improved Dueling Deep Q Network)	OULAD dataset (32,593 students, 22 courses); demographic, academic, and behavioral LMS data	Public No Access	Achieved ~97% accuracy and ~95% F1-score; outperformed PCT, MST-FaDe, Fuzzy SVM, and AISAR; effectively predicted at-risk students early
[8]	2020	Gradual At-Risk (GAR) model	The model was tested using data available in an institutional data mart, where curated data from six semesters are available.	Private	The model showed a high accuracy in correctly detecting at-risk students.
[20]	2021	Customized rule-based system (RBS)	It was collected from a programming course taught for undergraduate students at (CIT), United Arab Emirates University (UAEU).	Private - Need request-.	Visualized at-risk students to assist educators in identifying them and providing timely interventions.
[21]	2023	(KNN) algorithm	They used a tool built specifically for this project by aggregating data from several operational systems	Privates need to request access from the author	The findings were presented using a custom-developed Django web application, illustrating its effectiveness in supporting educational planning and informed decision-making.
[22]	2024	RADAR system	Students from a school district Dataset.	Private	Improved early identification and support for at-risk students.

Chapter 3

Methodology and Proposed Approach

Introduction

Ethical Considerations: The private LMS dataset used in this study was anonymized before analysis to ensure that no personally identifiable information was included. Permission to use the dataset was obtained from the corresponding educational institution for research purposes. All data were handled confidentially and used solely for academic research, with appropriate measures taken to protect student privacy throughout the study.

This section presents the methodology adopted to design, implement, and evaluate the proposed framework, AI-Based Early Warning Framework for Academic Risk Detection in education. Figure 7 shows the main steps of the proposed framework to identify the risk of academic achievement using a private educational dataset collected from an LMS. The proposed framework includes many steps: data collection, data preprocessing, feature engineering, train-test splitting, class balancing using SMOTE, machine learning model training, hyperparameter tuning, model evaluation, and student risk classification:

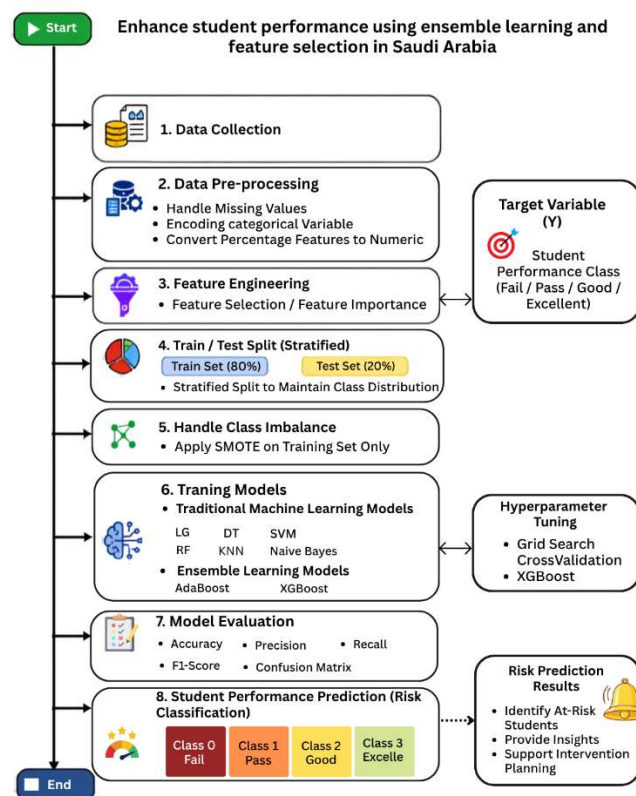


Figure 7 The proposed framework

3.1 Dataset Description

We used a private educational dataset collected from an LMS at a private educational institution in the Kingdom of Saudi Arabia. The dataset was collected over three months, from November 1, 2025, to January 1, 2026. It contains 1,395 anonymized student records from K–12 students across all educational levels and covers both scientific and literary subjects. The dataset is described by 17 academic and behavioral features related to participation, assessments, attendance, and student engagement. The dataset was used with permission for research purposes, and all personally identifiable information was removed prior to analysis to protect student privacy. Due to privacy and institutional restrictions, the dataset is not publicly available but may be made available upon reasonable request and institutional approval. Feature selection was performed using RF importance to rank the input features according to their predictive contribution. Based on the obtained rankings, the ten highest-ranked features (HomeWorks, Absence, Interactive Contents, AVG Participations %, Contents Participation %, HomeWorks Participation %, Virtual Classrooms, Discussion Rooms Participation %, Virtual Classrooms Participation %, and No. Active Courses) were selected to construct the selected feature set, while the original full feature set containing all 17 features was retained for comparison in the experimental evaluation. Consequently, the remaining seven lower-ranked features (Interactive Contents Phet, Discussion Rooms, Mailbox Participation %, Exams Participation %, Exams, Teachers' Messages, and Student Login %) were excluded only from the selected feature set. **Error! Not a valid bookmark self-reference.** presents the distribution of the target variable, which represents students' academic performance and classifies each student into four categories: Excellent, Good, Pass, and Fail. Table 2. Distribution of Student Performance Classes. Contains a description of the 17 Educational Features.

Table 2. Distribution of Student Performance Classes.

Class	Number of Students	Percentage
Good	725	51.97
Excellent	322	23.08
Fail	241	17.28
Pass	107	7.67

Table 3 Description of the 17 Educational Features

Feature	Description
HomeWorks	Student homework performance.
Absence	Number of student absences.
Interactive Contents	Student interaction with interactive learning materials.
AVG Participations %	Average student participation percentage.
Contents Participation %	Participation percentage in learning content.
HomeWorks Participation %	Participation percentage in homework activities.
Virtual Classrooms	Number of virtual classroom sessions attended.
Discussion Rooms Participation %	Participation percentage in discussion rooms.
Virtual Classrooms Participation %	Participation percentage in virtual classroom activities.

No. Active Courses	Number of active courses enrolled by the student.
Interactive Contents Phet	Participation in PhET interactive simulations.
Discussion Rooms	Number of discussion room activities.
Mailbox Participation %	Participation percentage in LMS mailbox communications.
Exams Participation %	Participation percentage in examinations.
Exams	Student examination performance.
Teachers' Messages	Number of teacher messages received or exchanged.
Student Login %	Student login frequency/percentage to the LMS.

3.2 Data Preprocessing

Data preprocessing is an important step in this process. It ensures the preparation of data for further analysis and improves its quality prior to modeling. The preprocessing includes the following steps:

- Clean data set: Remove percentage % symbols from some features.
- Encoding categorical variables into numerical representations to ensure compatibility with ML algorithms.

3.3 Data Splitting

The dataset was partitioned into training (80%) and testing (20%) sets using the Stratified Train-Test Split technique. This approach preserved the original distribution of the four academic performance categories across both subsets, ensuring a fair and reliable evaluation of the developed ML models.

