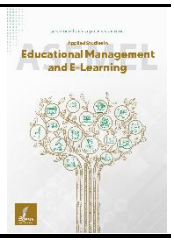
E-ISSN: | P-ISSN:

Applied Studies in Educational Management and E-Learning

www.asemel.reapress.com

Appl. Stud. Educ. Manag. E-Learn. Vol. 2, No. 3 (2025) 165–171.



Paper Type: Original Article

Comparison of Machine Learning Algorithms in Predicting Students' Academic Success and Failure with Emphasis on Performance Criteria

Sepideh Sabouri[1,*], Mohammad Shakouri[2]

[1] Science and Research Branch, Tehran Azad University, Tehran, Iran; Sepidesaboori46@gmail.com.
[2] Expert in Charge of Extracurricular Affairs, Iran University of Science and Technology, Tehran, Iran; M_shakouri@iust.ac.ir.

Citation:

Sabouri, S., & Shakouri, M. (2025). Comparison of machine learning algorithms in predicting students' academic success and failure with emphasis on performance criteria. *Applied studies in educational management and e-learning, 2*(3), 165-171.

Received: 12/ 10/2024 Reviewed: 17/11/2024 Revised: 19/12/2024 Accepted: 22/02/2025

Abstract

Purpose: Predicting students' academic status—including graduation, dropout, or retention—is an important issue in higher education management. This study aimed to evaluate the effectiveness of machine learning algorithms in predicting students' academic status.

Methodology: In this study, a dataset of 4424 samples and 35 individual, educational and socio-economic characteristics was used. Different machine learning algorithms including decision tree, random forest, support vector machine, XGBoost, LightGBM and CatBoost were implemented and compared. Data preprocessing was performed by scaling numerical features and balancing classes using SMOTE method. Cross-validation and Accuracy, F1-score and AUC metrics were used to evaluate performance.

Findings: The results showed that Boosting-based models, especially CatBoost and LightGBM, performed best in predicting educational status and were more accurate and stable than other algorithms. These algorithms were able to identify more complex patterns in the data and provide reliable performance.

Originality/Value: The findings indicate the high capacity of machine learning methods in analyzing educational data and supporting decision-making in higher education management. The use of advanced models such as CatBoost and LightGBM can be an effective tool for identifying students at risk of dropping out and improving educational planning.

Keywords: Machine learning, Academic status prediction, Student academic failure, CatBoost, LightGBM.

Corresponding Author: Sepidesaboori46@gmail.com https://doi.org/10.48313/asemel.v2i3.88

مطالعات کاربردی در مدیریت آموزشی و یادگیری الکترونیکی

دوره ۲، شماره (۳)، (۱۴۰۴)، ۱۷۱-۱۶۵

www.asemel.reapress.com

نوع مقاله: پژوهشی

مقایسه الگوریتم‌های یادگیری ماشین در پیش‌بینی موفقیت و افت تحصیلی دانشجویان با تاکید بر معیارهای عملکرد

سپیده صبوری[۱،*]، محمد شکوری[۲]

[۱] واحد علوم تحقیقات تهران، دانشگاه آزاد، تهران، ایران.

[۲] کارشناس مسئول امور فوق‌برنامه، دانشگاه علم و صنعت ایران، تهران، ایران.

چکیده

هدف: پیش‌بینی وضعیت تحصیلی دانشجویان، از جمله فارغ‌التحصیلی، ترک تحصیل یا ادامه تحصیل از موضوعات مهم در مدیریت آموزش عالی است. این پژوهش با هدف ارزیابی کارایی الگوریتم‌های یادگیری ماشین در پیش‌بینی وضعیت تحصیلی دانشجویان انجام شد.

روش‌شناسی پژوهش: در این مطالعه، مجموعه داده‌ای شامل ۴۴۲۴ نمونه و ۳۵ ویژگی فردی، تحصیلی و اجتماعی-اقتصادی مورد استفاده قرار گرفت. الگوریتم‌های مختلف یادگیری ماشین شامل درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان، *XGBoost*، *LightGBM* و *CatBoost* پیاده‌سازی و مقایسه شدند. پیش‌پردازش داده‌ها با مقیاس‌بندی ویژگی‌های عددی و متعادل‌سازی کلاس‌ها با روش *SMOTE* انجام شد. برای ارزیابی عملکرد، اعتبارسنجی متقابل و معیارهای *Accuracy*، *F1-score* و *AUC* به‌کار گرفته شد.

