Business Intelligence Management Studies
Vol. 12, No.47, Spring 2024; P 315-366
ims.atu.ac.ir
DOI: 10.22054/ims.2023.75523.2375

ISSN: 2821-0964 eISSN: 2821-0816 Received: 02 October 2023 Revised: 21 November 2023 Accepted: 05 December 2023 Original Research

Predicting Students' Performance Using Machine Learning Algorithms and Educational Data Mining (A Case Study of Shahed University)

Mozhdeh Salari Ph.D. Student of Information Technology Management Group, Department of Economics and Management, Science and Research Branch, Islamic Azad University, Tehran, Iran

Reza Radfar * Full Professor of Information Technology Management Group, Department of Economics and Management, Science and Research Branch, Islamic Azad University, Tehran, Iran

Mehdi Faqhihi Assistant Professor of Information Technology Management Group, Department of Economics and Management, Science and Research Branch, Islamic Azad University, Tehran, Iran

Abstract

The purpose of this research is to investigate the effective factors in predicting the academic performance of undergraduate students in the classification of four classes. To achieve this goal, the study follows the CRISP data mining method. The data set was extracted from the NAD educational system for the bachelor's degree in Shahed University for the entry of the years 2011 to 2021. 1468 records were used in data mining. First, the effective features on students' academic performance were extracted. Modeling was done using Rapidminer9.9 tool. To improve classification performance and satisfactory prediction accuracy, we use a combination of principal component analysis combined with machine learning algorithms and feature selection techniques and optimization algorithms. The performance of the prediction models is verified using 10-fold cross-validation. The results showed that the decision tree algorithm is the best algorithm in predicting students' performance with an accuracy of

* Corresponding Author: r.radfar@srbiau.ac.ir

How to Cite: Salari, M., Radfar, R. & Faghihi. M. (2024). Predicting Students' Performance Using Machine Learning Algorithms and Educational Data Mining (A Case Study of Shahed University), *Journal of Business Intelligence Management Studies*, 12(47), 315-366.

84.71%. This algorithm correctly predicted the graduation of 77.88% of excellent students, 85.26% of good students, 84.69% of medium students, and 85.96% of weak students based on the final GPA.

1. Introduction

The main problem in this research is to identify the factors that are effective in predicting the academic performance of undergraduate students in Shahed University. Choosing the best machine learning algorithm in predicting academic performance among different modeling methods based on validation and evaluation of models is another issue in the present research. The purpose of this research is to investigate the effective factors in predicting the academic performance of undergraduate students in Shahed University using educational data mining based on classification models.

Research questions

The main question in this research is what factors affect the prediction of undergraduate students' performance and improving their performance?

Sub questions

1- Which modeling algorithms have better results in predicting student performance?

2- What methods have been used to predict students' performance?

3- What is the validity of the developed model for Shahed University students?

2- Research background

1-2- Theoretical foundations

- **Educational data mining**

The processing of educational data improves the prediction of student behavior and new approaches to educational policies (Capuano & Toti, 2019) (Viberg et al., 2018)

- **Academic performance**

Academic performance of students means the extent to which they achieve educational goals (Banik & Kumar, 2019).

2-2- review of past studies

The highlighted cells in Table 1, based on past research, show the classification algorithms that have the most accuracy and effectiveness in predicting students' performance in the relevant research. The decision tree algorithm has been used the most in previous researches. The NB algorithm has been the most used in research after the decision tree. RF and ANN algorithms are next in use. After that, SVM and KNN algorithms have been used in research

Table 1. The results of research literature based on the use of classification algorithms

Accuracy	LR	Line RL	ANN	SVM	KNN	NB	RF	DT	Data mining algorithm
			*				*		(Batool et al., 2023)
			*	*	*	*	*	*	(Marjan et al., 2023)
	*			*	*		*		(Abdelmagid & Qahmash, 2023)
	*		*		*		*	*	(Manoharan et al., 2023)
99.34%						*	*	*	(Alghamdi & Rahman, 2023)
		*	*	*	*		*		(Alboaneen et al., 2022)
70-75%	*			*	*	*		*	(Yağcı, 2022)
83.44%		*		*				*	(Dabhade et al., 2021)
95%								*	(Najafi & etal,2021)
				*	*			*	(Soltani & etal,2021)
50-81%	*			*	*		*		(Cruz-Jesus et al., 2020)
				*		*	*	*	(Sokkhey & Okazaki, 2020)
							*	*	(Rebai et al., 2020)
						*	*	*	(Jayaprakash et al., 2020)
			*				*	*	(Zulfiker et al., 2020)
			*						(Musso et al., 2020)
85%			*						(Waheed et al., 2020)
			*	*	*	*		*	(Salal & Abdullaev, 2019)
			*		*	*		*	(Turabieh, 2019)
			*	*				*	(Xu et al., 2019)
						*		*	(ghodoosi & etal,2019)
95.84%						*			(fadavi & etal,2019)
91.5%				*	*	*		*	(Ajibade et al., 2019)

