

ماریة باوزیر

MARYA HALA 22 JULY

%D8%A7%D9%84%D9%85%D8%AC%D9%84%D8%A9.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:400586817

Submission Date

Jul 22, 2026, 1:30 AM GMT+3

Download Date

Jul 22, 2026, 7:49 PM GMT+3

File Name

MARYA HALA 22 JULY ملجمل.docx

File Size

1.2 MB

24 Pages

8,808 Words

52,416 Characters

26% Overall Similarity

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

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

-  **164 Not Cited or Quoted 22%**
Matches with neither in-text citation nor quotation marks
-  **37 Missing Quotations 4%**
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

- 15%  Internet sources
- 18%  Publications
- 18%  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

- 164 Not Cited or Quoted 22%**
Matches with neither in-text citation nor quotation marks
- 37 Missing Quotations 4%**
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

- 15% Internet sources
- 18% Publications
- 18% Submitted works (Student Papers)

Top Sources

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

1	Submitted works	Midocean University on 2026-05-02	4%
2	Internet	www.mdpi.com	2%
3	Internet	dev-journal-bucket.s3.amazonaws.com	<1%
4	Internet	link.springer.com	<1%
5	Publication	Ossama H. Embarak, Shatha Hawarna. "Automated AI-driven System for Early De...	<1%
6	Publication	Sameh Taqatqeh, Mary Agoyi. "Early warning and recommendation system for un...	<1%
7	Internet	elar.urfu.ru	<1%
8	Internet	nova.newcastle.edu.au	<1%
9	Submitted works	Computer Engg on 2026-03-23	<1%
10	Publication	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Co...	<1%

11	Submitted works	Higher Education Commission Pakistan on 2025-10-15	<1%
12	Internet	mdpi-res.com	<1%
13	Internet	mfuir.mfu.ac.th	<1%
14	Publication	V. Sekhar, P. Ajay Kumar Reddy, K. Logesh, T. Raghunadha Reddy, Dara Raju. "Ne...	<1%
15	Internet	mental.jmir.org	<1%
16	Internet	assets-eu.researchsquare.com	<1%
17	Internet	dokumen.pub	<1%
18	Publication	Narayanan, Usha. "Patient Behavior Prediction Using Big Data Analytics and Mac...	<1%
19	Submitted works	Universiti Tenaga Nasional on 2022-10-03	<1%
20	Submitted works	University of Newcastle on 2021-06-02	<1%
21	Internet	www.researchsquare.com	<1%
22	Publication	Yueming Hu, Abhnil Prasad, Merlinde Kay. "Smart integration of aerosols for DNI ...	<1%
23	Publication	L. J. Cao, Francis E. H. Tay. "Chapter 38 Feature Selection for Support Vector Machi...	<1%
24	Internet	web.realinfo.tv	<1%

25	Submitted works	Dublin Business School on 2025-08-28	<1%
26	Publication	Lilibeth Azuela, Herbert Azuela, Anna Lyza Felipe, Armin Aragoncillo, Mia Eleazar. ...	<1%
27	Internet	ijece.iaescore.com	<1%
28	Submitted works	Bournemouth University on 2025-09-04	<1%
29	Publication	Shajid Hossain Hemal, Md. Ashikur Rahman Khan, Ishtiaq Ahammad, Masudur Ra...	<1%
30	Submitted works	University of Hull on 2025-04-24	<1%
31	Submitted works	University of Westminster on 2026-07-02	<1%
32	Publication	Hongwen Gu, Yuqi Zhang. "Efficient and Interpretable Machine Learning for Stud...	<1%
33	Internet	www.tandfonline.com	<1%
34	Publication	Paolo Ferro, Harinadh Vemanaboina, Chander Prakash. "Computational Techniqu...	<1%
35	Publication	Shankar Babu, Mahesh Babu Kota. "Synergies in Smart and Virtual Systems using...	<1%
36	Internet	jirbdai.com.ng	<1%
37	Publication	Hamad Almaghrabi, Ben Soh, Alice Li, Idrees Alsolbi. "SoK: The Impact of Educatio...	<1%
38	Publication	N. V. R. Naidu, N. L. Ramesh, A. Jagannatha Reddy, Nagaraju Kottam et al. "Applie...	<1%

39	Publication	Sukhpreet Kaur, Amanpreet Kaur, Manish Kumar. "Recent Advances in Computati...	<1%
40	Internet	doaj.org	<1%
41	Internet	www.preprints.org	<1%
42	Internet	www.seruvenyayinevi.com	<1%
43	Publication	Niall Sclater. "Learning Analytics Explained", Routledge, 2017	<1%
44	Internet	arxiv.org	<1%
45	Internet	e-jurnal.pnl.ac.id	<1%
46	Internet	www.ijprems.com	<1%
47	Publication	Albert Sire Langa, Ramon Reig-Bolaño, Clara Masó-Albareda, Ariadna Farrés-Serr...	<1%
48	Submitted works	KCA University on 2025-04-22	<1%
49	Publication	Pankaj Chittora, Sandeep Chaurasia, Prasun Chakrabarti, Gaurav Kumawat et al. ...	<1%
50	Publication	Paul A. Schutz, Krista R. Muis. "Handbook of Educational Psychology", Routledge, ...	<1%
51	Publication	Salma Boujmiraz, Hassane Darhmaoui, Ahmed Drissi el maliani. "Predicting stude...	<1%
52	Submitted works	The University of Notre Dame on 2024-07-29	<1%

53	Internet	slejournal.springeropen.com	<1%
54	Submitted works	Cavite State University on 2026-06-05	<1%
55	Submitted works	College of Public Health on 2026-03-17	<1%
56	Publication	Park, Eun Soo. "Evaluating Aspect-Based Sentiment Analysis in Healthcare Drug R..."	<1%
57	Submitted works	University of Warwick on 2026-02-09	<1%
58	Internet	ijisrt.com	<1%
59	Internet	mafiadoc.com	<1%
60	Submitted works	Liberty University on 2026-07-20	<1%
61	Submitted works	Midocean University on 2026-05-02	<1%
62	Submitted works	Midocean University on 2026-05-02	<1%
63	Publication	Romero-Zaldivar, V.A.. "Monitoring student progress using virtual appliances: A c..."	<1%
64	Publication	Saknee, Ayad. "Predictive Relationships Between Learning Management System E..."	<1%
65	Publication	Yar Muhammad, Muhammad Abul Hassan, Sultan Almotairi, Kawsar Farooq, Fabr...	<1%
66	Internet	achieverpapers.com	<1%