3.4 Handling imbalanced dataset using SMOTE:

SMOTE is a data balancing technique designed to address the class imbalance problem by generating synthetic samples for minority classes rather than simply duplicating existing instances [9]. In this study, the dataset was first divided into training and testing sets using a stratified train-test split. SMOTE was then applied only to the training set to balance the class distribution, while the testing set remained unchanged to prevent data leakage and ensure a fair and reliable evaluation of the machine learning models.[9]

A synthetic sample is generated according to the following formula:

$$x_{\text{new}} = x_i + \lambda(x_{nn} - x_i)$$

where x_i represents a sample from a minority class, x_{nn} represents one of its nearest-neighbor samples from.

3.5 Feature Selection

To avoid feature-selection leakage, we first divided the dataset into training and testing sets. After that, we calculated the feature importance on the basis of the training dataset only. In this case, an RF classifier containing 100 decision trees was trained using the entire training features. On the other hand, the test set was left completely untouched during the feature selection process.

The selected features were subsequently applied to both the training and testing sets:

Feature Importance: RF feature importance was used to evaluate the importance of each input feature. The importance of each individual feature was calculated by averaging the

node impurity across 100 decision trees. The features were sorted in decreasing order of their importance.

A fixed random seed of 42 was used to ensure reproducibility.

3.6 Optimization Models

Model optimization is an essential step in developing robust and high-performing machine learning models. Appropriate hyperparameter selection can significantly improve prediction accuracy, enhance model generalization, and reduce the risk of overfitting. Various optimization techniques have been proposed to identify the optimal hyperparameter configuration for machine learning models. In this study, Grid Search combined with Stratified Cross-Validation was employed to systematically optimize the selected machine learning models prior to the final evaluation.

3.6.1 Grid Search

An exhaustive hyperparameter optimization technique that systematically evaluates all possible combinations of predefined hyperparameter values within a specified search space. The objective is to identify the hyperparameter configuration that achieves the best predictive performance based on a predefined evaluation metric[3].

3.6.2 Stratified Cross-Validation

It is used to estimate the generalization capability of ML models by partitioning the training dataset into multiple folds. During each iteration, one fold is used for validation, while the remaining folds are used for training. This procedure is repeated until every fold has been used once for validation. The overall model performance is calculated as the average evaluation score across all folds[3].

The average cross-validation score is computed as follows:

$$CV = \frac{1}{k} \sum_{i=1}^k M_i$$

where:

- k is the number of folds.
- M_i is the evaluation metric obtained from i the fold.
- CV represents the average cross-validation score.

3.6.3 Hyperparameter Optimization Procedure

In this study, Grid Search was applied to optimize the hyperparameters of the evaluated ML models. Stratified cross-validation preserves the original class distribution within each fold, making it suitable for the imbalanced multi-class educational dataset used in this research. Hyperparameter optimization was performed after data preprocessing, feature selection (when applicable), and class balancing using SMOTE. The weighted F1-score was adopted as the optimization criterion because it provides a balanced evaluation for imbalanced multi-class classification problems. After identifying the optimal hyperparameter configuration, each optimized model was retrained using the entire training dataset before being evaluated on the independent testing dataset. The hyperparameter search spaces explored during the optimization process are summarized in **Error! Reference source not found.**

Table 4. Hyperparameter Search Space Used for Grid Search Optimization

Model	Hyperparameter Search Space
LR	$C = \{0.01, 0.1, 1, 10\}$; solver = {lbfgs, liblinear}; penalty = {l2}; class_weight = {None, balanced}
SVM	$C = \{0.1, 1, 10\}$; kernel = {linear, rbf}; gamma = {scale, auto}; class_weight = {None, balanced}
KNN	n_neighbors = {3, 5, 7, 9, 11}; weights = {uniform, distance}; metric = {euclidean, manhattan}
DT	criterion = {gini, entropy}; max_depth = {5, 10, 15, 20, None}; min_samples_split = {2, 5, 10}; min_samples_leaf = {1, 2, 4}; class_weight = {None, balanced}
RF	n_estimators = {100, 200}; criterion = {gini, entropy}; max_depth = {10, 20, None}; min_samples_split = {2, 5}; min_samples_leaf = {1, 2}; max_features = {sqrt, log2}; class_weight = {None, balanced}
NB	var_smoothing = $\{10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}\}$

Table 4 Error! Reference source not found. summarizes the hyperparameter search spaces evaluated during the Grid Search optimization process for all eight ML models. The optimal hyperparameter configuration for each model was automatically selected using Stratified 3-Fold Cross-Validation based on the weighted F1-score. The optimized models were then retrained using the entire training dataset and subsequently evaluated on the independent testing dataset to obtain the final performance results.

3.7 Machine Learning Models

ML was employed to develop predictive models for students' academic performance using LMS data. In this study, five traditional ML models, namely Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), and Naïve Bayes (NB), together with three ensemble learning models, namely Random Forest (RF), AdaBoost (AB), and Extreme Gradient Boosting (XGB), were implemented and evaluated. Model performance was assessed under four experimental configurations using full and selected feature sets, with and without SMOTE. Hyperparameter optimization was performed using Grid Search with Stratified 3-Fold Cross-Validation, and the weighted F1-score was used as the optimization criterion. Pattern Recognition and ML)[3]

Six traditional classification algorithms were investigated in this study:

1. **DT:** It represents decision-making using a tree-like structure. In this structure, each internal node corresponds to a feature, each branch represents a decision rule, and each leaf node denotes a class label. Due to their simplicity and interpretability, DT are widely used for classification tasks and are easy to understand and implement[5]
2. **RF:** DT algorithms are used in the RF to increase the efficiency of predicting results and prevent overfitting. All DTs learn from a random sample of the training dataset, and the output result is produced based on the aggregation of all DT predictions.[5]

3. **KNN:** It uses the classes of a data point's closest neighbors in the feature space to predict the data point's class. It uses Euclidean distance, and other distance metrics are used by the algorithm to determine similarity [3]
4. **SVM:** It is a supervised ML algorithm employed in classification and regression. It operates by identifying the optimal hyperplane that separates data points from different classes while maximizing the margin between them, thereby improving the model's generalization performance.[3].
5. **NB:** It is an algorithm of classification. It makes the assumption that features are conditionally independent; it makes computation easier and enables the model to work well even with massive datasets[4]
6. **LR:** It is a supervised ML algorithm primarily used for classification tasks. It evaluates the probability that an instance belongs to a particular class by applying the logistic (sigmoid) function to a linear combination of the input features.

These models represent different statistical, probabilistic, distance-based, and tree-based learning approaches, providing a comprehensive comparison of classical machine learning techniques for predicting students' academic performance.

3.8 Ensemble Learning Models

Ensemble learning is a type of machine learning approach wherein a combination of several base learners is employed to enhance prediction accuracy, as opposed to employing only one learner model. Instead of depending on a single classifier, an ensemble approach involves using the output of several learners in order to reach better accuracy and efficiency of the process [6, 8].

The widely used boosting algorithms include AdaBoost and XGBoost, which were considered in this paper.