Accuracy	LR	Line RL	ANN	SVM	KNN	NB	RF	DT	Data mining algorithm
85%			*						(Ahmad & Shahzadi, 2018)
						*		*	(Hasani & Bazrafshan, 2018)
		*				*	*	*	(Hussain et al., 2018)
	*				*	*	*	*	(Umer et al., 2017)
						*		*	(Khasanah, 2017)
								*	(Asif et al., 2017)
92.34%	*		*				*		(Hoffait & Schyns, 2017)
			*					*	(khosravi &etal,2017)
86%			*			*		*	(Mueen et al., 2016)
					*	*		*	(Amrieh et al., 2015)
92.34%						*		*	(Yehuala, 2015)
	*			*				*	(zahedi & etal,2015)
								*	(Punlumjeak & Rachburee, 2015)
71%							*	*	(Osmanbegović et al., 2014)
								*	(Shamloo & et al.,2014)
								*	(Asadi & et al.,2013)
60-75%					*	*		*	(Kabakchieva, 2013)
96%			*			*	*	*	(Oskouei & Askari, 2014)
						*		*	(Nghe et al., 2007)
94.17%			*	*	*	*	*	*	present research

3- Method

This study follows the popular training data mining method CRISP. The data collection of Nad educational system for bachelor's degree in non-medical fields of Shahed University has been extracted from 2011 to 2021. We used the Label Encoder technique to encode the features. In this research, C4.5 and ID3 decision tree classification algorithms, random forest, Naïve Bayes, k-nearest neighbor and artificial neural network and gradient enhanced tree were used to analyze and classify students and predict the final GPA. Modeling was done using RapidMiner 9.9. To improve the classification performance and solve the misclassification problem, we use a combination of principal component analysis and feature selection techniques and optimization algorithms. In this research, prediction accuracy was evaluated using

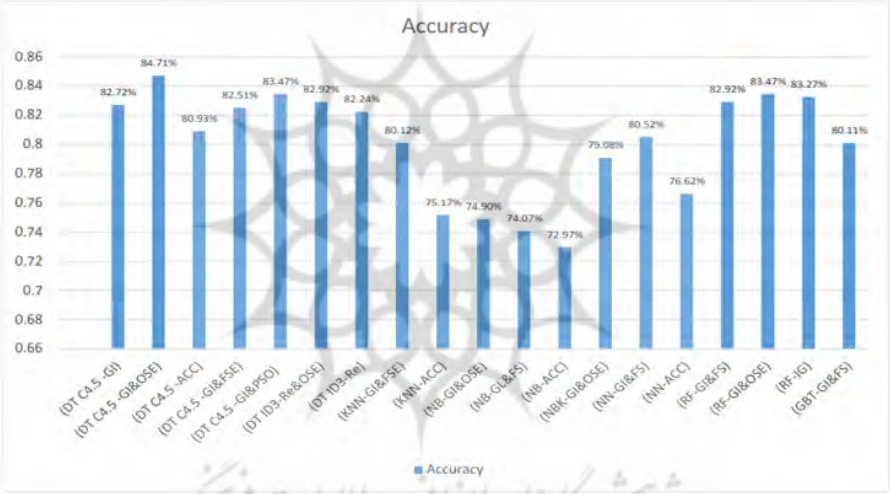
10-fold cross-validation method for all algorithms. Also, different algorithms were compared using the analytical descriptive method and based on evaluation criteria, and the best prediction model was introduced in this research.

4-Data analysis

4-1 Introduction

The best model is the model that has the best values for the selected performance measurement criteria(Lever et al., 2016). Figure 1 is a graph that compares the accuracy of the algorithms used in this research.

Figure 1. Comparative chart of the accuracy of the algorithms



According to Table 2, the DTC4.5 algorithm is able to predict the class of 1235 objects out of 1458, which gives it an accuracy value of 84.71%.

Table 2. Confusion matrix of DT C4.5-GI&OSE research model

	Students with excellent performance	Students with good performance	Students with average performance	Students with poor performance	precision
Prediction 1	81	22	0	0	78.64%
Prediction 2	22	295	49	9	78.67%
Prediction 3	1	27	498	50	86.46%

	Students with excellent performance	Students with good performance	Students with average performance	Students with poor performance	precision
Prediction 4	0	2	41	361	89.36%
Recall	77.88%	85.26%	84.69%	85.95%	

4-2 important features

The prioritization of predictive variables based on their weight is as follows:
Diploma GPA: 0.262
Semester 1 GPA: 0.201
Semester 2 GPA: 0.197
Number of honors semesters: 0.122
Conditional number: 0.114
Year of entry: 0.104

4-3 The results of the implementation of the student performance prediction model

The results of the prediction model are shown in Table 3:

Table 3. The results of the DT C4.5-GI&OSE model implementation

Row No.	شماره دانشجویی	معدل کل طبقه بند...	prediction(مع...	confidence(1)	confidence(2)	confidence(3)	confidence(4)
1	932106010	3	3	0	0	1	0
2	982174021	3	4	0	0	0	1
3	972106010	1	1	1	0	0	0
4	902161006	3	3	0	0	1	0
5	962155025	2	3	0	0.455	0.545	0
6	992151020	3	3	0	0.071	0.857	0.071
7	992151006	2	2	0	1	0	0
8	902161013	3	3	0	0.059	0.941	0