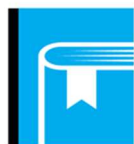
67	Internet	download.bibis.ir	<1%
68	Internet	elibrary-dev.nusamandiri.ac.id	<1%
69	Internet	nhess.copernicus.org	<1%
70	Internet	www.eriesjournal.com	<1%
71	Internet	www.researchgate.net	<1%
72	Publication	"Artificial Intelligence in Education Technologies: New Development and Innovati...	<1%
73	Publication	"Intelligent System Design Based on Soft Computing Models", Springer Science a...	<1%
74	Publication	Ayman Abu hajaj, Nadine Sheikh. "AI-Driven Intelligent Assessment System for Ev...	<1%
75	Submitted works	Bangladesh University of Professionals on 2022-02-11	<1%
76	Publication	Chetna Sharma, Amandeep Kaur, Harpreet Singh. "Artificial Intelligence for Medi...	<1%
77	Publication	Ghareebaih, Yaman Abu. "Car Price Prediction Using Machine Learning: Analyzin...	<1%
78	Publication	Huong Nguyen Thi Cam, Aliza Sarlan, Noreen Izza Arshad. "A hybrid model integr...	<1%
79	Publication	Ilyas Ibrahim Ilyas, Souley Boukari, Abdulsalam Ya'u Gital. "A Comparative Analy...	<1%
80	Submitted works	Islington College,Nepal on 2026-05-15	<1%

81	Submitted works	Liverpool John Moores University on 2023-04-29	<1%
82	Publication	Md. Tahidur Rahman, Md. Lotifur Rohman, Shyam Sundar Sarkar, Md. Fahim, Md....	<1%
83	Publication	Minh Nguyen, Tong He, Lijun An, Daniel C. Alexander, Jiashi Feng. "Predicting Alz...	<1%
84	Publication	Mohammed Fathy El-Amin. "Advances and Applications of Machine Learning in FI...	<1%
85	Submitted works	Strathmore University (Main Account) on 2026-03-27	<1%
86	Publication	Xi Jin, Zhonghui Wang, Xin Zhao, Yang Wen, Chunbo Qin. "Effect of Physical Fitnes...	<1%
87	Publication	Xinyi Wang, He Zhang. "A transformer-enhanced framework for estimating stude...	<1%
88	Internet	conference.uis.edu.my	<1%
89	Internet	educationaltechnologyjournal.springeropen.com	<1%
90	Internet	public-pages-files-2025.frontiersin.org	<1%
91	Internet	repository.aaup.edu	<1%
92	Internet	rsisinternational.org	<1%
93	Internet	www.gepsoft.com	<1%
94	Internet	www.ifets.info	<1%

95	Internet	www.ijraset.com	<1%
96	Publication	Balqis Albreiki, Zahiriddin Rustamov, Jaloliddin Rustamov, Nazar Zaki. "Chapter 1...	<1%
97	Publication	Bhabani Prasanna Pattanaik, Dharendra Kumar Shukla, Sambit Satpathy, Kamala...	<1%
98	Publication	C. Guedes Soares, F.C. Salvado. "Innovations in Maritime Technology and Enginee...	<1%
99	Submitted works	Kazakh-British Technical University on 2026-06-01	<1%
100	Submitted works	Leeds Trinity and All Saints on 2026-04-17	<1%
101	Publication	Mourade Azrou, Jamal Mabrouki, Sultan Ahmad. "IoT and Advanced Intelligence ...	<1%
102	Publication	Norhalina Senan. "Rough Set Approach for Attributes Selection of Traditional Mal...	<1%
103	Publication	Poonam Nandal, Tapas Kumar, Meeta Singh, Mamta Dahiya. "Progressive Compu...	<1%
104	Submitted works	Sim University on 2021-08-16	<1%
105	Submitted works	Teesside University (Blackboard Ultra) on 2026-05-01	<1%
106	Submitted works	University of Bedfordshire on 2022-05-20	<1%
107	Submitted works	University of East London on 2026-05-07	<1%
108	Submitted works	University of Leeds on 2023-09-04	<1%

109	Submitted works	University of Sunderland on 2026-07-02	<1%
110	Submitted works	University of Wolverhampton on 2026-05-30	<1%
111	Internet	acervodigital.ufpr.br	<1%
112	Internet	github.com	<1%
113	Internet	ijds.mtu.edu.iq	<1%
114	Submitted works	info.administrator on 2026-07-13	<1%
115	Internet	jurnal.univrab.ac.id	<1%
116	Internet	odr.chalmers.se	<1%
117	Internet	ojs.uajy.ac.id	<1%
118	Internet	sitams.ac.in	<1%
119	Internet	www.coursehero.com	<1%
120	Internet	www.frontiersin.org	<1%
121	Internet	www.inderscience.com	<1%
122	Publication	Abdelkarim Bettahi, Hamid Harroud, Fatima-Zahra Belouadha. "Early Student Ris...	<1%

123	Submitted works	Asia Pacific University College of Technology and Innovation (UCTI) on 2026-06-04	<1%
124	Submitted works	Coventry University on 2025-08-15	<1%
125	Publication	Danish Javed, NZ Jhanjhi, Farzeen Ashfaq, Navid Ali Khan, Shampa Rani Das, Sure...	<1%
126	Publication	Kaur, Davinder. "Trustworthy AI: Ensuring Explainability and Acceptance", Purdu...	<1%
127	Publication	M. Skittou, M. Merrouchi, T. Gadi. "Development of an Early Warning System to S...	<1%
128	Submitted works	Tilburg University on 2026-07-06	<1%
129	Submitted works	University of Newcastle on 2023-08-28	<1%
130	Publication	"Smart Trends in Computing and Communications", Springer Science and Busine...	<1%
131	Publication	Anshul Verma, Pradeepika Verma. "Research Advances in Intelligent Computing",...	<1%
132	Submitted works	College of Public Health on 2026-03-12	<1%
133	Submitted works	Leeds Beckett University on 2026-04-22	<1%
134	Publication	Nilanjan Dey, Bitan Misra, Sayan Chakraborty. "Smart Medical Imaging for Diagn...	<1%
135	Submitted works	University of East London on 2026-02-24	<1%
136	Submitted works	University of Education on 2026-07-08	<1%



An AI-Based Early Warning System for Academic Risk Prediction Using Machine Learning and Learning Management System Data

Marya Mohammed Bawazir– Hala Haytham Hajjar

ORCID ID: 253000692 - ORCID ID: 253001456

Specialization – College – University Affiliation – Country (Font Size 12)

Email (Font Size 12 in Black)

Dr. Hager Saleh (ORCID ID) Specialization, College, University, Country, Email:

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 System 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 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 two 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: Artificial Intelligence, Early Warning System, Academic Risk Prediction, Learning Management System (LMS), Machine Learning(ML), Feature Selection, SMOTE, XGBoost.