یافته‌ها: نتایج نشان داد که مدل‌های مبتنی بر *Boosting*، به‌ویژه *CatBoost* و *LightGBM*، بهترین عملکرد را در پیش‌بینی وضعیت تحصیلی ارایه کردند و نسبت به سایر الگوریتم‌ها دقت و پایداری بیشتری داشتند. این الگوریتم‌ها توانستند الگوهای پیچیده‌تری را در داده‌ها شناسایی کرده و عملکردی قابل‌اتکا ارایه دهند.

اصالت/ارزش افزوده علمی: یافته‌ها بیانگر ظرفیت بالای روش‌های یادگیری ماشین در تحلیل داده‌های آموزشی و پشتیبانی از تصمیم‌گیری در مدیریت آموزش عالی است [1]. استفاده از مدل‌های پیشرفته مانند *CatBoost* و *LightGBM* می‌تواند ابزار موثری برای شناسایی دانشجویان در معرض ترک تحصیل و بهبود برنامه‌ریزی آموزشی باشد.

کلیدواژه‌ها: یادگیری ماشین، پیش‌بینی وضعیت تحصیلی، افت تحصیلی دانشجویان، CatBoost، LightGBM.

۱- مقدمه

افت تحصیلی دانشجویان یکی از چالش‌های اساسی نظام‌های آموزش عالی در سراسر جهان است. شناسایی و پیش‌بینی دانشجویانی که در معرض ترک تحصیل قرار دارند، می‌تواند به اتخاذ سیاست‌های حمایتی و افزایش بهره‌وری آموزشی منجر شود. از این‌رو، پیش‌بینی وضعیت تحصیلی دانشجویان به یکی از موضوعات مهم در حوزه تحلیل داده‌های آموزشی تبدیل شده است [1].

در سال‌های اخیر، پژوهش‌های متعددی از روش‌های یادگیری ماشین برای پیش‌بینی موفقیت یا شکست تحصیلی استفاده کرده‌اند. به‌طور مثال، برخی مطالعات از الگوریتم‌های درخت تصمیم[1] یا رگرسیون لجستیک استفاده کرده‌اند و به نتایجی با دقت متوسط دست یافته‌اند. در مقابل، پژوهش‌های جدیدتر با بهره‌گیری از روش‌های یادگیری عمیق یا الگوریتم‌های پیشرفته‌تر مانند *XGBoost* نشان داده‌اند که می‌توان دقت پیش‌بینی را بهبود بخشید. با این‌حال، بخش عمده‌ای از این پژوهش‌ها تنها به یک یا دو الگوریتم محدود بوده و مقایسه جامع بین چندین الگوریتم پرکاربرد کمتر موردتوجه قرار گرفته است [2].

این پژوهش با هدف پر کردن این خلا، به مقایسه شش الگوریتم مطرح یادگیری ماشین شامل درخت تصمیم، جنگل تصادفی[2]، ماشین بردار پشتیبان[3]، *XGBoost*، *LightGBM* و *CatBoost* پرداخته است. نوآوری اصلی این مقاله در مقایسه جامع الگوریتم‌ها بر اساس معیارهای کلیدی عملکرد *Accuracy*، *F1-score* و *AUC* بر روی یک مجموعه داده واقعی شامل ۴۴۲۴ دانشجو و استفاده از روش‌های متعادل‌سازی داده[4] و تنظیم ابر پارامترها با *Optuna* است. نتایج این مطالعه علاوه بر مقایسه کمی الگوریتم‌ها، می‌تواند راهنمایی عملی برای انتخاب مدل مناسب در مسایل مشابه آموزشی ارایه دهد [3–1].

پیش‌بینی وضعیت تحصیلی دانشجویان یکی از چالش‌های مهم و بنیادین در مدیریت آموزشی و برنامه‌ریزی تحصیلات عالی است. دانشگاه‌ها معمولا با طیف گسترده‌ای از دانشجویان مواجه هستند که هرکدام شرایط فردی، اجتماعی، تحصیلی و اقتصادی متفاوتی دارند و این تنوع، پیش‌بینی مسیر تحصیلی آن‌ها را پیچیده می‌کند. در این راستا، درک درست از مفاهیم کلیدی مرتبط با موفقیت یا ناکامی تحصیلی نقش مهمی در تحلیل داده‌ها و طراحی مدل‌های یادگیری ماشین ایفا می‌کند.