5- Discussion

In the main method of research, namely DT C4.5-GI&OSE, in the classification mode of four classes, it is observed that the average of the diploma has the greatest effect on the process of predicting student performance. In response to the sub-question of a research, the best algorithm in the four-class mode is Decision Tree C4.5-GI&OSE with a prediction accuracy of 84.71. This model showed 84.17% accuracy,

83.42% sensitivity and 0.780 kappa. DT C4.5-GI&OSE technique correctly predicted the graduation of 77.88% of excellent students, 85.26% of good students, 84.69% of average students, and 85.96% of poor students.

6-Conclusion

The obtained results show that there is a relationship between students' social and academic characteristics and their academic performance. DT C4.5-GI&OSE algorithm was the best algorithm for predicting the final GPA scores of students at the end of studies with a prediction accuracy of 84.71%. In this model, the average grade point average of the diploma has the greatest effect on the prediction process. Using machine learning models as a decision support tool improves the academic level of students and reduces the number of potential unsuccessful and dropout students. This study was carried out at the undergraduate level, which can be used in future research for the master's and doctoral level.

Keywords: student performance prediction, data mining, machine learning, modeling, improving the quality of education

مقاله پژوهشی تاریخ ارسال: ۱۴۰۲/۰۷/۱۰ تاریخ بازنگری: ۱۴۰۲/۰۸/۳۰ تاریخ پذیرش: ۱۴۰۲/۰۹/۱۴ ISSN: 2821-0964 eISSN: 2821-0816

مطالعات مدیریت کسب و کار هوشمند
سال دوازدهم، شماره ۴۷، بهار ۱۴۰۳، ص ۳۱۵ تا ۳۶۶
ims.atu.ac.ir
DOI: 10.22054/ims.2023.75523.2375

پیش‌بینی عملکرد دانشجویان با استفاده از الگوریتم‌های یادگیری ماشین و داده‌کاوی آموزشی (مطالعه موردی دانشگاه شاهد)

مژده سالاری دانشجوی گروه مدیریت فناوری اطلاعات، دانشکده اقتصاد و مدیریت، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران

رضا رادفر * استاد گروه مدیریت فناوری اطلاعات، دانشکده اقتصاد و مدیریت، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران

مهدی فقیهی دانشیار گروه مدیریت فناوری اطلاعات، دانشکده اقتصاد و مدیریت، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران

چکیده

هدف این تحقیق بررسی عوامل مؤثر در پیش‌بینی عملکرد تحصیلی دانشجویان مقطع کارشناسی در طبقه‌بندی چهار کلاسه می‌باشد. برای دستیابی به این هدف، مطالعه از روش داده‌کاوی کریسپ پیروی می‌کند. مجموعه داده‌ها از سیستم آموزشی ناد برای مقطع کارشناسی در دانشگاه شاهد برای ورودی سال‌های ۱۳۹۰ تا ۱۴۰۰ استخراج شده است. تعداد ۱۴۶۸ رکورد در داده‌کاوی استفاده شده است. ابتدا شاخص‌های مؤثر بر عملکرد تحصیلی دانشجویان استخراج شد. مدل‌سازی با استفاده از ابزار رپیدماینر ۹,۹ انجام شد. برای بهبود عملکرد طبقه‌بندی و دقت پیش‌بینی رضایت‌بخش، از ترکیبی از تجزیه‌وتحلیل مؤلفه اصلی [i] همراه با الگوریتم‌های یادگیری ماشین و تکنیک‌های انتخاب ویژگی و الگوریتم‌های بهینه‌سازی استفاده می‌کنیم. عملکرد مدل‌های پیش‌بینی با استفاده از اعتبارسنجی متقاطع ۱۰ برابری تأیید شده است. نتایج نشان داد که الگوریتم درخت تصمیم بهترین الگوریتم در پیش‌بینی عملکرد دانشجویان با دقت ۸۴,۷۱ درصد است. این الگوریتم به‌درستی فارغ‌التحصیلی ۷۷,۸۸ درصد از دانشجویان عالی و ۸۵,۲۶ درصد از دانشجویان خوب و ۸۴,۶۹ درصد از دانشجویان متوسط و ۸۵,۹۶ درصد از دانشجویان ضعیف را بر اساس معدل نهایی پیش‌بینی کرد. متغیر معدل دیپلم بیشترین تأثیر را در پیش‌بینی عملکرد دانشجویان دارد.

کلیدواژه‌ها: پیش‌بینی عملکرد دانشجویان، داده‌کاوی، یادگیری ماشینی، مدل‌سازی، بهبود کیفیت آموزش.