1. Introduction:

Improving the quality of education has become a strategic priority worldwide. In Vision 2030 of Saudi Arabia, the transformation of the education system involves integrating advanced technologies to enhance teaching, learning, and educational outcomes. This transformation has been supported through platforms such as **Madrasati**, the national learning platform serving K–12 education, and **Classera**, an LMS used by educational institutions to manage learning activities, assessments, communication, and student engagement. Six million students benefited from Madrasati during the 2020–2021 academic year, and nearly 98% accessed the platform of digital learning across the Kingdom (Boni). Classera illustrates the growing use of integrated LMS that provide detailed academic and behavioral data suitable for learning analytics and student-performance prediction (Alhujayri, 2024).

LMS have evolved content to become valuable sources of educational data. It collects students' academic and behavioral information, including attendance, assignments, examinations, participation, and interactions with learning resources. These data provide educational institutions with information that can help to monitor students' progress and support evidence-based educational decisions (Romero & Ventura, 2020)

Despite these advances, student assessment in many educational settings still uses traditional evaluation methods such as examinations, assignments, and instructors' observations. While these methods measure academic achievement, they often provide retrospective evaluations and cannot effectively utilize the large amount of data generated by LMS platforms. Consequently, it will be a challenge to identify at-risk students in the early stage (Alwarthan et al., 2022; Romero & Ventura, 2020) AI, particularly ML, has emerged as an effective solution for educational prediction tasks. By analyzing historical LMS data, ML models can identify hidden learning patterns, predict students' academic performance, and support early intervention strategies that improve learning outcomes. As a result, AI has become an important component of Educational Data Mining and Learning Analytics (Baker et al., 2016; Romero & Ventura, 2020).

Although numerous studies have applied machine learning techniques to predict student performance, most have relied on public datasets or data collected from institutions outside Saudi Arabia. However, relatively limited research has investigated prediction models using private datasets from Saudi LMS education (Alwarthan et al., 2022; Romero & Ventura, 2020) To address these challenges, we propose an AI-Based Early Warning System for predicting students' academic risk using private educational data extracted from KSA LMS. The proposed framework integrates data preprocessing, using RF for feature selection, class balancing using the Oversampling Technique (SMOTE), supervised ML algorithms, and Grid Search-based hyperparameter optimization.

2 Previous Studies: Several studies have developed and applied machine learning models to improve students' academic performance through early prediction.

CR-NODE was proposed by Bettahi et al (Bettahi et al., 2025), 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. (Ujkani et al., 2024) 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%.

The proposed RADAR system by Embarak et al (Embarak & Hawarna, 2024) 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 the 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 machine learning system incorporating XAI can process multiple kinds of data and take into account other factors associated with the risks posed to students.

Macfadyen (Macfadyen & Dawson, 2010) 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 Learning Management System (LMS) at the University of British Columbia in Canada. The data sample contained logs of student activities such as participation in discussions, completion of assessments, sending emails, interactions, and frequency of logging into the LMS. Major predictors of student performance were established; they included participation in discussions, messaging, performance in tests/ quizzes, and frequency of logging into the LMS. Data cleaning was conducted, which entailed the elimination of observations that had missing values and structuring for predictive modeling (PM). The regression model (RM) technique was adopted for developing the predictive model (PM) for academic success using activity data from the LMS. According to the findings, activity data of students from the Learning Management System 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.

Queiroga et al. (Queiroga et al., 2022) presented a research paper that aimed to predict academically at-risk secondary school students using machine learning 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 (ANEP) 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, encompassing their grades, attendance, demographics, and other features. 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. Machine learning algorithms such as RF, DT, AdaBoost, LR, NB, and MLP were investigated. However, the latter model was preferred since it had a better performance and was more interpretable. From the analysis, it is clear that the ML models have done remarkably well through achieving a very high accuracy rate when predicting at-risk students. The AUROC and F1 scores produced a result above 0.90 and 0.88, respectively.

Bagunaid (Bagunaid et al., 2024) 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 the performance of the students as well as 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. The dataset comprised demographic information, performance information, and learning behaviors data. 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 machine learning 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%.

Johora (Johora et al., 2025) 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. It included demographic, academic, and behavioral data of the students. Preprocessing steps were followed, including handling of the missing data, elimination of the 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 it can be seen, it was proved that 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.

Christou et al. (Christou et al., 2023) have proposed a machine learning 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, Greece, including Physics, Philosophy, Primary Education, History, Mathematics, and Chemistry. It contained students' academic information such as grades, course-related data, and study time. The preprocessing stage involved organizing, cleaning, and transforming the academic records into suitable input features for machine learning models. The proposed FSC4RBF method was evaluated against several classification techniques, including Decision Tree (DT), Bayesian Network, Support Vector Machine (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. (Realinho et al., 2022) 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 student dropout and academic success. The data were collected from multiple institutional sources, merged, cleaned, and preprocessed to support machine learning classification tasks. Several ML techniques were applied to analyse student

performance and identify at-risk students. The study used a public dataset from the Polytechnic Institute of Portalegre, Portugal, containing 4424 student records with 35 attributes. The dataset included demographic, macroeconomic, socioeconomic, and academic performance data from enrollment until the end of the first and second semesters. Data collected from multiple sources, such as academic systems, national databases, The data were then combined into one single dataset. Data cleaning was done by eliminating duplicates, managing missing data, managing outliers, renaming the variable names, and selecting suitable features. The study applied RF, XGBoost, LightGBM, and CatBoost after performing data integration, preprocessing, and feature importance analysis using permutation feature importance. The project did not include the 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 student drop-out and student achievement and can be used for benchmarking ML models and building learning analytics systems.

The employment of AI techniques to predict student academic performance within the BL environment of SEU was first done by Hamadneh (Hamadneh et al., 2022)). 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 the Blackboard LMS. The data contained the following predictors: type of study, participation in live lectures, exam results, and completion of assignments. Final exam marks were the target variable. Various 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.

The project of Skittou et al (Skittou et al., 2023) 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.

David Bañeres et al. (Bañeres et al., 2020) 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. They can see that, starting with the fall semester of 2016, the number of training registries rises every semester (20161). For every course and semester, they suggested a way to find the optimal classifier and training set. 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. The EWS uses GAR to modify the kind of message that is sent to the student. The third contribution involved using the 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 examine 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.

Albreiki, Balqis, et al. (Albreiki et al., 2021) 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. Time-wise, 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.

With various ML algorithms, Khan, Ijaz, et al. (Khan et al., 2021) created a set of prediction models. The dataset was collected from a course delivered at Buraimi University College, Sultanate of Oman, and covered three academic semesters. It consisted of 151 student records described by ten predictive features and one target class used for model development. Data cleaning: 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 machine learning 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. The improvement of the struggling students served as a gauge for the overall effects. Four out of five students were able to produce an end result that was acceptable. From the beginning of the course, students' performance is closely monitored as part of the supportive measures. Additionally, a series of preventive measures is implemented to provide extra support to students who are having difficulty.