ترک تحصیل

به دانشجویی گفته می‌شود که پیش از اتمام دوره تحصیلی، فرآیند آموزش را متوقف کرده و دانشگاه را ترک می‌کند. ترک تحصیل نه‌تنها باعث آسیب آموزشی و فردی می‌شود، بلکه از منظر مدیریتی نیز هزینه‌های زیادی بر نظام آموزش عالی تحمیل می‌کند. شناسایی دانشجویان در معرض خطر *Dropout*، امکان طراحی برنامه‌های حمایتی هدفمند، مداخله‌های آموزشی به‌موقع و کاهش نرخ ترک تحصیل را فراهم می‌سازد.

موفقیت تحصیلی

موفقیت تحصیلی معمولا با شاخص‌هایی مانند معدل، پاس کردن دروس، پیشرفت ترمی و میزان مشارکت در فعالیت‌های آموزشی سنجیده می‌شود. این شاخص‌ها تصویری شفاف از وضعیت علمی دانشجو ارایه می‌کنند و به مدیران کمک می‌کنند تا دانشجویان نیازمند حمایت آموزشی را زودتر شناسایی کرده و برنامه‌های بهبود عملکرد برای آنان طراحی کنند.

یادگیری ماشین نظارت شده

یادگیری نظارت شده یکی از روش‌های اصلی در علم داده است که هدف آن، آموزش یک مدل بر اساس داده‌هایی است که هم ویژگی‌های ورودی و هم برچسب خروجی مشخص دارند. در این روش، مدل الگوهای موجود در داده‌ها را یاد می‌گیرد تا بتواند برای نمونه‌های جدید پیش‌بینی دقیق انجام دهد. در پژوهش حاضر، الگوریتم‌های نظارت شده برای پیش‌بینی احتمال *Dropout* یا *Success* به‌کار گرفته شده‌اند و بررسی شده که کدام مدل دقت، *F1-score* و *AUC* بالاتری ارایه می‌دهد.

---

[1] Decision Tree
[2] Random Forest (RF)
[3] Support Vector Machine (SVM)
[4] Synthetic Minority Over-sampling Technique (SMOTE)

به‌طور کلی، ترکیب داده‌های تحصیلی، ویژگی‌های فردی و اطلاعات اقتصادی-اجتماعی با الگوریتم‌های یادگیری ماشین، امکان تحلیل عمیق‌تری از رفتار تحصیلی دانشجویان فراهم می‌کند. درک این مفاهیم پایه، مبنای توسعه یک سیستم پیش‌بینی موثر و علمی است که بتواند به بهبود کیفیت آموزش و کاهش نرخ ترک تحصیل کمک کند [4–1].

۲- پیشینه تحقیق

تحقیقات پیشین عمدتا بر الگوریتم‌های پایه مانند درخت تصمیم و رگرسیون لجستیک[1] تمرکز داشته‌اند و دقت متوسطی را در پیش‌بینی وضعیت تحصیلی گزارش کرده‌اند. با پیشرفت الگوریتم‌های یادگیری ماشین، پژوهش‌های جدیدتر از الگوریتم‌های تقویتی[2] مانند *LightGBM*، *XGBoost* و *CatBoost* برای افزایش دقت استفاده کرده‌اند و نشان داده‌اند که این روش‌ها در تشخیص دانشجویان در معرض ترک تحصیل و پیش‌بینی موفقیت تحصیلی بسیار موثرتر هستند.

با این‌حال، بیشتر مطالعات گذشته تنها به یک یا دو الگوریتم محدود بوده و مقایسه جامع بین چندین الگوریتم با معیارهای عملکرد متنوع Accuracy، *F1-score* و *AUC* بر روی یک دیتاست واقعی کمتر انجام شده است. این محدودیت نشان‌دهنده نیاز به پژوهش‌هایی است که الگوریتم‌های مختلف را به‌طور هم‌زمان و با استفاده از معیارهای دقیق عملکرد، ارزیابی و مقایسه کنند [4].