* نویسنده مسئول : r.radfar@srbiau.ac.ir

مقدمه

مؤسسات آموزش عالی اولویت بالایی برای بهبود عملکرد تحصیلی دانشجویان قائل هستند و پیش‌بینی موفقیت دانشجو یک چالش بزرگ برای مدیران آموزش عالی است (Arcinas et al., 2021). مطالعات متعددی اهمیت شناسایی عوامل کلیدی منجر به عملکرد را نادیده گرفته‌اند. چنین شناسایی برای توانمندسازی رهبران برای شناخت نقاط قوت و ضعف برنامه‌های آموزشی و در نتیجه اجرای مداخلات اصلاحی برای بهبود پیشرفت دانشجویان ضروری است (Alshanqiti & Namoun, 2020). با توجه به تحقیقات اندکی که در زمینه کشف عوامل مؤثر بر عملکرد دانشجویان در ایران وجود دارد، انجام چنین تحقیقاتی برای پر کردن خلأ پژوهشی نسبت به مطالعات متعدد در سایرکشورها ضروری است. یکی از مسائل اصلی پیش‌بینی دستاوردهای آینده دانشجویان قبل از شرکت در امتحانات نهایی است، بنابراین ما می‌توانیم به طور فعال به دانشجویان در دستیابی به عملکرد بهتر و جلوگیری از ترک تحصیل کمک کنیم(Batool et al., 2023) . یکی از موضوعات قابل‌توجه در عملکرد تحصیلی، پیش‌بینی صحیح عملکرد تحصیلی دانشجویان و اقدام به‌موقع نسبت به دانشجویان در معرض خطر افت تحصیلی است (Asif et al., 2017). یکی از مهم‌ترین شاخص‌های پیشرفت هر جامعه، میزان توسعه و کاربرد فناوری اطلاعات و ارتباطات در آموزش است (Rostami et al., 2015). استفاده از داده‌کاوی در این رابطه برای استخراج عوامل مؤثر بر عملکرد تحصیلی دانشجویان مفید است (Alboaneen et al., 2022). داده‌کاوی آموزشی می‌تواند به سؤالاتی مانند پیش‌بینی نتایج دانشجویان پاسخ دهد (El Aissaoui et al., 2019). ازآنجایی‌که در سیستم‌های آموزشی فعلی به پیش‌بینی عملکرد دانشجویان اهمیت داده نمی‌شود، این سیستم‌ها از ناکارآمدی رنج می‌برند(Cuevas et al., 2018). شناسایی دانشجویان در معرض خطر، افزایش نرخ فارغ‌التحصیلی، نظارت مؤثر بر عملکرد سازمانی، مدیریت منابع دانشگاه و بهینه‌سازی تجدید برنامه درسی(Ampadu, 2023)، یافتن دلایل احتمالی ادامه یا انصراف از تحصیل دانشجو؛ ساخت مدل رفتاری دانشجو و پیش‌بینی بر اساس آن، فراهم کردن زمینه برای کار

با حجم زیادی از داده‌ها و کسب دانش عمیق در مورد روش‌ها، فرآیندها و الگوریتم‌های یادگیری(Garcia & Skrita, 2019) ، ارزیابی مدیران در مورد اقدامات آموزشی اتخاذشده، نظارت و ارزیابی فرآیند یاددهی-یادگیری(Debang & Hassan, 2023) . برخی از کاربردهای داده‌کاوی آموزشی هستند. طبقه‌بندی مؤثرترین تکنیک داده‌کاوی است که برای پیش‌بینی استفاده می‌شود که می‌تواند به دانشجویان و اساتید در بهبود کیفیت آموزش با شناسایی افراد موفق و ضعیف در ابتدای ترم/سال کمک کند(Saa, 2016). پیش‌بینی اولیه تحصیلی در داده‌کاوی آموزشی، راهبرد جدیدی در پیش‌بینی عملکرد تحصیلی است (Waheed et al., 2020). در عصر داده‌های انبوه که در آن میزان اطلاعات به طور تصاعدی در حال افزایش است، اهمیت داده‌کاوی هرگز بیشتر از این نبوده است (Ampadu, 2023) . در آینده‌ای نزدیک، به لطف داده‌های انبوه و اینترنت اشیا، اطلاعات هر دانشجو از بدو تولد تا این لحظه می‌تواند در دسترس باشد. این اطلاعات شامل داده‌های سنتی عملکرد تحصیلی دانشجو از تمامی محیط‌های آموزشی قبلی و اطلاعات مربوط به وضعیت شخصی هر دانشجو مانند اطلاعات پزشکی، خانوادگی، اقتصادی، مذهبی، ارتباطی، عاطفی، روان‌شناختی و غیره است(Al-Emran et al., 2020; Romero & Ventura, 2020). این داده‌ها را می‌توان از منابع متعدد جمع‌آوری کرد و سپس برای بهبود و شخصی‌سازی فرآیند یادگیری در هر لحظه خاص از زندگی دانشجویان ترکیب کرد(Ding et al., 2019). با انجام پیش‌بینی در صورت مشاهده مشکل در عملکرد دانشجو ، کمک به‌موقع به او ارائه می‌شود (Kumar & Salal, 2019). ارزیابی وضعیت کنونی برنامه آموزشی، پیش‌بینی نتایج مطلوب و ترسیم نقشه راه برای تغییرات برنامه‌ریزی‌شده ضروری است (Batirovna, 2023). حجم زیاد اطلاعات و ساختار نداشتن آن‌ها برای اساتید و دانشجویان چندان مؤثر نیست. با داده‌کاوی آموزشی می‌توان با کشف الگوها و قوانین موجود با این مشکلات مقابله کرد. استفاده از این الگوهای کشف‌شده می‌تواند راهگشای مشکلات آموزشی دانشجویان باشد و در بهبود نظام آموزشی و نحوه ارائه دروس تأثیرگذار باشد. مسئله اصلی در این تحقیق شناسایی