Akçapınar et al. (Akçapınar et al., 2019) 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. Both final term academic achievement scores and interaction among students data from the virtual classroom were taken into account in the analysis. 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**.

Table 1 Of The Reviewed Papers

Paper	Year	Models	Limitation	Dataset	Type	The best model	Result
(Bettahi et al., 2025)	2025	CR-NODE LR, RF, XGBoost, Gradient Boosting, LSTM, and Basic Neural ODE	It focuses exclusively on predictive accuracy without evaluating whether deployment actually improves student outcomes.	large public university Dataset.	Private - Need request-.	CR-NODE	CR-NODE achieves the best balance
(Ujkani et al., 2024)	2024	RF Classifier, Gradient Boosting Classifier, KNN, Multi-Layer Perceptron (Neural Network), First Custom Neural Network Architecture, Second Custom Neural Network Architecture, XAI, and LSTM.	The dataset was from a pre-pandemic time. it is vital to consider the broader context of how the COVID-19 pandemic may have impacted students' habits, motivations, and learning practices.	Open University Learning Analytics Dataset	Public https://analyse.kmi.open.ac.uk/open_dataset	Custom Neural Network Architecture	The proposed models performed with the best accuracy of up to 94% compared to models for predicting students' success
(Embarak & Hawarna, 2024)	2024	Data analysis, ML, feature selection, XAI, RADAR	It relied on a specific dataset from a specific computer major, which may limit the generalizability of the findings to other contexts.	Students from a school district Dataset.	Private	RADAR system	The RADAR system can help students who are at risk receive better

							support and direction
(Macfadyen & Dawson, 2010)	2010	Correlation analysis, Regression model, LR, Network analysis	Applied to one course only, a small dataset, not a full AI model, limited generalization	LMS dataset from Blackboard (student activity, messages, logins, assessments)	Private	LR model	Predicted at-risk students with 81% accuracy and explained 30% of the variation in final grades
(Queiroga et al., 2022)	2022	RF, DT, AdaBoost, MLP, NB, LR	Needs annual data preprocessing and retraining; results based on Uruguay data only; one model showed bias and weak performance.	Nationwide K-12 dataset (Uruguay), 258,440 students; demographic data, grades, absences, social welfare, school data (2015–2020)	Private	RF	Achieved high performance (AUROC > 0.90, F1 > 0.88) and successfully predicted at-risk and dropout students early
(Bagunaid et al., 2024)	2024	ID2QN (Deep Reinforcement Learning), MV-MOC Clustering, Att-CapsNet (DL), Federated Learning (CGAN), TSE (Entropy), LWEA + Blockchain	High computational complexity; requires multi-modal biometric data; privacy concerns; tested mainly on the OULAD dataset	OULAD dataset (32,593 students, 22 courses); demographic, academic, and behavioural LMS data	Public No Access	ID2QN (Improved Duelling Deep Q Network)	Achieved ~97% accuracy and ~95% F1-score; outperformed PCT, MST-FaDe, Fuzzy SVM, and AISAR; effectively predicted at-risk students early
(Johora et al., 2025)	2025	RF, Gradient Boosting, XGBoost, SVC, KNN, DT, LR, Stacking Ensemble, SHAP, LIME	Small dataset (872 students) from Bangladesh only; limited generalizability; based on survey data	872 undergraduate students from multiple universities; demographic, academic, and behavioural features	Private	Stacking Ensemble (RF + GB + SVC)	Achieved 86.38% accuracy; SHAP and LIME identified key factors such as attendance affecting student performance

				(31 features)			
(Christou et al., 2023)	2023	RBF Neural Network, SVM, DT, Bayesian Network, Neural Networks	Data from one university only; limited generalization	Student academic records from several departments, University of Ioannina	Private	Feature selection and construction method for RBF network	Achieved the highest accuracy and the lowest error compared with other models in predicting grades and dropout
(Realinho et al., 2022)	2022	Random Forest, XGBoost, LightGBM, CatBoost	An imbalanced dataset, limited to one institution, and mainly focused on dataset description rather than proposing a strong predictive model.	Student Dropout and Academic Success Dataset – Polytechnic Institute of Portalegre	Public No Access	Not specified	Identified the most important factors for dropout prediction, such as grades, approved courses, tuition status, and course type
(Hamadneh et al., 2022)	2022	ANN, Multilayer Perceptron Neural Network, Multilayer Perceptron Neural Network with Firefly Algorithm	Limited to one university, a small dataset, and a few features	Dataset from the Blackboard Learning Management System at the Saudi Electronic University	Private	Neural Network trained by Firefly Algorithm	Reported strong predictive performance of 0.39 and highlighted midterm exams, assignments, attendance, and study mode as key factors
(Skittou et al., 2023)	2023	SVM, RF, SGD, and KNN.	It is limited by the absence of administrative interfaces that enable dynamic, interactive, and efficient data management, because aggregating data from several operational systems	They used a tool built specifically for this project by aggregating data from several operational systems	Private-need to request access from the author	(KNN) algorithm	The findings were presented using a custom-developed Django web application, illustrating its effectiveness in supporting educational planning and informed decision-making.
(Bañeres et al., 2020)	2020	DT, NB, SVM, LR,	The model was course-specific	The model	Private	Gradual At-Risk	The model showed a

2020)et al., 2020)		Hierarchical Mixed models, KNN, Neural Network models	and becomes invalid if the assessment procedure is modified, including changes to the number or substance of assessment activities.	was tested using data available in an institutional data mart, where curated data from six semesters are available.		(GAR) model	high accuracy in correctly detecting at-risk students.
(Albreiki et al., 2021)	2021	Statistical and ML techniques	still lack The results of efficiency in terms of time and have high false predictions. They propose a customized rule-based model that offers timely identification for at-risk students.	It was collected from a programming course taught for undergraduate students at (CIT), United Arab Emirates University (UAEU).	Private - Need request-.	Customized rule-based system (RBS)	Visualized at-risk students to assist educators in identifying them and providing timely interventions.
(Khan et al., 2021)	2021	ML algorithms KNN RepTree DT (MLP) (ANN)	The training dataset may contain a limited number of instances due to constraints on classroom size.	The study utilized students' academic records from a course offered at Buraimi University College (BUC).	Private - Need request-.	Developed a set of prediction models	The prediction model proved to be an effective and valuable tool for supporting both instructors and students.
(Akçapınar et al., 2019)	2019	Prediction model Classification model	It had limited Transferability and generalizability of prediction models into different courses.	The study involved 76 undergraduate students from a public university in Turkey.	Private - Need request-	(KNN)	The kNN model achieved an accuracy of 89% in identifying students at risk of academic failure by the end of the semester.