۲-۱- نمادهای ریاضی

برای مدل‌سازی پیش‌بینی وضعیت تحصیلی دانشجویان، می‌توان رابطه کلی زیر را در نظر گرفت:

$$Y = f\ (X1, X2, \dots, Xn).$$

۱. *Xi*: ویژگی‌های دانشجو شامل ۳۵ متغیر که اطلاعات فردی، تحصیلی و اجتماعی-اقتصادی را پوشش می‌دهند.
۲. *Y*: کلاس خروجی که نشان‌دهنده وضعیت تحصیلی دانشجو است *Dropout* یا *Success*.

این رابطه نشان می‌دهد که وضعیت تحصیلی *Y* یک تابع از ویژگی‌های مختلف *X1* تا *Xn* است و هدف الگوریتم‌های یادگیری ماشین، یافتن تابع *f* به‌گونه‌ای است که بتواند وضعیت دانشجویان جدید را با دقت بالا پیش‌بینی کند.

مساله پژوهش این است که وضعیت تحصیلی دانشجویان، فارغ‌التحصیلی، ترک تحصیل یا ادامه تحصیل با استفاده از داده‌های تحصیلی، شخصی و اجتماعی- اقتصادی پیش‌بینی شود. اهمیت این موضوع در شناسایی به‌موقع دانشجویان در معرض ترک تحصیل، کاهش نرخ *Dropout* و حمایت از تصمیم‌گیری موثر در مدیریت آموزش عالی است. محدوده پژوهش شامل ۴۴۲۴ نمونه با ۳۵ ویژگی و بررسی شش الگوریتم یادگیری ماشین (*DT*، *RF*، *SVM*، *XGBoost*، *LightGBM* و *CatBoost*) با معیارهای عملکرد *Accuracy*، *F1-score* و *AUC* است [5]، [6].

۲-۲- آماده‌سازی داده‌ها[3]

در مرحله اولیه، دیتاست شامل ۴۴۲۴ نمونه و ۳۵ ویژگی مورد بررسی قرار گرفت. ویژگی‌ها شامل اطلاعات فردی، تحصیلی و اجتماعی-اقتصادی هستند و انواع داده‌ها به دو دسته عددی و دسته‌ای تقسیم شدند.

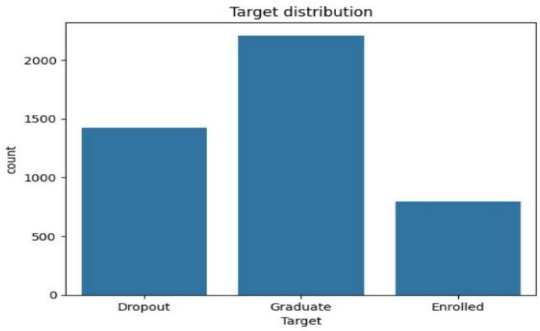
مراحل اصلی آماده‌سازی داده‌ها

۱. مدیریت داده‌های گمشده: مقادیر از دست رفته با میانگین (برای ویژگی‌های عددی) یا میانه/مد (برای ویژگی‌های دسته‌ای) جایگزین شدند.
۲. بررسی داده‌های پرت: نقاط غیرعادی شناسایی شدند و تصمیم به حذف یا نگه‌داشتن آن‌ها بر اساس تاثیر روی توزیع داده‌ها گرفته شد.

---

[1] Logistic Regression
[2] Boosting
[3] Data understanding and cleaning

۳. بررسی توزیع کلاس‌ها: نسبت کلاس‌ها نشان داد که دیتاست نامتوازن است، به‌طوری که تعداد دانشجویان *Dropout* کمتر از *Success* است. این مرحله باعث شد برای مراحل بعدی از روش *SMOTE* برای بالانس کلاس‌ها استفاده شود.



شکل ۱- توزیع کلاس‌ها در دیتاست قبل از پیش‌پردازش نشان‌دهنده نامتوازن بودن کلاس‌ها است.
Figure 1- The distribution of classes in the dataset before preprocessing indicates that the classes are unbalanced.

جدول ۱- اطلاعات پایه دانشجویان.
Table 1- Basic information of students.