عواملی است که بر پیش‌بینی عملکرد تحصیلی دانشجویان مقطع کارشناسی در دانشگاه شاهد مؤثر هستند. انتخاب بهترین الگوریتم یادگیری ماشین در پیش‌بینی عملکرد تحصیلی از میان روش‌های مختلف مدل‌سازی بر اساس اعتبارسنجی و ارزیابی مدل‌ها مسئله دیگر در تحقیق حاضر است. هدف این پژوهش بررسی عوامل مؤثر در پیش‌بینی عملکرد تحصیلی دانشجویان مقطع کارشناسی در دانشگاه شاهد با استفاده از داده‌کاوی آموزشی بر اساس مدل‌های طبقه‌بندی می‌باشد. همچنین در این تحقیق می‌توان بهترین مدل را برای پیش‌بینی عملکرد تحصیلی دانشجویان برای بهبود عملکرد و موفقیت تحصیلی آن‌ها انتخاب کرد. داده‌کاوی صورت گرفته بر روی یک سری عواملی که داده‌های در دسترس در سامانه آموزشی ناد دارند، صورت پذیرفته است.

پیشینه پژوهش

مبانی نظری پژوهش

- **داده‌کاوی آموزشی**

تکنیک‌های خاص داده‌کاوی آموزشی می‌توانند بهترین ابزارها را برای حل برخی مشکلات یادگیری فراهم کنند(Aldowah et al., 2019). حوزه داده‌کاوی آموزشی از تکنیک‌های داده‌کاوی برای بررسی داده‌های آموزشی استفاده می‌کند (Kaur & Dahiya, 2023). در اکثر دانشگاه‌ها پایگاه‌های اطلاعاتی متعددی از ویژگی‌های جمعیتی و سوابق تحصیلی دانشجویان وجود دارد. در این حجم از داده‌ها، الگوها و روابط قابل‌توجهی پنهان می‌ماند که می‌توان آن‌ها را با استفاده از دانش داده‌کاوی استخراج و تحلیل کرد (Rostami et al., 2015). در سال‌های اخیر، تحقیقات آموزشی به ابزاری مؤثر برای شناسایی الگوهای پنهان در داده‌های آموزشی، پیش‌بینی پیشرفت تحصیلی و بهبود محیط یادگیری/تدریس تبدیل شده است(Waheed et al., 2020). اکنون دانشگاه‌ها باید ظرفیت خود را برای استفاده از داده‌ها برای پیش‌بینی موفقیت تحصیلی و تضمین پیشرفت دانشجویان بهبود بخشند(Bernacki et al., 2020). پردازش داده‌های آموزشی، پیش‌بینی

رفتار دانشجویان و رویکردهای جدید به سیاست‌های آموزشی را بهبود می‌بخشد (Viberg et al., 2018)(Capuano & Toti, 2019) .

- **ویژگی‌ها در داده‌کاوی آموزشی**

ویژگی‌های به کار گرفته‌شده در تحقیقات قبلی بر اساس جدول ۱ شامل موارد زیر می‌باشد:

- اطلاعات جمعیت شناختی قبل از ورود به دانشگاه شامل سن، جنسیت، محل زندگی، سطح تحصیلات والدین، صلاحیت مادر، وضعیت خانوادگی دانشجو، درآمد سالانه خانواده، شغل، مذهب، نظم فردی، تسلط به زبان انگلیسی و عادات دیگر دانشجویان
- وضعیت تحصیلی قبلی شامل نمره دانشجو در دبیرستان یا محیط آموزشی، نمرات کل از تحصیلات قبلی، رشته تحصیلی، نتایج امتحان ورودی
- وضعیت تحصیلی کنونی شامل نمرات دوره، معدل در نیمسال اول، معدل نیمسال دوم، معدل ترم جاری و معدل کل دانشگاه، نتایج امتحانات قبلی، وضعیت دسترسی به اینترنت، داشتن کامپیوتر، حضور در کلاس، غیبت، تعداد دوره‌های گذرانده شده، تعداد دانشجویان در یک کلاس، تعداد دروس ارائه‌شده در یک‌ترم و رشته اصلی، نمرات ارزیابی، زمان ارائه راه‌حل، گزارش مشارکت سخنرانی ویدئویی، زمان صرف شده در هفته، اندازه مدرسه، اندازه کلاس، رقابت، فشار والدین
- رفتارهای استفاده از اینترنت شامل زمان آنلاین بودن، حجم ترافیک اینترنت و ردپای دیجیتالی دانشجویان در اینترنت (مرور، زمان درس، درصد مشارکت)
- ویژگی‌های شخصیتی و روان‌شناختی و رفتاری دانشجویان شامل عوامل روان رنجوری، برونگرایی، گشودگی به تجربه، موافق بودن و وظیفه‌شناسی و تنوع درون فردی و تفاوت‌های فردی بین افراد- شدت مطالعه، ارزشیابی موقت، مشارکت دانشجویان، انگیزه، عادات، مسائل اجتماعی و مالی، عدم پیشرفت و جابجایی شغلی دانشجویان- راهبردهای یادگیری، درک حمایت اجتماعی، انگیزه، جمعیت‌شناسی اجتماعی، وضعیت سلامت.