1- AdaBoost

Adaptive Boosting (AdaBoost) is a boosting algorithm that constructs a strong classifier by iteratively combining multiple weak classifiers and assigning greater weight to previously misclassified samples. The final prediction is computed as:

$$H(x) = \text{sign} \left(\sum_{m=1}^M \alpha_m h_m(x) \right)$$

where $H(x)$ denotes the final predicted class for input x , $h_m(x)$ represents the prediction produced by the m -th weak classifier, α_m is the weight assigned to that classifier according to its performance, and M is the total number of weak classifiers included in the ensemble.

2- XGBoost

It is an efficient implementation of the Gradient Boosting algorithm that improves prediction accuracy while reducing overfitting through regularization. The algorithm builds an ensemble of decision trees sequentially, where each new tree is trained to minimize the residual errors of the previous trees [8]

The prediction of XGBoost is computed as:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i)$$

where:

- $\hat{y}_i$ is the predicted output for the i -th sample.
- x_i is the input feature vector of the i -th sample.
- $f_k(x_i)$ is the prediction produced by the k -th decision tree.
- K is the total number of decision trees in the ensemble.

The objective function of XGBoost combines the training loss with a regularization term to improve generalization and reduce overfitting:

$$\text{Obj} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

where:

- **Obj** is the overall objective function.
- $l(y_i, \hat{y}_i)$ is the training loss that measures the difference between the true value y_i and the predicted value $\hat{y}_i$.
- $\Omega(f_k)$ is the regularization term that controls the complexity of the k -th decision tree.
- n is the total number of training samples.
- K is the total number of trees in the ensemble.

3.9 Model Evaluation Metrics

The optimized models were retrained using the optimal hyperparameter values obtained through Grid Search and then evaluated on the testing dataset. The evaluation was performed using the same performance metrics adopted throughout this study, including:

1. **Accuracy:** it is the proportion of correctly classified instances among all predictions made by the model. It is calculated as:

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

where TP denotes True Positives, TN denotes True Negatives, FP denotes False Positives, and FN denotes False Negatives. Accuracy is simple and intuitive; however, it may not provide a reliable evaluation for imbalanced datasets (Hastie et al., 2009). Since this study addresses a multi-class classification problem with imbalanced classes, Accuracy was reported together with weighted Precision, weighted Recall, and weighted F1-score. Weighted averaging was used for Precision, Recall, and F1-score to account for the class imbalance.[6].

2. **Precision:** It is the percentage of correctly predicted positive cases among all predicted positive cases.

$$\text{Precision} = \frac{TP}{TP + FP}$$

In our study, which addresses a four-class classification problem (Excellent, Good, Pass, and Fail), the overall precision was computed using the **weighted average**, where the precision of each class is weighted according to its number of instances. This provides an overall measure

that accounts for class imbalance. Precision is particularly important when the cost of false positives is high [6].

3. **Recall:** also known as Sensitivity, measures the model's ability to correctly identify actual positive instances. It is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

where TP denotes True Positives and FN denotes False Negatives. In this study, weighted Recall was used to evaluate the multi-class classification problem. Since the primary objective is the early identification of academically at-risk students, particular attention was given to improving the Recall of the minority classes (Fail and Pass), ensuring that students requiring intervention were correctly identified. [6].

4. **F1-Score:** The F1-score is the harmonic mean of Precision and Recall and provides a balanced evaluation of model performance, particularly for imbalanced datasets. It is calculated as:

$$\text{F1-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

In this study, the weighted F1-score was used as the primary evaluation metric because it accounts for class imbalance by considering the contribution of each class according to its support. [6].

5. **Confusion Matrix:** We used it to evaluate the performance of a classification model by comparing the predicted labels with the actual labels. It provides a comprehensive summary of the classification outcomes; it includes True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). In addition, it serves as the foundation for calculating key evaluation metrics, such as accuracy, precision, recall, and the F1-score.

Chapter 4 Results.

4.1 Experimental Setup

Experiments were run using Google Colab. Python 3.12.13 was used as the implementation environment. The proposed framework was implemented using Pandas 2.2.2 and NumPy 2.0.2 for data preprocessing, Scikit-learn 1.6.1 for ML model development, and XGBoost 3.3.0 for implementing the XGBoost classifier. The experiments were executed on an x86_64 processor architecture. The performance was evaluated on a private dataset that consists of 1,395 student instances and includes 17 academic and behavioral features. For data preprocessing, Pandas and NumPy libraries were used, while the ML models were built using the Scikit-learn framework and the XGB library. To conduct a fair experiment, we split the dataset into 80% train and 20% test using stratified sampling. Feature selection was applied in two experimental settings to investigate the impact of reducing the input feature set on model performance. To address the class imbalance problem, SMOTE was applied to the training set in the two experiments. To achieve a fair comparison, all ML models were optimized using Grid Search with cross-validation. To ensure reproducibility, a fixed random seed (random_state = 42) was used throughout data splitting, SMOTE, cross-validation, and model training whenever applicable.

4.2 Class Distribution

From the information represented in *Figure 8*, it is clear that the data distribution in the four academic performance classes was imbalanced. The percentage of the Good class constituted 51.97%, while the Pass class constituted only 7.67%. The rest of the data was related to the Excellent and Fail classes (23.08% and 17.28%). The problem of imbalance was solved using SMOTE. In the SMOTE-oversampled training dataset, there were equal numbers of observations in each class, meaning that the dataset was balanced.

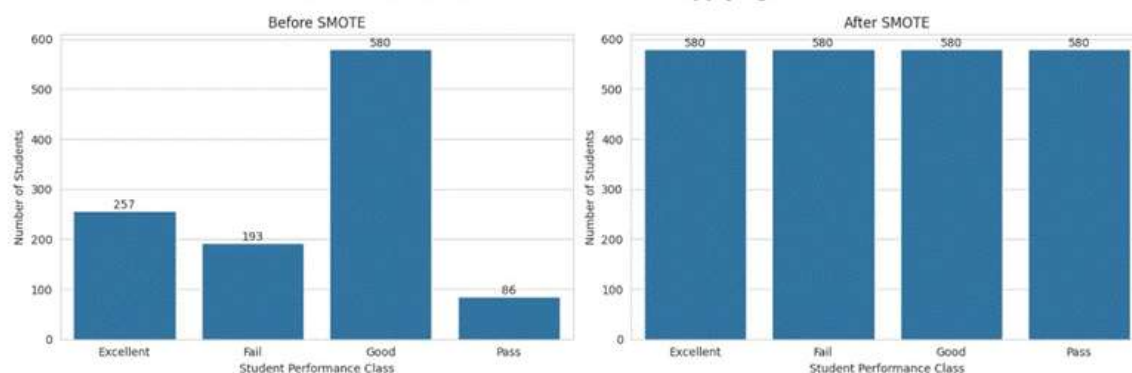


Figure 8 Distribution of Student Performance Classes with SMOTE.