شماره دانشجو	سن	معدل	جنسیت	حضور در کلاس	وضع اجتماعی و اقتصادی	وضعیت تحصیل
1001	20	3.2	مذکر	85%	متوسط	موفقیت‌آمیز
1002	21	2.5	مونث	70%	پایین	ترک تحصیل
1003	19	3.8	مونث	%92	بالا	موفقیت‌آمیز

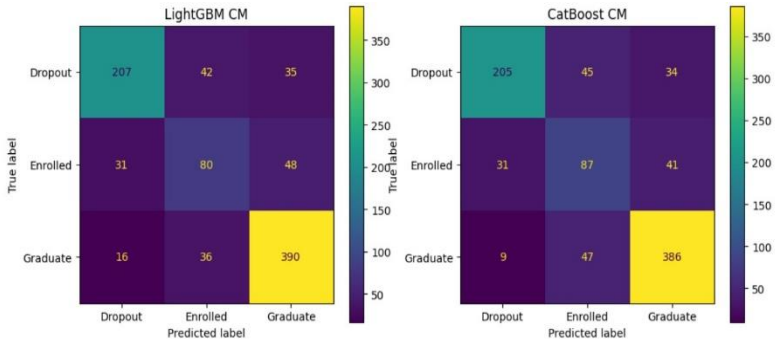
۳-۲- پیش‌پردازش و تقسیم داده‌ها

پس از آماده‌سازی داده‌ها، مراحل پیش‌پردازش و تقسیم داده‌ها به شرح زیر انجام شد:

۱. کدگذاری متغیرهای دسته‌ای: ویژگی‌های متنی مانند جنسیت، رشته تحصیلی و وضعیت شغلی با استفاده از *One-Hot Encoding* یا *Label Encoding* به مقادیر عددی تبدیل شدند.

۲. نرمال‌سازی ویژگی‌های عددی: ویژگی‌های عددی مانند سن، معدل و درصد حضور در کلاس با *StandardScaler* استاندارد شدند تا الگوریتم‌هایی مانند *SVM* عملکرد بهتری داشته باشند.

۳. متعادل‌سازی کلاس‌ها: با توجه به نامتوازن بودن دیتاست، از روش *SMOTE* برای افزایش نمونه‌های کلاس اقلیت[1] استفاده شد تا نسبت کلاس‌ها متعادل گردد.



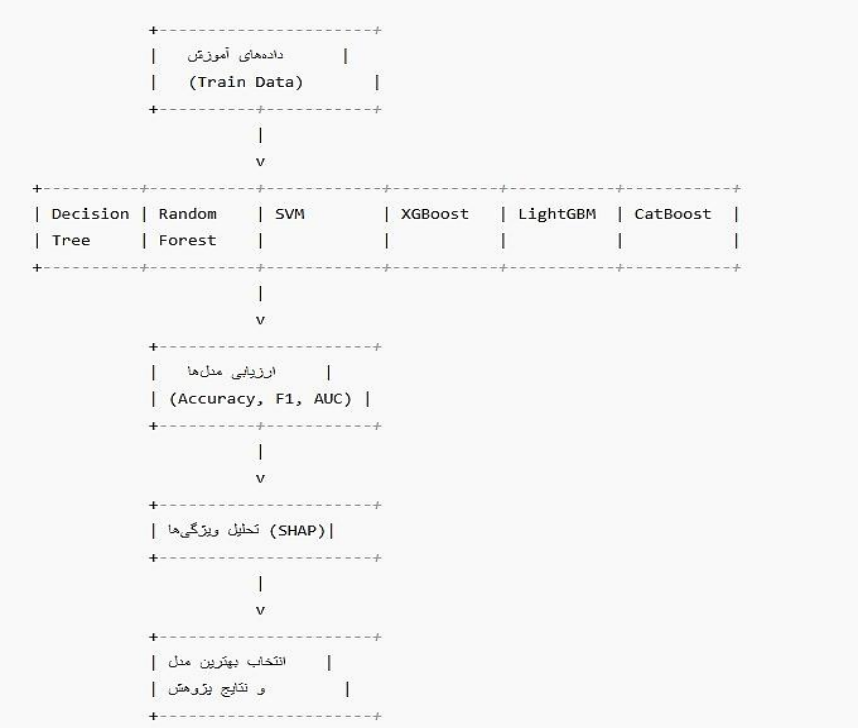
شکل ۲- مقایسه توزیع کلاس‌ها قبل و بعد از اعمال روش SMOTE برای متعادل‌سازی دیتاست [7]، [8].
Figure 2- Comparison of class distribution before and after applying the SMOTE method to balance the data [7], [8].