- **معدل نهایی معیار سنجش عملکرد تحصیلی**

در پژوهش حاضر، معیار سنجش عملکرد تحصیلی دانشجویان، معدل نهایی آن‌ها در مقطع کارشناسی است. در پژوهش یقینی و همکاران (۱۳۸۷) نیز وضعیت تحصیلی آینده دانشجویان از طریق معدل نهایی زمان فارغ‌التحصیلی پیش‌بینی شده است (یقینی و همکاران، ۱۳۸۷). سایر محققان همچون موسو و همکاران (۲۰۲۰) و پانلومجیک و راچبور (۲۰۱۵) نیز وضعیت تحصیلی آینده دانشجویان را بر اساس معدل نهایی در زمان فارغ‌التحصیلی پیش‌بینی کرده‌اند(Musso et al., 2020) (Punlumjeak & Rachburee, 2015).

جدول ۱. ویژگی‌های مؤثر بر پیش‌بینی عملکرد تحصیلی دانشجویان در تحقیقات

انواع ویژگی‌ها	مطالعات مربوطه
نمرات امتحانات میان‌ترم	(Yağcı, 2022)
مشخصه‌های جمعیتی	(Cruz-Jesus et al., 2020)
سوابق تحصیلی	(Fernandes et al., 2017)
مدرسه و عوامل محیطی	(Rebai et al., 2020)
عوامل روانی اجتماعی	(Musso et al., 2020)
سوابق تحصیلی	(Kabakchieva, 2013)(Harwati, 2014),
مشارکت تحصیلی	(Salal & Abdullaev, 2019)
عملکرد تحصیلی اولیه	(Khasanah, 2017)
سوابق تحصیلی قبلی	(Punlumjeak & Rachburee, 2015)
خانواده و پیشینه	(Zulfiker et al., 2020)
عوامل تحصیلی	(Nghe et al., 2007)
سوابق خانوادگی و تحصیلی	(Oskouei & Askari, 2014)
ردپای دیجیتال	(Waheed et al., 2020)
نتایج آزمون ورودی	(Yehuala, 2015)
ویژگی‌های رفتاری	(Amrieh et al., 2015)
داده‌های ارزیابی	(Umer et al., 2017)
الگوهای استفاده از اینترنت	(Xu et al., 2019)
اطلاعات پیش‌دانشگاهی	(Hasani & Bazrafshan, 2018)

انواع ویژگی‌ها	مطالعات مربوطه
خانواده و عوامل مالی	(Pandey & Pal, 2011)
عوامل روانشناسی	(Galla et al., 2014)
ویژگی‌های شخصیتی	(Chamorro-Premuzic & Furnham, 2009)
عوامل ترک تحصیل	(Hellas et al., 2018)

- **عملکرد تحصیلی**

عملکرد به‌عنوان «معیار شایستگی دانشجو برای دوره‌های آینده» تعریف می‌شود (Lei & Li, 2015). عملکرد تحصیلی دانشجویان به معنای میزان دستیابی آن‌ها به اهداف آموزشی است (Banik & Kumar, 2019). عملکرد تحصیلی کلیه فعالیت‌هایی است که فرد برای کسب دانش و گذراندن مدارج علمی انجام می‌دهد (Seif, 2016). افراد با عملکرد تحصیلی خوب اعتمادبه‌نفس به دست می‌آورند و مورد تأیید همه قرار می‌گیرند و افراد با عملکرد تحصیلی نامطلوب به دلیل عدم عزت‌نفس و احساس بی‌لیاقتی از ادامه تحصیل باز می‌مانند (Phan & Ngu, 2014).

- **افت تحصیلی**

یکی از مشکلات نظام آموزشی افت قابل‌توجه عملکرد تحصیلی دانشجویان است که آسیب‌های اقتصادی به جامعه و آسیب‌های روحی و روانی به دانشجویان وارد می‌کند که عموماً غیرقابل‌جبران است. یک نظام آموزشی کارآمد کمترین میزان ترک تحصیل و بیشترین کارایی را دارد (Romero & Ventura, 2007).افت تحصیلی به این معنی است که دانشجو نمره ضعیفی می‌گیرد و استانداردهای تحصیلی را برآورده نمی‌کند (Widyastuti et al., 2017).