4.3 Feature Importance Analysis

Feature importance analysis was conducted to identify the variables that contribute most to predicting students' academic performance. An RF classifier was used for the evaluation of the relative importance of each input feature. From the feature importance results shown in Figure 9, the high predictive importance of student academic performance includes Homework, Absence, and Interactive Contents. Other participation variables, such as Average Participation Percentage,

Contents Participation Percentage, and Homework Participation Percentage, are also demonstrated to be highly important. In contrast, Teacher Messages and Student Login Percentage had limited importance; these features contributed minimally to the model's predictive capability. From an educational view, the described characteristics are considered essential predictors of engagement and learning behaviors. Homework is a sign of constant practice and on-time completion of learning tasks, while absence denotes lower involvement in class activities and less exposure to the course material. In addition, interaction with the course materials available in the LMS shows student engagement in the learning process. Such behaviors correlate directly with academic performance and thus offer important insights into at-risk students.

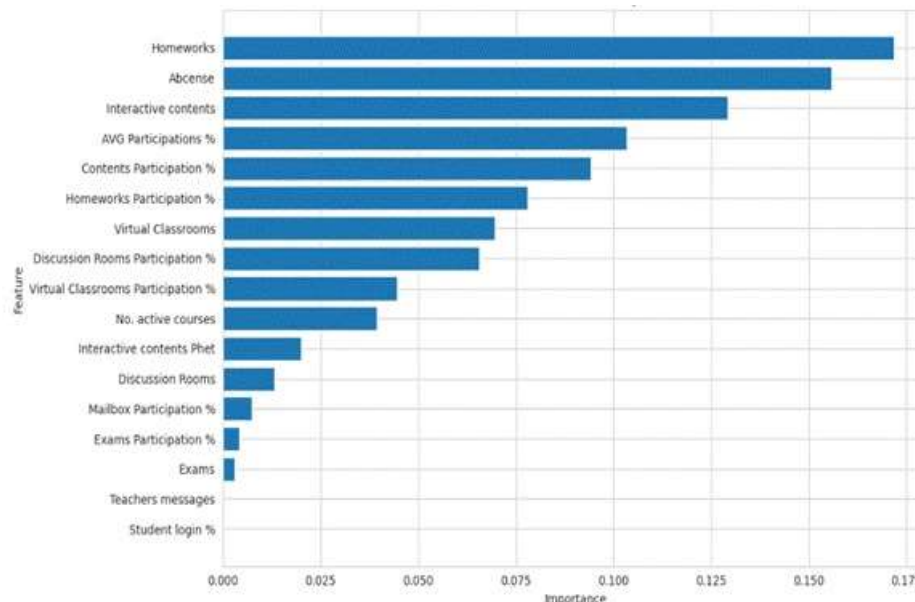


Figure 9. Random Forest Feature Importance Ranking.

4.4 Experimental Configurations

To evaluate the influence of feature selection and class balancing on prediction performance, the proposed models were assessed under experimental configurations: 1) **Results without applying SMOTE**: Full Feature Set+ Selected Features. 2) **Results with applying SMOTE**: Full Feature Set+ Selected Features. The two experimental setups were created and tested separately. We implemented them for comparison of the impact of both preprocessing steps individually and together. The following table **Error! Reference source not found.** and **Error! Reference source not found.** describe the two experimental setups. Each experiment was conducted under similar conditions using the same ML models and the same evaluation criteria for comparison.

4.5 Results without Applying SMOTE

Table 5 shows the performance of Applying Traditional and Ensemble Learning Models Using Full and Selected Features without SMOTE.

First, we can see that with full features. DT recorded the highest performance percentage, with 80.65 accuracy; RF achieved 79.21 accuracy. In contrast, the lowest performance across all evaluation metrics was NB. For the ensemble learning models, XGB recorded the highest accuracy of 82.44, whereas AB achieved an accuracy of 74.91%.

Second, the same models were evaluated with the selected feature set. DT and RF both achieved an accuracy of 81.00 accuracy. The lowest performance across all evaluation metrics was NV, while SVM and NB showed marginal performance gains. In contrast, LG, KNN, and AB exhibited unchanged performance across the full and selected feature configurations.

Overall, XGB ranked as the top-performing model in both cases, with an accuracy of 82.44% using the full feature set and 82.08% using the selected feature set.

Table 5 Results of applying models using full and selected features without SMOTE.

Configuration	Approach	Model	Accuracy	Precision	Recall	F1-Score
Full Features	Traditional ML	LR	70.97	72.36	70.97	69.85
		SVM	75.27	75.23	75.27	73.88
		KNN	76.70	76.64	76.70	75.55
		DT	80.65	80.20	80.65	80.21
		RF	79.21	79.64	79.21	77.86
		NB	14.34	35.52	14.34	10.89
	Ensemble Learning	AB	74.91	74.85	74.91	73.12
		XGB	82.44	82.93	82.44	81.81
Selected Features	Traditional ML	LR	70.97	72.36	70.97	69.85
		SVM	75.63	75.46	75.63	74.47
		KNN	76.70	76.64	76.70	75.55
		DT	81.00	80.63	81.00	80.55
		RF	81.00	82.12	81.00	79.85
		NB	15.77	43.32	15.77	14.98
	Ensemble Learning	AB	74.91	74.85	74.91	73.12
		XGB	82.08	81.99	82.08	81.46

The confusion matrix of

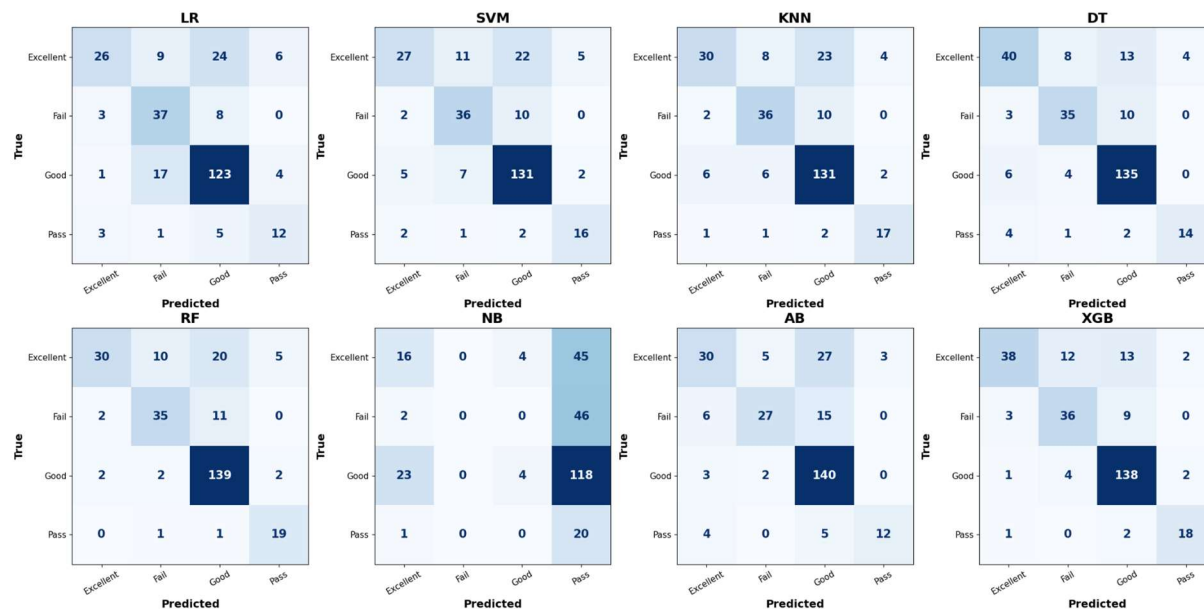


Figure 10 and Figure 11 shows additional understanding of the classification ability of the model. Based on each class (Excellent, Good, Pass, and Fail). We can see that XGB and DT get the best accuracies. The color intensity in each confusion matrix represents the number of observations assigned to each class, with darker colors indicating larger counts. XGB achieved the highest number of correctly classified observations, followed by DT and RF, as evidenced by the strong concentration of correctly classified instances along the main diagonal. However, NB recorded the largest number of misclassified observations, while LR, SVM, and KNN showed moderate classification performance with greater confusion between neighboring classes.