3. Methodology: This section presents the methodology adopted to design, implement and evaluate the proposed framework, AI-Based Early Warning System for Academic Risk Detection in education.

Figure 1 shows the main steps of the proposed framework to identify the risk of academic achievement using a private educational dataset collected from 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.

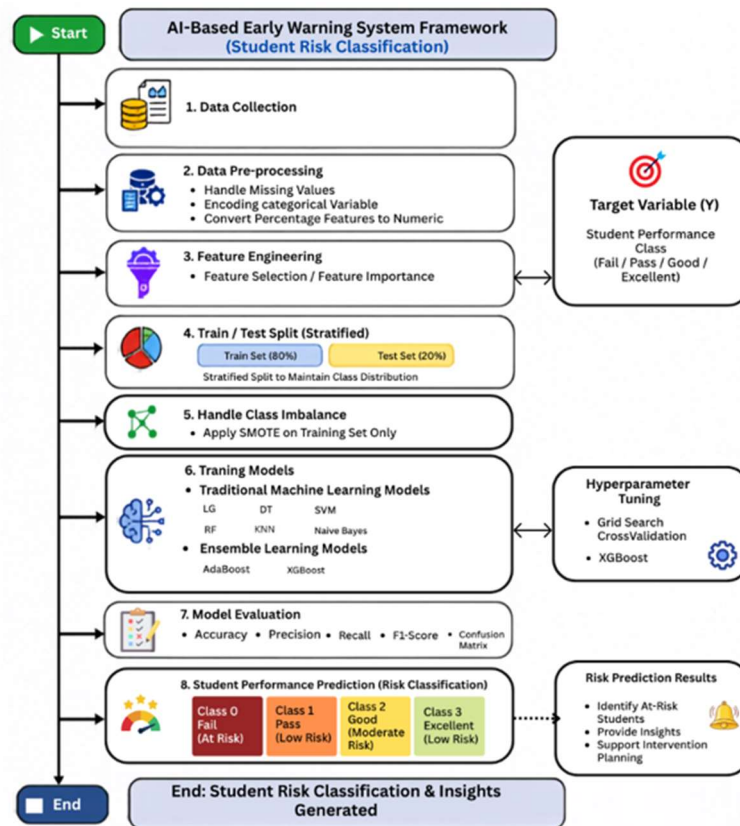


Figure 1 . Proposed AI-Based Early Warning Framework

The steps will be detailed as follows:

4.1 Dataset Description

We used a private educational dataset collected from an LMS. The dataset contains **1,395 student records** represented by 17 educational attributes: participation, assessments, attendance, and student engagement. The feature selection experiments selected 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), while the remaining seven features (Interactive Contents, Discussion Rooms, Mailbox Participation %, Exams Participation %, Exams, Teachers' Messages, and Student Login %) were excluded based on their low importance scores, such as participation, assessments, attendance, and engagement measures. *Table 2* The target variable represents students' academic performance and classifies each student into one of the following four categories: Excellent, Good, Pass and Fail.

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

4.2 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.

4.3 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.

4.4 Handling imbalance dataset using SMOTE:

SMOTE is a technique for data balancing proposed as an improvement for 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 (Ujkani et al., 2024)

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.

4.5 Feature Selection

Feature selection was evaluated to determine which educational features made significant contributions in predicting the students' academic performance. Elimination of the less important variables makes the prediction process easier and less complicated. The steps involved in the feature selection include:

- **Feature Importance** is a method for assessing the relative contribution of each feature to an ML model. It assigns an importance score to every feature, where higher scores indicate a greater influence on the model's predictive performance.

4.6 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.

4.6.1 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 (Bishop & Nasrabadi, 2006)

4.6.2 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 (Bishop & Nasrabadi, 2006).

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.

4.6.3 Hyperparameter Optimization Procedure

In this study, it was applied to optimize the hyperparameters of the selected 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, the optimized models were retrained using the entire training dataset before being evaluated on the independent testing dataset.

The hyperparameter search space explored during the optimization process is summarized in Table 3.

Table 3 Hyperparameter Search Space Used for Grid Search Optimization

Model	Hyperparameter Search Space
Decision Tree	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}
Random Forest	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}
XGBoost	n estimators = {100, 200}; learning rate = {0.01, 0.05, 0.1}; max depth = {3, 5, 7}; subsample = {0.8, 1.0}; colsample_bytree = {0.8, 1.0}

Table 3 summarizes the range of hyperparameter values evaluated during the Grid Search optimization process for the selected 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 selected hyperparameter values were subsequently used to retrain the optimized models before conducting the final performance evaluation.

4.6 Machine Learning Models

ML 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)(Bishop & Nasrabadi, 2006)

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.(Hastie, 2009)
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.(Hastie, 2009)
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 (Bishop & Nasrabadi, 2006)
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.(Bishop & Nasrabadi, 2006).
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(Studies, 2024)
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.

4.7 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 (Bañeres et al., 2020; Hastie et al., 2009).

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

1- AdaBoost

One of the oldest and most 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.

$$H(x) = \text{sign}(m=1 \sum \alpha_m h_m(x))$$

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- 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 (Bañeres et al., 2020)

The output of XGBoost is calculated as a sum of the outputs of all decision trees:

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

where

$\hat{y}_i$ refers to the predicted output, K It refers to the total number of prediction trees used, $f_k(x_i)$ Refers to the output of the k^{th} decision tree.

The objective function of XGBoost is composed of two parts: the training loss and the regularization part.

4.8 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.

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

Where: (TP) True Positives, (TN) True Negatives, (FP) False Positives, (FN) False Negatives. It is simple and intuitive; however, not deal with imbalanced datasets (Hastie et al., 2009).

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

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

It is important when the cost of false positives is high (Hastie et al., 2009).

3. **Recall:** also known as sensitivity, evaluates the model's ability to correctly identify actual positive instances.

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

When there is a positive case (like a student who is at risk), recall is very important (Hastie et al., 2009).

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.

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

It is particularly valuable when evaluating models on imbalanced datasets because it provides a balanced assessment by considering both precision and recall (Hastie et al., 2009).

5. **Confusion Matrix:** we used 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.

4. Results and Discussion:

4.1 Experimental Setup

Experiments were run using Google Colab. Python 3 was used as an implementation environment. 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 ensure a fair comparison, all ML models were optimized using Grid Search with cross-validation.

4.2 Class Distribution

From the information represented in Figure 2, 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 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.