مرور مطالعات گذشته

بررسی سوابق پژوهشی در جدول ۲ نشان می‌دهد که الگوریتم‌های داده‌کاوی می‌توانند با دقت بالا عملکرد تحصیلی دانشجویان را پیش‌بینی کرده و موجبات پیشرفت تحصیلی

دانشجویان را فراهم کنند. دقت پیش‌بینی مدل‌ها به سطوح بسیار بالایی می‌رسد و برای یک پیش‌بینی خوب، نمرات اولیه لازم است (Zimmermann et al., 2015). پیش‌بینی دقیق عملکرد تحصیلی دانشجو مستلزم درک عمیق عوامل و ویژگی‌های مؤثر بر نتایج و پیشرفت دانشجو است (Alshanqiti & Namoun, 2020).

جدول ۲. نتایج ادبیات تحقیق بر اساس استفاده از الگوریتم‌های طبقه‌بندی

Accuracy	LR	Line RL	ANN	SVM	KNN	NB	RF	DT	Data mining algorithm
			*				*		(Batool et al., 2023)
			*	*	*	*	*	*	(Marjan et al., 2023)
	*			*	*		*		(Abdelmagid & Qahmash, 2023)
	*		*		*		*	*	(Manoharan et al., 2023)
99.34%						*	*	*	(Alghamdi & Rahman, 2023)
		*	*	*	*		*		(Alboaneen et al., 2022)
70-75%	*			*	*	*		*	(Yağcı, 2022)
83.44%		*		*				*	(Dabhade et al., 2021)
95%								*	(Najafi & etal,2021)
				*	*			*	(Soltani & etal,2021)
50-81%	*			*	*		*		(Cruz-Jesus et al., 2020)
				*		*	*	*	(Sokkhey & Okazaki, 2020)
							*	*	(Rebai et al., 2020)
						*	*	*	(Jayaprakash et al., 2020)
			*				*	*	(Zulfiker et al., 2020)
			*						(Musso et al., 2020)
85%			*						(Waheed et al., 2020)
			*	*	*	*		*	(Salal & Abdullaev, 2019)
			*		*	*		*	(Turabieh, 2019)
			*	*				*	(Xu et al., 2019)
						*		*	(ghodoosi & etal,2019)
95.84%						*			(fadavi & etal,2019)
91.5%				*	*	*		*	(Ajibade et al., 2019)
85%			*						(Ahmad & Shahzadi, 2018)
						*		*	(Hasani & Bazrafshan, 2018)

Accuracy	LR	Line RL	ANN	SVM	KNN	NB	RF	DT	Data mining algorithm
		*				*	*	*	(Hussain et al., 2018)
	*				*	*	*	*	(Umer et al., 2017)
						*		*	(Khasanah, 2017)
								*	(Asif et al., 2017)
92.34%	*		*				*		(Hoffait & Schyns, 2017)
			*					*	(khosravi &etal,2017)
86%			*			*		*	(Mueen et al., 2016)
					*	*		*	(Amrieh et al., 2015)
92.34%						*		*	(Yehuala, 2015)
	*			*				*	(zahedi & etal,2015)
								*	(Punlumjeak & Rachburee, 2015)
71%							*	*	(Osmanbegović et al., 2014)
								*	(Shamloo & et al.,2014)
								*	(Asadi & et al.,2013)
60-75%					*	*		*	(Kabakchieva, 2013)
96%			*			*	*	*	(Oskouei & Askari, 2014)
						*		*	(Nghe et al., 2007)
94.17%			*	*	*	*	*	*	present research

بررسی ادبیات تحقیق نشان داد که عملکرد تحصیلی توسط متغیرهای متعدد و متنوعی از جمله ویژگی‌های جمعیت شناختی دانشجویان پیش‌بینی گردیده است (Costa-Mendes et al., 2021) (Rebai et al., 2020; Rizvi et al., 2019; Musso et al., 2020) . سوابق تحصیلی و عوامل جمعیت شناختی دانشجویان بهترین ویژگی برای پیش‌بینی عملکرد تحصیلی هستند (Batool et al., 2023). در این تحقیق با توجه به پیشینه پژوهش به این نتیجه رسیدیم که داده‌های اقتصادی- اجتماعی از اهمیت کمتری نسبت به داده‌های تحصیلی و رفتاری دانشجویان برخوردار بوده و استفاده از آن‌ها کمتر توصیه می‌شود. تعدادی از محققین مانند یاغجی(۲۰۲۲)، هافایت و شینز(۲۰۱۷) و برناکی(۲۰۲۰) معتقدند که به دلیل حجم زیاد داده‌های آموزشی و دشواری در جمع‌آوری داده‌ها، داده‌های اجتماعی- اقتصادی برای پیش‌بینی غیرضروری هستند(Hoffait & Schyns, 2017)