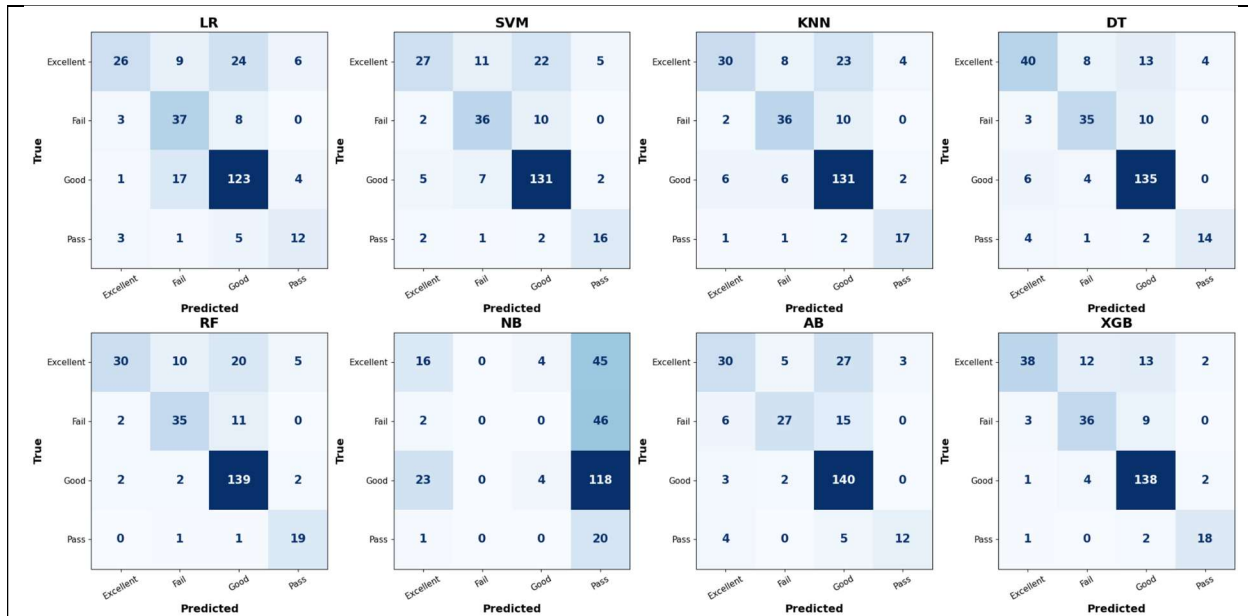


Figure 10 Confusion Matrices of the Models without SMOTE-full features.

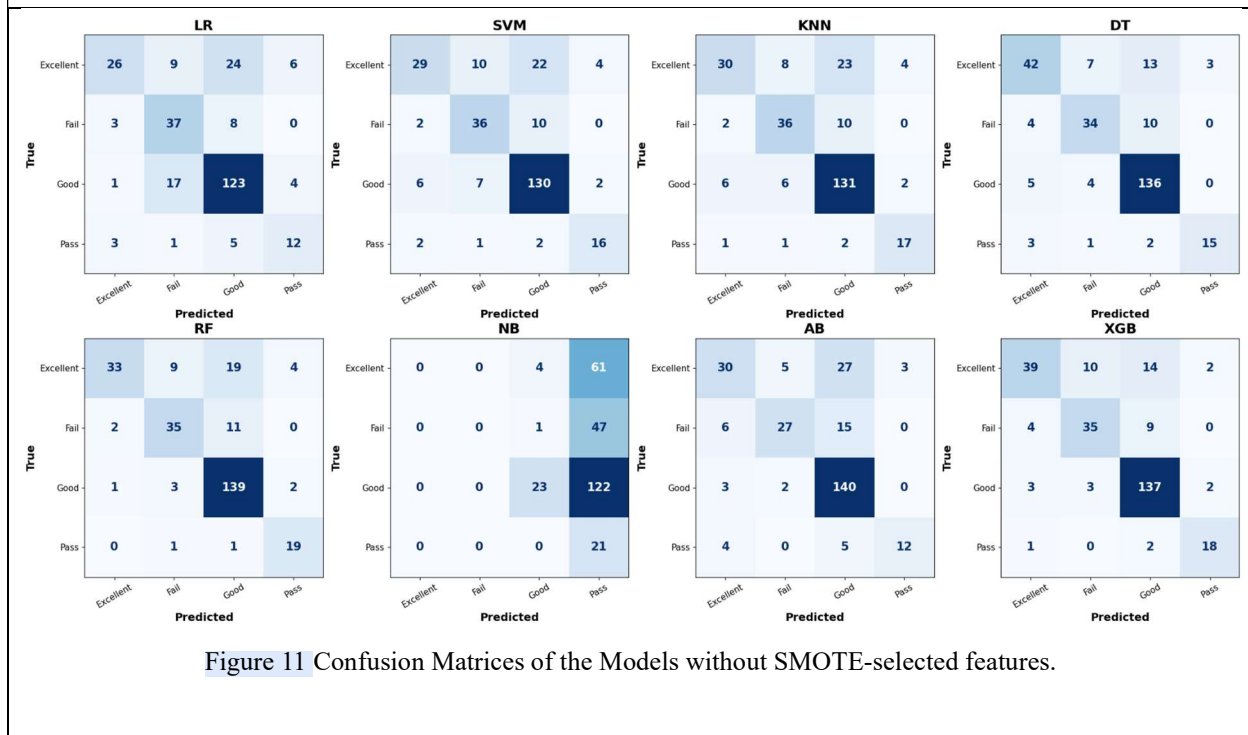


Figure 11 Confusion Matrices of the Models without SMOTE-selected features.

4.6 Results with Applying SMOTE.

Table 6 presents the performance results of Traditional and Ensemble Learning Models Using Full and Selected Features with SMOTE.