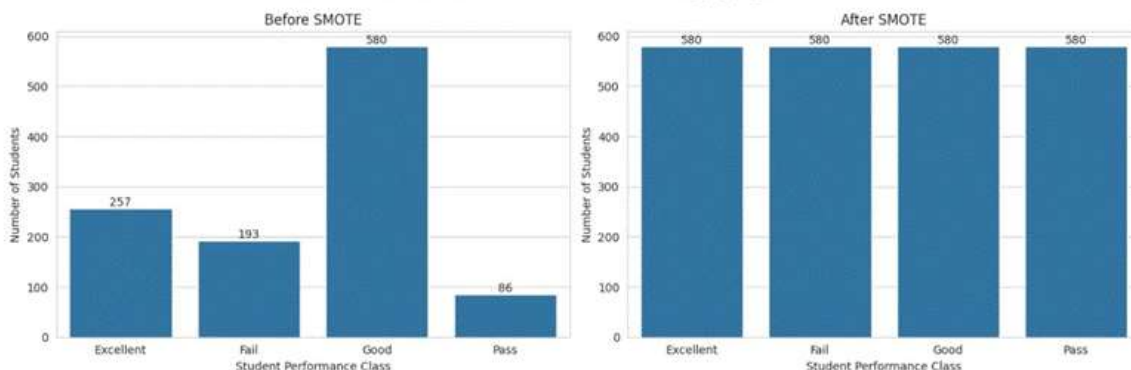


Figure 2 . Distribution of Student Performance Classes with SMOTE.

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 Figure 3, 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.

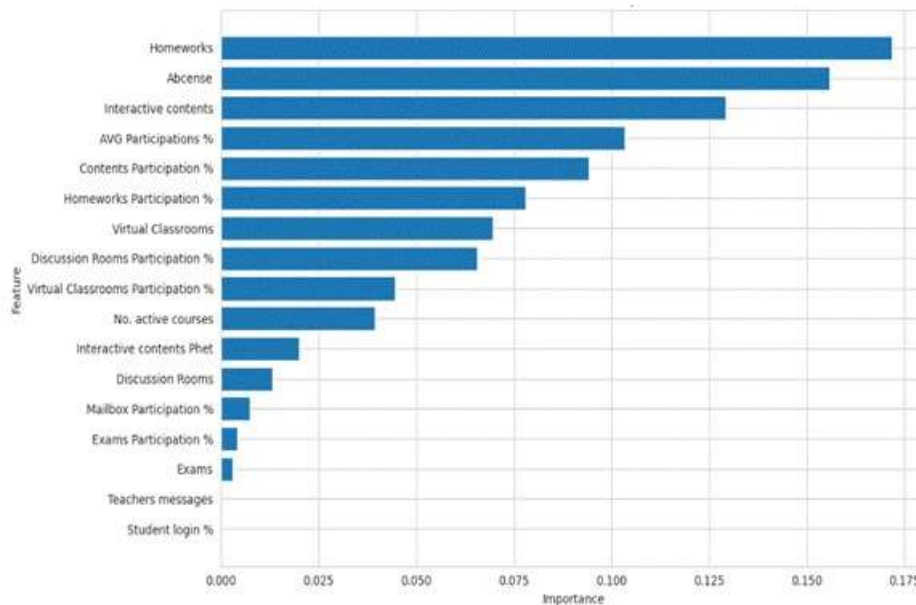


Figure 3. Random Forest Feature Importance Ranking.

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; 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 4 and 5 describe the two experimental setups in brief:

- **Result without applying SMOTE:** Full Feature Set+ Selected Features.
- **Result with applying SMOTE:** Full Feature Set+ Selected Features.

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 4 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 ACC; RF achieved 79.21 ACC. In contrast, the lowest performance across all evaluation metrics was NV. 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 81.00ACC. 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 4 Results of Applying Optimized Traditional and Ensemble Learning 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 5* and *Figure 4* Show 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.

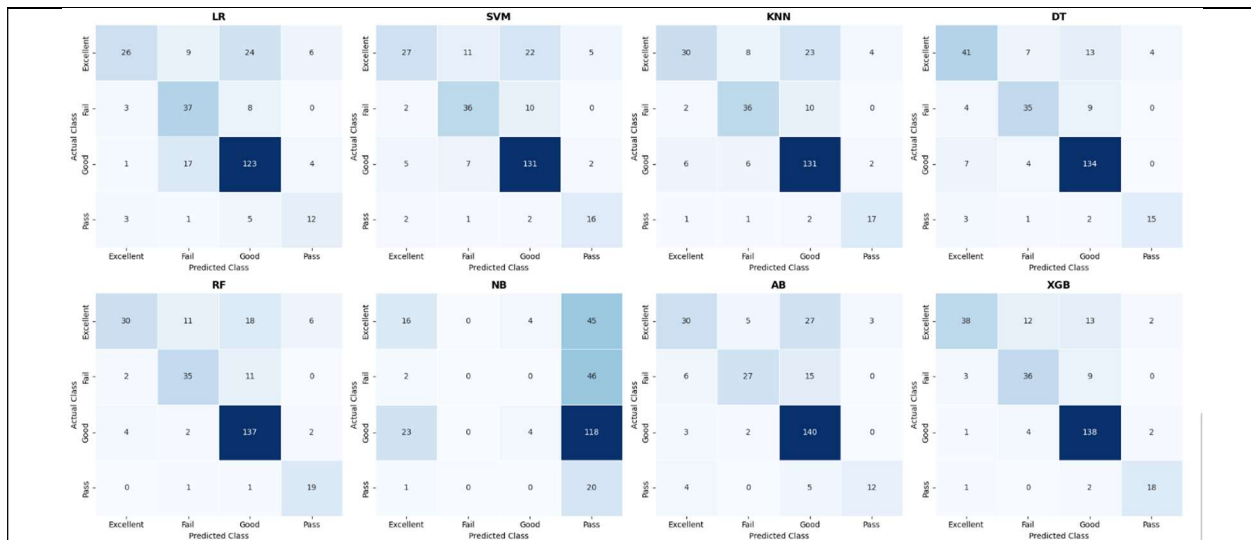


Figure 4 Confusion Matrices of the Optimized Traditional and Ensemble Learning Models without SMOTE-full feature.