[1] Dropout

۴-۲- تقسیم داده‌ها به Train و Test

۱. داده‌ها به دو بخش *Train* ۸۰% و *Test* ۲۰% تقسیم شدند.

۲. برای حفظ نسبت کلاس‌ها از *Stratified Split* استفاده شد تا توزیع *Dropout* و *Success* در هر بخش حفظ شود.



شکل ۳- بلوک دیاگرام مراحل آموزش شش الگوریتم یادگیری ماشین، ارزیابی عملکرد آن‌ها با معیارهای Accuracy، F1-scoreو AUC، تحلیل اهمیت ویژگی‌ها با SHAP و انتخاب بهترین مدل برای پیش‌بینی وضعیت تحصیلی دانشجویان [9]، [10].

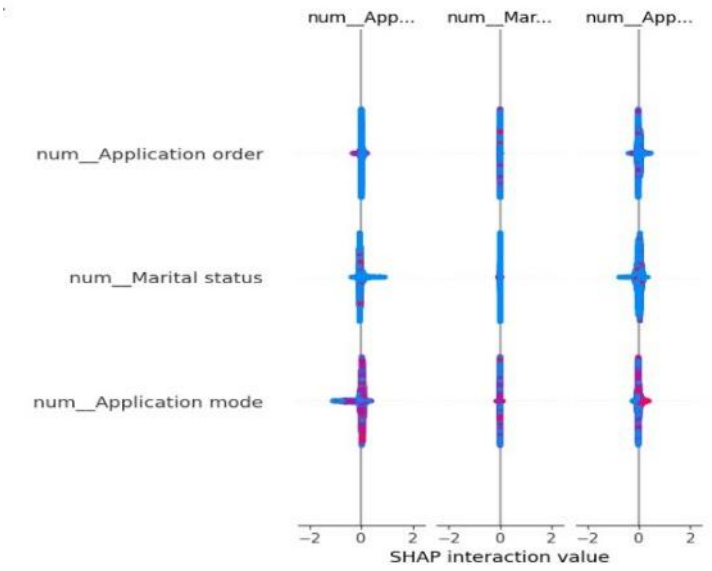
Figure 3- Block diagram of the steps of training six machine learning algorithms, evaluating their performance with Accuracy, F1-score, and AUC metrics, analyzing the importance of features with SHAP, and selecting the best model to predict students' academic status [9], [10].

در این پژوهش شش الگوریتم یادگیری ماشین شامل سه مدل پایه *Decision Tree*، *Random Forest*، *SVM* و سه مدل تقویتی *XGBoost*، *LightGBM*، *CatBoost* برای پیش‌بینی وضعیت تحصیلی دانشجویان پیاده‌سازی شدند. به‌منظور بهبود عملکرد، ابر پارامترهای هر مدل با استفاده از *Optuna*تنظیم شدند که منجر به افزایش دقت و امتیاز *F1*، به‌ویژه در مدل‌های مبتنی بر *Boosting* گردید [5].

جدول ۲- مقایسه عملکرد مدل‌ها قبل و بعد از Tuning با استفاده از Optuna.

Table 2- Comparison of model performance before and after tuning using Optuna.

الگوریتم	دقت قبل	دقت بعد	امتیاز F1قبل	امتیاز F1 بعد
درخت تصمیم	0.72	0.75	0.70	0.74
جنگل تصادفی	0.78	0.82	0.77	0.81
ماشین بردار پشتیبان	0.76	0.80	0.75	0.79
XGBoost	0.81	0.85	0.80	0.84
LightGBM	0.83	0.87	0.82	0.86
CatBoost	0.84	0.88	0.83	0.87



شکل ۴- نمودار SHAP در مدل LightGBM که اهمیت و اثرگذاری ویژگی‌های اصلی بر پیش‌بینی وضعیت تحصیلی دانشجویان را نشان می‌دهد.