First, after applying SMOTE, we evaluated traditional and ensemble ML models using the full feature set. Among the traditional models, RF achieved the highest performance percentage (accuracy 81.00%, precision 81.89%, recall 81.00%, and F1-score 80.66%). DT also achieved the second- highest performance percentage (79.21% accuracy, 79.92% precision, 79.21% recall, and 79.36% F1-score). In ensemble learning, XGB achieved the best overall performance percentage (82.80% accuracy, 83.00% precision, 82.80% recall, and 82.46% F1-score). In contrast, NV recorded the lowest results across the evaluation metrics.

Second, using the selected feature set after applying SMOTE, the same models were evaluated. RF achieved (80.29% accuracy, 80.38% precision, recall of 80.29%, and F1-score of 79.87%), while DT achieved (79.57% accuracy, 80.39% precision, 79.57% recall, and 79.78% F1-score). For the ensemble learning models, XGB achieved the highest performance (82.80% accuracy, 83.00% precision, 82.80% recall, and 82.46% F1-score). AB achieved (73.84% accuracy, 74.11% precision, 73.84% recall, and 73.03% F1-score), whereas NV produced the lowest performance in this configuration. Overall, XGB achieved the highest performance across both experimental settings, achieving an accuracy of 82.80% with both the full feature set and the selected feature set.

Table 6 Results of applying models using full and selected features with SMOTE.

Configuration	Approach	Model	Accuracy	Precision	Recall	F1-Score
Full Features	Traditional ML	LR	64.16	69.51	64.16	64.19
		SVM	74.19	74.62	74.19	73.61
		KNN	77.42	77.47	77.42	77.21
		DT	79.21	79.92	79.21	79.36
		RF	81.00	81.89	81.00	80.66
		NB	17.56	49.84	17.56	17.18
	Ensemble Learning	AB	72.40	72.12	72.40	71.27
		XGB	82.80	83.00	82.80	82.46
Selected Features	Traditional ML	LR	64.16	69.51	64.16	64.19
		SVM	72.76	72.92	72.76	72.25
		KNN	77.42	77.47	77.42	77.21
		DT	79.57	80.39	79.57	79.78
		RF	80.29	80.38	80.29	79.87
		NB	16.49	43.94	16.49	16.02

	Ensemble Learning	AB	73.84	74.11	73.84	73.03
		XGB	82.80	83.00	82.80	82.46

Figure 12 and Figure 13 The confusion matrices demonstrate that applying SMOTE improved the classification balance across the four classes. The classification performance of the traditional ML and ensemble learning models. Based on (Excellent, Good, Pass, and Fail), XGB achieved the highest performance, the second highest performance RF and DT. The darker diagonal values show a higher number of correctly classified. In contrast, NV recorded the highest number of misclassified observations, whereas LR, SVM, KNN, and AB showed moderate classification performance.

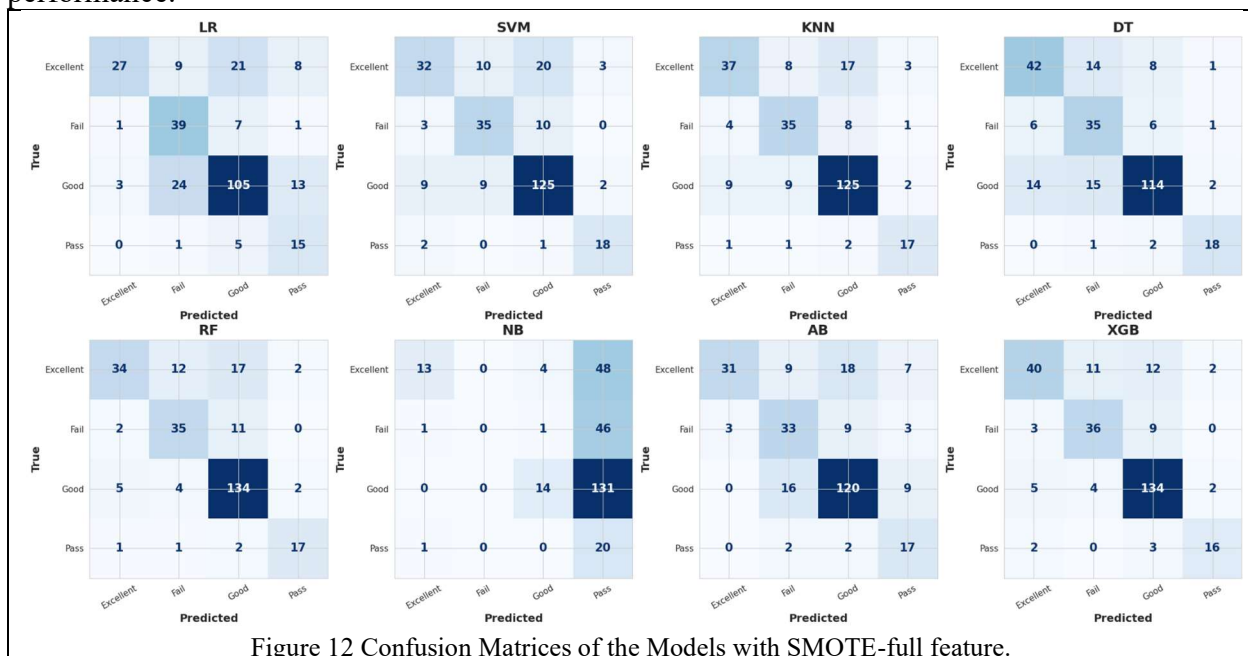
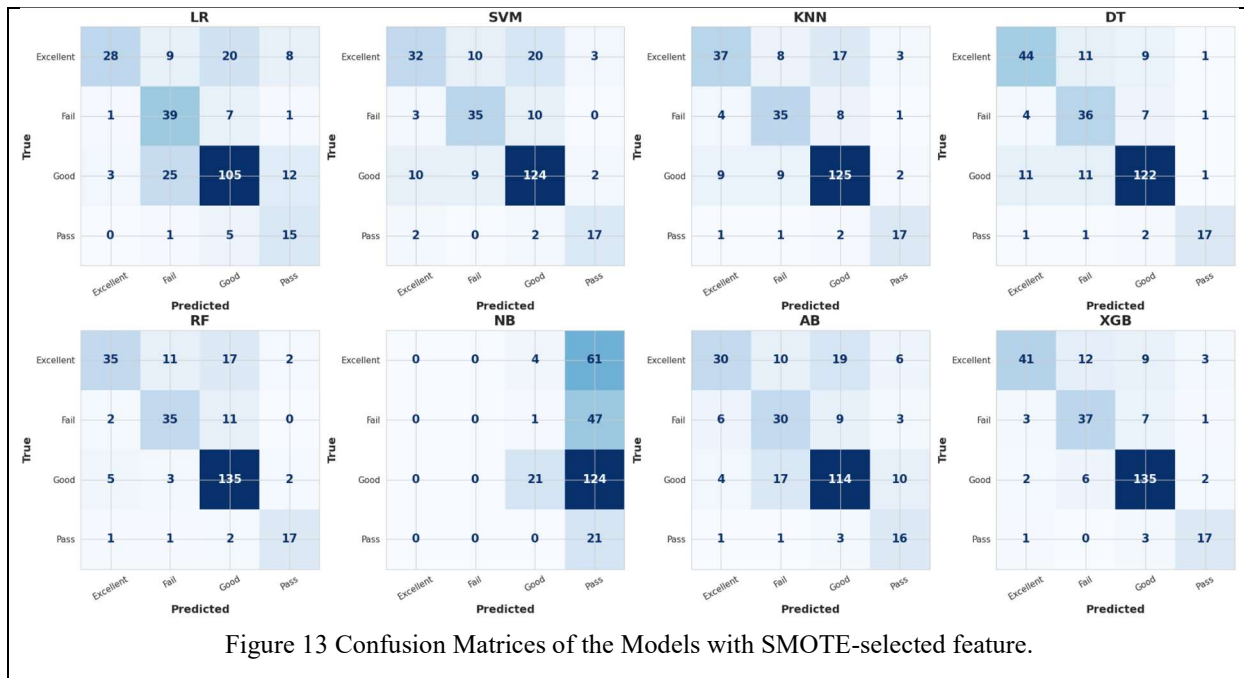