Full feature without applying SMOTE

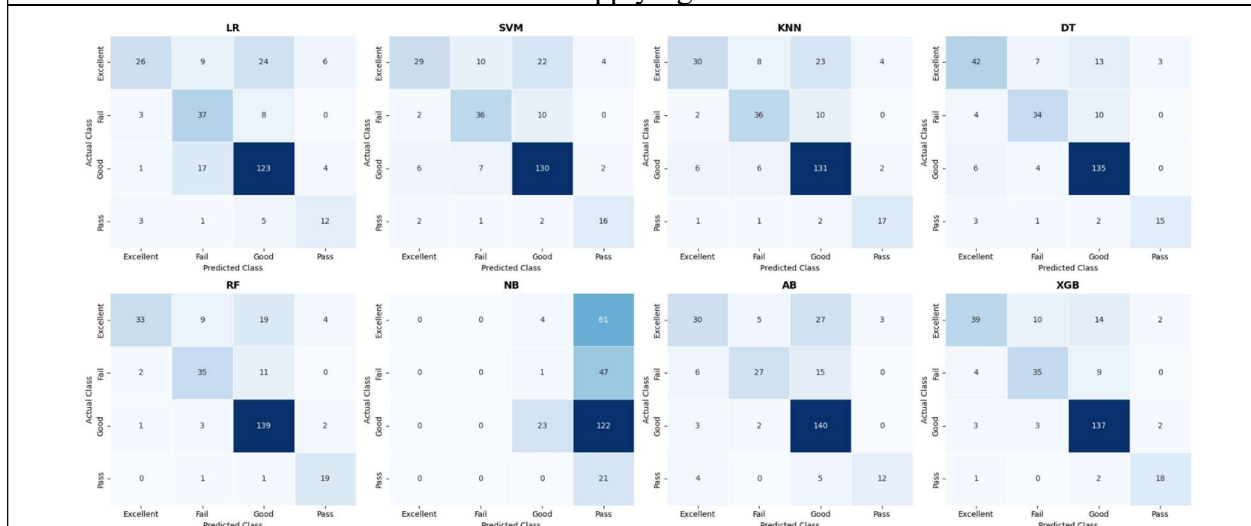


Figure 5 Confusion Matrices of the Optimized Traditional and Ensemble Learning Models without SMOTE-selected features.

Selected feature without applying SMOTE

4.6 Results with Applying SMOTE.

Table 5 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, NB 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 5 Results of Applying Optimized Traditional and Ensemble Learning 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 6 and Figure 7 The confusion matrices demonstrate that applying SMOTE improved the classification balance across the four classes. Figure X illustrates 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.

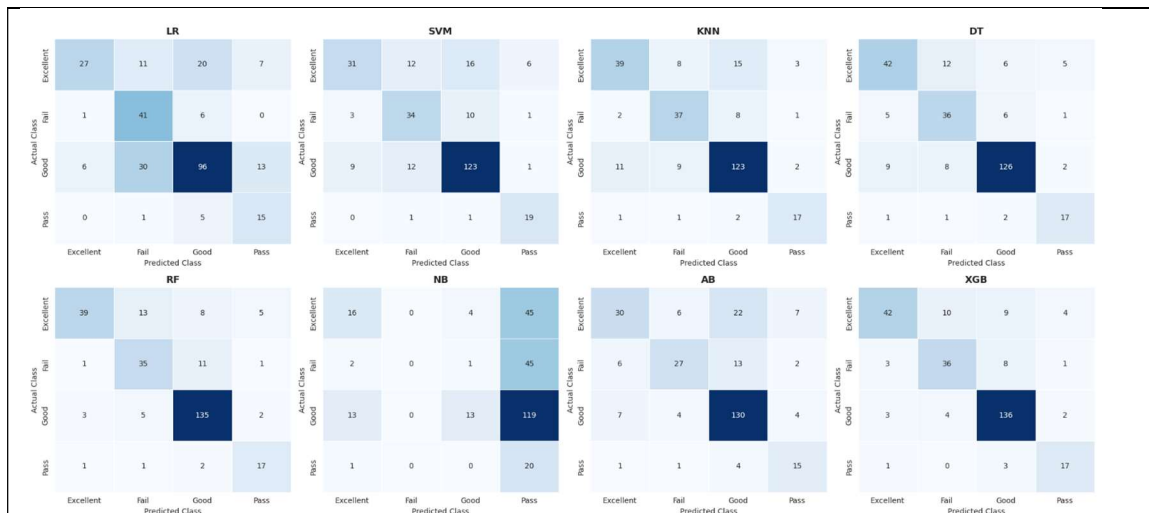


Figure 6 Confusion Matrices of the Optimized Traditional and Ensemble Learning Models with SMOTE-full feature.

Full feature with applying SMOTE

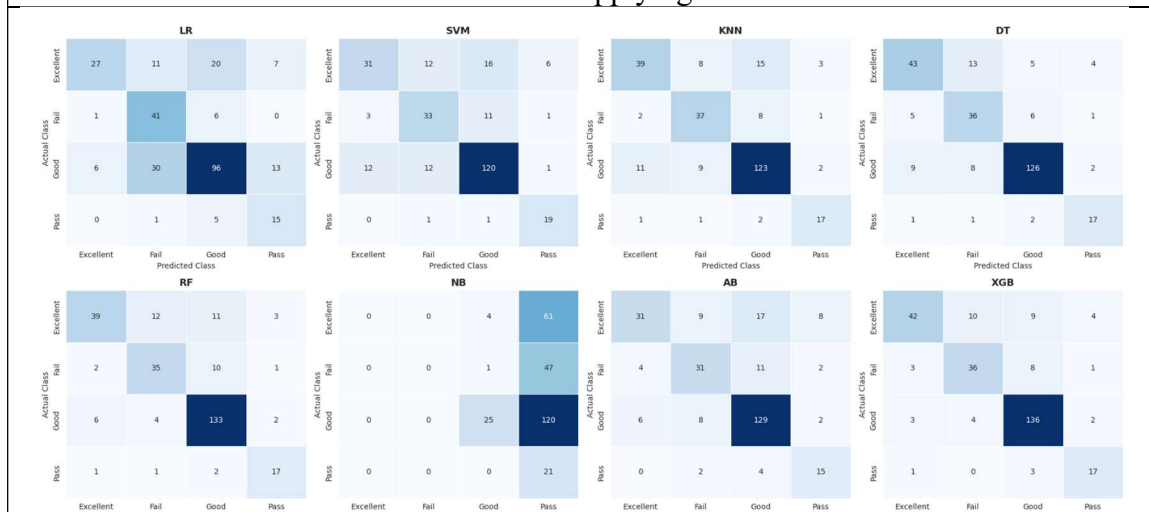


Figure 7 Confusion Matrices of the Optimized Traditional and Ensemble Learning Models with SMOTE-full feature.

Selected feature with applying SMOTE

Discussion

Figure 8 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. Overall, the results show that the effect of SMOTE and feature selection varied depending on the learning algorithm.

XGB achieved the highest accuracy in all configurations, with performance percentage 82.08% of ACC to 82.80% of ACC, showing stable performance regardless of the feature set or the use of SMOTE. DT and RF also demonstrated competitive results, with accuracies 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. NV 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 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.