Figure 4- SHAP diagram in the LightGBM model showing the importance and impact of the main features on predicting students' academic status.

۳- نتیجه و جمع‌بندی

این پژوهش نشان داد که الگوریتم‌های *Boosting*، به‌ویژه *CatBoost* و *LightGBM*، بالاترین سطوح دقت و *F1-score* را در پیش‌بینی وضعیت تحصیلی دانشجویان ارایه می‌کنند. این مدل‌ها با استفاده کارآمد از ویژگی‌های عددی و دسته‌ای، توانستند الگوهای پیچیده‌تری از رفتار تحصیلی دانشجویان را شناسایی کنند. تحلیل ویژگی‌ها نشان داد که عواملی مانند میزان حضور در کلاس، معدل ترم قبل، وضعیت اقتصادی-اجتماعی خانواده و سابقه عملکرد تحصیلی از مهم‌ترین شاخص‌های موثر در افزایش احتمال *Dropout* محسوب می‌شوند. در مقابل، الگوریتم‌های پایه مانند *Decision Tree* و *Random Forest* عملکرد پایین‌تری داشتند و نسبت به نامتوازن بودن داده‌ها حساسیت بیشتری نشان دادند.

این نتایج تاکید می‌کند که انتخاب الگوریتم مناسب، اعمال روش‌های موثر پیش‌پردازش، استفاده از تکنیک‌های متعادل‌سازی مانند *SMOTE* و تحلیل اهمیت ویژگی‌ها نقش تعیین‌کننده‌ای در بهبود عملکرد مدل‌ها دارد. یافته‌های این پژوهش می‌تواند به مدیران و تصمیم‌گیرندگان حوزه آموزش عالی کمک کند تا با تکیه بر مدل‌های دقیق‌تر، دانشجویان در معرض خطر را زودتر شناسایی کرده و برنامه‌های حمایتی هدفمند ارایه دهند.

منابع

[1] Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover artificial intelligence*, *4*(1), 2. https://doi.org/10.1007/s44163-023-00079-z

[2] Mduma, N., Kalegele, K., & Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction. *Data science journa*, *18*(14), 1–10. https://doi.org/10.5334/dsj-2019-014

[3] Sulak, S. A., & Koklu, N. (2024). Predicting student dropout using machine learning algorithms. *Intelligent methods in engineering sciences*, *3*(3), 91–98. https://doi.org/10.58190/imiens.2024.103

[4] Goren, O., Cohen, L., & Rubinstein, A. (2024). Early prediction of student dropout in higher education using machine learning models. *Proceedings of the 17th international conference on educational data mining* (pp. 349–359). International (CC BY-NC-ND 4.0) License. https://doi.org/10.5281/zenodo.12729834

[5] Cho, C. H., Yu, Y. W., & Kim, H. G. (2023). A study on dropout prediction for university students using machine learning. *Applied sciences*, *13*(21), 1204. https://doi.org/10.3390/app132112004

[6] Quimiz-Moreira, M., Delgadillo, R., Parraga-Alava, J., Maculan, N., & Mauricio, D. (2025). Factors, prediction, explainability, and simulating university dropout through machine learning: A systematic Review, 2012–2024. *Computation*, *13*(8), 198. https://doi.org/10.3390/computation13080198

[7] Mduma, N. (2023). Data balancing techniques for predicting student dropout using machine learning. *Data*, *8*(3), 49. https://doi.org/10.3390/data8030049

[8] Wijayaningrum, V. N., Kirana, A. P., & Putri, I. K. (2024). Student academic performance prediction framework with feature selection and imbalanced data handling. *Jurnal ilmiah kursor*, *12*(3), 123–134. http://kursorjournal.org/index.php/kursor/article/view/356/159

[9] Joshi, A., Saggar, P., Jain, R., Sharma, M., Gupta, D., & Khanna, A. (2021). CatBoost — an ensemble machine learning model for prediction and classification of student academic performance. *Advances in data science and adaptive analysis*, *13*(03n04), 2141002. https://doi.org/10.1142/S2424922X21410023

[10] Syed Mustapha, S. M. F. D. (2023). Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods. *Applied system innovation*, *6*(5), 86. https://doi.org/10.3390/asi6050086