Figure 12 Confusion Matrices of the Models with SMOTE-full feature.



Discussion

Overall, the results show that the effects of SMOTE and feature selection varied depending on the ML algorithm. While Figure 14 Compares the accuracy of eight ML models across all experimental configurations using the optimized hyperparameters identified through Grid Search: full features + selected features without SMOTE, full features + selected features with SMOTE. The final evaluation of the proposed framework also considered precision, recall, F1-score, and confusion matrices. Emphasis was placed on recall because correctly identifying students in the Fail and Pass classes is more important than achieving high overall accuracy.

XGB achieved the highest accuracy, precision, recall, F1-score, and confusion matrices. In all configurations, performance ranged from 82.08% to 82.80%, showing stable performance regardless of the feature set or the use of SMOTE. Figure 15 presents the comparison before and after applying SMOTE to XGB. The findings show that the effect of SMOTE on the classes was not consistent, as the recall of the Fail class increased from 72.92% to 77.08%, demonstrating more effective identification of academically at-risk students after class balancing. Moreover, the Excellent class had a small improvement in terms of both the recall rate and F1 score. In contrast, the Recall rate of the Pass class decreased from 85.71% to 80.95%, while its F1-score dropped from 83.72% to 77.27%, while changes for the Good class were rather small. Such results demonstrate that the effect of SMOTE on the classes was not uniform, but that SMOTE altered the learning process and shifted the balance of classes, possibly identifying the minority classes better while having some negative effect on others. DT and RF also demonstrated competitive results, with accuracy close to 80%. RF benefited from SMOTE when the full feature set was used, whereas DT showed better performance without SMOTE.

KNN produced consistent results across all settings, with only a small improvement after applying SMOTE. In contrast, LR, SVM, and AB showed the lowest rate of accuracy after SMOTE was applied. NB remained the lowest-performing model in all configurations, although its accuracy improved slightly when SMOTE was used.

Feature selection caused marginal changes in most models. Reducing the number of features may help simplify the models without affecting their performance. However, the findings also show that SMOTE did not consistently improve accuracy, recall, and F-score across all models. Its main contribution is more likely related to improving the recognition of minority classes and producing more balanced predictions, which should also be examined through recall, F1-score, and the confusion matrices.

In all experimental setups, XGB demonstrated better results compared to other ML algorithms and achieved the best overall predictive efficiency of models.

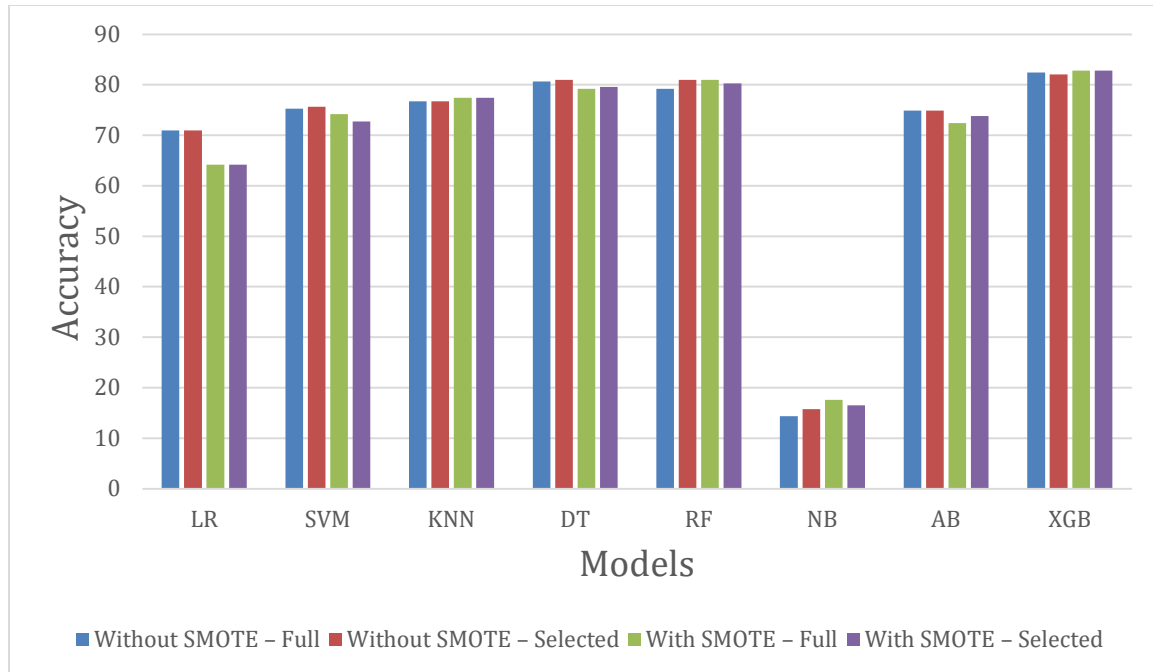


Figure 14 Accuracy comparison of Optimized ML models using full and selected feature sets with and without SMOTE.