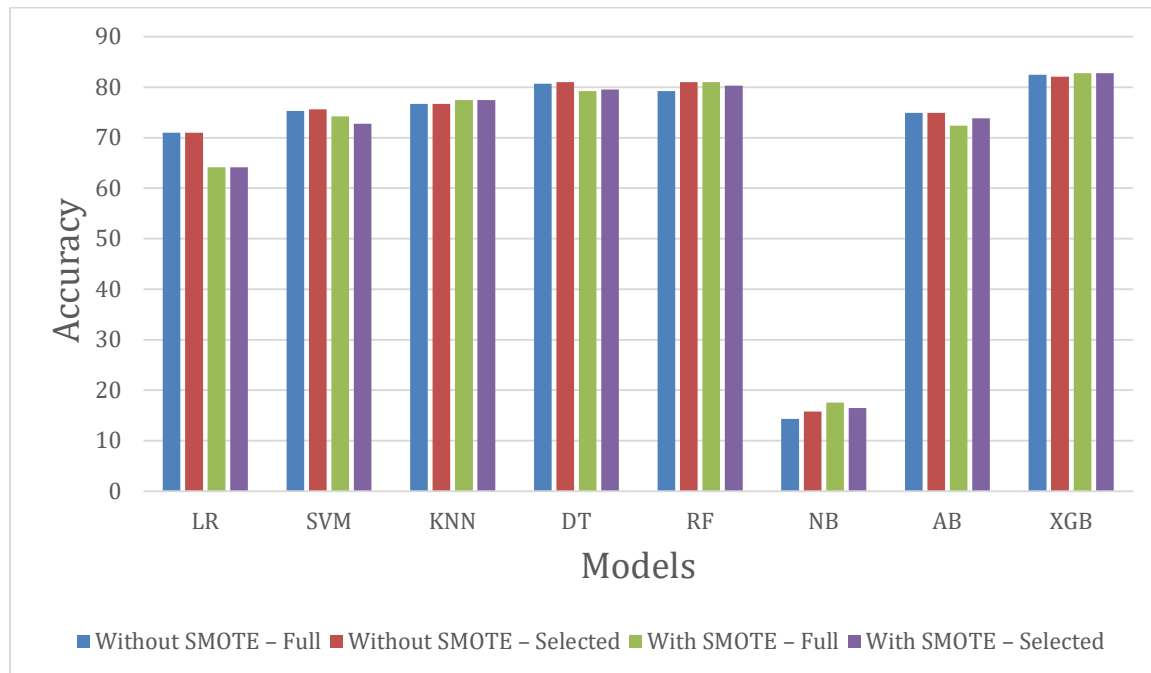


Figure 8 Accuracy comparison of Optimized ML models using full and selected feature sets with and without SMOTE.

5. Conclusion: This study presented an AI-based Early Warning System 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. 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 Learning Management Systems to support real-time academic risk prediction and timely educational intervention.

References list:

- Akçapınar, G., Altun, A., & Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education*, 16(1), 40.
- Albreiki, B., Habuza, T., Shuqfa, Z., Serhani, M. A., Zaki, N., & Harous, S. (2021). Customized rule-based model to identify at-risk students and propose rational remedial actions. *Big Data and Cognitive Computing*, 5(4), 71.

- Alhujayri, I. A. (2024). Transforming teaching through Classera: Insights from Saudi Arabian secondary schools. *Journal of Language and Cultural Education*, 12(1).
- Alwarthan, S. A., Aslam, N., & Khan, I. U. (2022). Predicting student academic performance at higher education using data mining: a systematic review. *Applied Computational Intelligence and Soft Computing*, 2022(1), 8924028.
- Bagunaid, W., Chilamkurti, N., Shahraki, A. S., & Bamashmos, S. (2024). Visual data and pattern analysis for smart education: a robust DRL-based early warning system for student performance prediction. *Future Internet*, 16(6), 206.
- Baker, R. S., Martin, T., & Rossi, L. M. (2016). Educational data mining and learning analytics. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, 379–396.
- Bañeres, D., Rodríguez, M. E., Guerrero-Roldán, A. E., & Karadeniz, A. (2020). An early warning system to detect at-risk students in online higher education. *Applied Sciences*, 10(13), 4427.
- Bettahi, A., Harroud, H., & Belouadha, F.-Z. (2025). Early Student Risk Detection Using CR-NODE: A Completion-Focused Temporal Approach with Explainable AI. *Algorithms*, 18(12), 781.
- Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.
- Boni, A. Saudi Arabia's Digital and Distance Education.
- Christou, V., Tsoulos, I., Loupas, V., Tzallas, A. T., Gogos, C., Karvelis, P. S., Antoniadis, N., Glavas, E., & Giannakeas, N. (2023). Performance and early drop prediction for higher education students using machine learning. *Expert Systems with Applications*, 225, 120079.
- Embarak, O. H., & Hawarna, S. (2024). Automated AI-driven system for early detection of at-risk students. *Procedia Computer Science*, 231, 151–160.
- Hamadneh, N. N., Atawneh, S., Khan, W. A., Almejalli, K. A., & Alhomoud, A. (2022). Using artificial intelligence to predict students' academic performance in blended learning. *Sustainability*, 14(18), 11642.
- Hastie, T. (2009). The elements of statistical learning: data mining, inference, and prediction. In: springer.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. In: Springer series in statistics New-York.
- Johora, F. T., Hasan, M. N., Rajbongshi, A., Ashrafuzzaman, M., & Akter, F. (2025). An explainable AI-based approach for predicting undergraduate students academic performance. *Array*, 26, 100384.
- Khan, I., Ahmad, A. R., Jabeur, N., & Mahdi, M. N. (2021). An artificial intelligence approach to monitor student performance and devise preventive measures. *Smart Learning Environments*, 8(1), 17.
- Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an “early warning system” for educators: A proof of concept. *Computers & education*, 54(2), 588–599.
- Queiroga, E. M., Batista Machado, M. F., Paragarino, V. R., Primo, T. T., & Cechinel, C. (2022). Early prediction of at-risk students in secondary education: A countrywide k-12 learning analytics initiative in uruguay. *Information*, 13(9), 401.
- Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. *Data*, 7(11), 146.
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 10(3), e1355.
- Skittou, M., Merrouchi, M., & Gadi, T. (2023). Development of an early warning system to support educational planning process by identifying at-risk students. *IEEE access*, 12, 2260–2271.

Studies, S. U. S. o. I. (2024). *Data Science vs. Machine Learning*. Syracuse University. Retrieved july 4 from <https://ischool.syracuse.edu/data-science-vs-machine-learning/>

Ujkani, B., Minkovska, D., & Hinov, N. (2024). Course success prediction and early identification of at-risk students using explainable artificial intelligence. *Electronics*, 13(21), 4157.