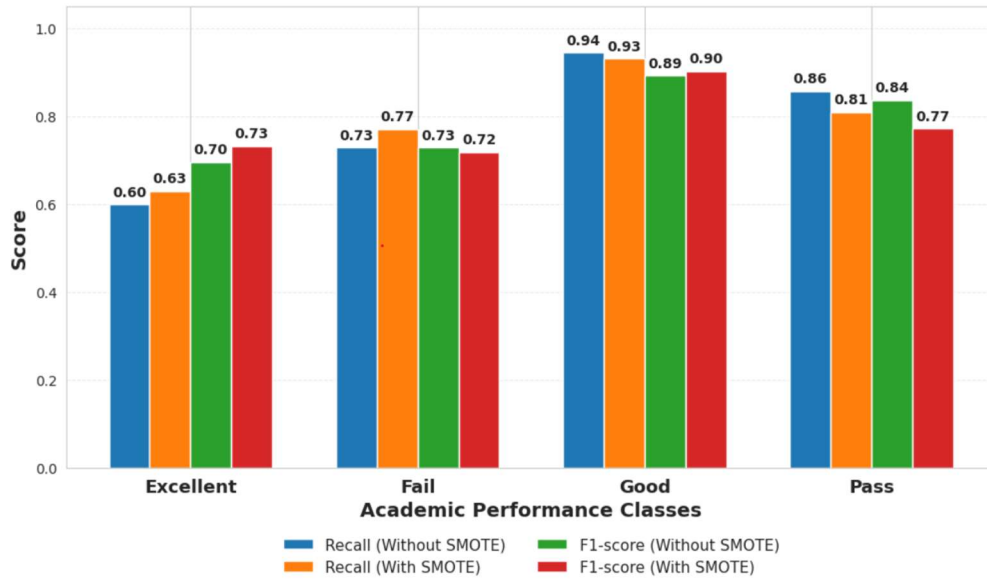


Figure 15. Class-wise Performance of the XGBoost Model with and without SMOTE

Chapter 5 Conclusion and Future Work

5.1 Conclusion

This study presented an AI-based Early Warning framework for predicting students' academic performance using a private LMS dataset collected in the Kingdom of Saudi Arabia. The proposed approach combined data preprocessing, Random Forest-based feature selection, SMOTE for class balancing, Grid Search with Stratified Cross-Validation for hyperparameter optimization, and the evaluation of multiple machine learning models. Eight supervised machine learning models were assessed under four experimental configurations. Among the evaluated models, XGBoost achieved the highest predictive performance, reaching an accuracy of 82.80%. In addition, feature selection reduced model complexity by identifying the most influential educational attributes, while SMOTE improved class balance and enhanced the recognition of minority classes. The proposed framework can be integrated into an LMS to automatically notify instructors when students are identified as academically at risk, allowing timely interventions through personalized guidance, targeted learning resources, and continuous monitoring. Although the proposed framework produced promising results, its applicability is limited by the use of a private dataset collected from a single educational institution. Future research should validate the proposed framework using larger and more diverse educational datasets, investigate advanced deep learning approaches, and integrate the framework into LMS to support real-time academic risk prediction and timely educational intervention. Although the proposed framework demonstrated promising results, further validation using larger multi-institutional datasets and integration with existing LMS platforms are required before real-world deployment.

5.2 Future Work

Although the proposed framework achieved the desired predictive performance, several directions can be explored in the future to enhance its effectiveness:

- Using larger and more diverse educational datasets collected from multiple educational sectors to improve its generalizability and robustness.
- Merge additional educational data sources, such as attendance records, behavioral indicators, discussion forum activities, assignment submission times, and temporal learning analytics, to provide a more comprehensive representation of students' learning behaviors.
- Investigate advanced deep learning approaches, including Long Short-Term Memory (LSTM), Transformer-based models, and hybrid machine learning-deep learning architectures for modeling sequential learning behaviors.
- Develop a real-time AI-based Early Warning System that continuously monitors student engagement throughout the semester and automatically alerts instructors, academic advisors, and students when academic risks are detected.
- Integrate Explainable Artificial Intelligence (XAI) techniques, such as SHAP and LIME, to improve the transparency and interpretability of prediction results and support educators' decision-making.

- Design a personalized intervention recommendation system that recommends appropriate academic support strategies according to each student's predicted risk level.
- Investigate online learning and automatic model updating techniques to continuously adapt the predictive model as new educational data become available.

References:

- [1] Y. Li, "Artificial Intelligence Technologies for Education," ed: MDPI-Multidisciplinary Digital Publishing Institute, 2025.
- [2] A. Taghvaie, T. Warnakulasuriya, D. Kumar, F. Zare, R. Sharma, and D. M. Vilathgamuwa, "A comprehensive review of harmonic issues and estimation techniques in power system networks based on traditional and artificial intelligence/machine learning," *Ieee Access*, vol. 11, pp. 31417–31442, 2023.
- [3] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning* (no. 4). Springer, 2006.
- [4] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [5] T. Hastie, "The elements of statistical learning: data mining, inference, and prediction," ed: springer, 2009.
- [6] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning," ed: Springer series in statistics New-York, 2009.
- [7] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.
- [8] D. Bañeres, M. E. Rodríguez, A. E. Guerrero-Roldán, and A. Karadeniz, "An early warning system to detect at-risk students in online higher education," *Applied Sciences*, vol. 10, no. 13, p. 4427, 2020.
- [9] B. Ujkani, D. Minkovska, and N. Hinov, "Course success prediction and early identification of at-risk students using explainable artificial intelligence," *Electronics*, vol. 13, no. 21, p. 4157, 2024.
- [10] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an “early warning system” for educators: A proof of concept," *Computers & education*, vol. 54, no. 2, pp. 588–599, 2010.
- [11] G. Akçapınar, A. Altun, and P. Aşkar, "Using learning analytics to develop early-warning system for at-risk students," *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, p. 40, 2019.
- [12] I. Khan, A. R. Ahmad, N. Jabeur, and M. N. Mahdi, "An artificial intelligence approach to monitor student performance and devise preventive measures," *Smart Learning Environments*, vol. 8, no. 1, p. 17, 2021.
- [13] N. N. Hamadneh, S. Atawneh, W. A. Khan, K. A. Almejalli, and A. Alhomoud, "Using artificial intelligence to predict students’ academic performance in blended learning," *Sustainability*, vol. 14, no. 18, p. 11642, 2022.
- [14] E. M. Queiroga, M. F. Batista Machado, V. R. Paragarino, T. T. Primo, and C. Cechinel, "Early prediction of at-risk students in secondary education: A countrywide k-12 learning analytics initiative in uruguay," *Information*, vol. 13, no. 9, p. 401, 2022.
- [15] V. Christou *et al.*, "Performance and early drop prediction for higher education students using machine learning," *Expert Systems with Applications*, vol. 225, p. 120079, 2023.
- [16] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," *Data*, vol. 7, no. 11, p. 146, 2022.
- [17] F. T. Johora, M. N. Hasan, A. Rajbongshi, M. Ashrafuzzaman, and F. Akter, "An explainable AI-based approach for predicting undergraduate students academic performance," *Array*, vol. 26, p. 100384, 2025.
- [18] A. Bettahi, H. Harroud, and F.-Z. Belouadha, "Early Student Risk Detection Using CR-NODE: A Completion-Focused Temporal Approach with Explainable AI," *Algorithms*, vol. 18, no. 12, p. 781, 2025.

- [19] W. Bagunaid, N. Chilamkurti, A. S. Shahraki, and S. Bamashmos, "Visual data and pattern analysis for smart education: a robust DRL-based early warning system for student performance prediction," *Future Internet*, vol. 16, no. 6, p. 206, 2024.
- [20] 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.
- [21] M. Skittou, M. Merrouchi, and T. Gadi, "Development of an early warning system to support educational planning process by identifying at-risk students," *IEEE access*, vol. 12, pp. 2260–2271, 2023.
- [22] O. H. Embarak and S. Hawarna, "Automated AI-driven system for early detection of at-risk students," *Procedia Computer Science*, vol. 231, pp. 151–160, 2024.