(Bernacki et al., 2020) (Yağcı, 2022). در این تحقیق نیز مانند بسیاری از تحقیقات قبلی از اعتبارسنجی متقاطع برای ارزیابی نتایج خود استفاده کرده‌ایم. بررسی ادبیات تحقیق نشان داد که با مجموعه داده‌های بزرگ‌تر می‌توان به آموزش مدل بهتری دست یافت(Dabhade et al., 2021). سلول‌های هایلایت شده در جدول ۲ بر اساس تحقیقات گذشته الگوریتم‌های طبقه‌بندی را نشان می‌دهد که بیشترین دقت و اثربخشی را در پیش‌بینی عملکرد دانشجویان در تحقیق مربوطه داشته‌اند. الگوریتم درخت تصمیم در تحقیقات قبلی بیشترین استفاده را داشته است. الگوریتم NB پس از درخت تصمیم بیشترین استفاده را در تحقیقات داشته است. الگوریتم‌های RF و ANN در رتبه بعدی استفاده قرار دارند. پس‌ازآن الگوریتم‌های SVM و KNN در تحقیقات استفاده شده‌اند. در تحقیق کومار و همکاران(۲۰۱۸)، همچنین نشان داده شد که اکثر محققان از الگوریتم درخت تصمیم C4.5 (J48) برای پیش‌بینی عملکرد دانشجو استفاده می‌کنند و سپس الگوریتم‌های ID3، CART و Naïve Bayes اغلب توسط محققان استفاده می‌شوند. برخی از محققان همچنین از ترکیب این الگوریتم‌ها برای پیش‌بینی عملکرد دانشجویان در داده‌کاوی آموزشی استفاده کردند(Kumar et al., 2018). در مقاله بتول و همکاران(۲۰۲۳)، نتایج نشان می‌دهد که ANN و Random Forest بیشترین استفاده را در بین الگوریتم‌های داده‌کاوی دارند(Batool et al., 2023). نتایج تحقیق الشنقیتی و نامون(۲۰۲۰) نیز نشان داد که الگوریتم‌های جنگل تصادفی در تحقیقات داده‌کاوی آموزشی در سال‌های اخیر بیشتر مورداستفاده قرار گرفته است (Alshanqiti & Namoun, 2020). با توجه به بررسی ادبیات تحقیق متوجه تحقیقات اندک در ایران شدیم و به لزوم انجام تحقیقات جدید برای پر کردن شکاف تحقیقاتی موجود در ایران با سایر کشورها پی بردیم.

روش تحقیق

هدف این پژوهش بررسی عوامل مؤثر در پیش‌بینی عملکرد تحصیلی دانشجویان مقطع کارشناسی در دانشگاه شاهد با استفاده از داده‌کاوی آموزشی بر اساس مدل‌های طبقه‌بندی می‌باشد. در این تحقیق عملکرد دانشجویان در پایان تحصیل با توجه به معدل نهایی آن‌ها

پیش‌بینی گردیده است. برای دستیابی به این هدف، این مطالعه از روش رایج داده‌کاوی آموزشی CRISP-DM پیروی می‌کند و شامل شش مرحله است. در این روش از طریق پیش‌پردازش داده‌ها، داده‌ها را از حالت خام خارج کرده و به داده‌هایی تبدیل می‌کنیم که بتوان از آن‌ها برای داده‌کاوی و مدل‌سازی استفاده کرد. مجموعه داده‌ها از پایگاه موجود در سامانه آموزشی دانشجویی ناد برای مقطع کارشناسی در رشته‌های غیرپزشکی دانشگاه شاهد و ورودی سال‌های ۱۳۹۰ تا ۱۴۰۰ استخراج شده است. مشخصات دانشجویان مانند معدل و جزئیات پذیرش آن‌ها در یک فایل اکسل قرار داده شده است. همچنین بر اساس تحقیقات قبلی و نظرات کارشناسان، ویژگی‌های مؤثر بر عملکرد دانشجویان شناسایی شده است. پس از پیش‌پردازش و تجزیه‌وتحلیل دقیق، ۱۹ ویژگی برای مطالعه انتخاب شد. برای کدگذاری ویژگی‌های مجموعه داده به ویژگی‌های عددی، از تکنیک Label Encoder استفاده کردیم. در این تحقیق از الگوریتم‌های طبقه‌بندی درخت تصمیم C4.5 و ID3، جنگل تصادفی، Naïve Bayes، k- نزدیک‌ترین همسایه و شبکه عصبی مصنوعی و درخت تقویت‌شده گرادیان برای تحلیل و طبقه‌بندی دانشجویان و پیش‌بینی معدل نهایی استفاده شد. مدل‌سازی با استفاده از ابزار رپیدماینر ۹,۹ انجام شد. در مدل‌های پیش‌بینی، کار چالش‌برانگیز انتخاب تکنیک‌های مؤثری است که بتواند دقت پیش‌بینی رضایت‌بخشی را ایجاد کند؛ بنابراین، برای بهبود عملکرد طبقه‌بندی و حل مشکل طبقه‌بندی اشتباه، از ترکیبی از تجزیه‌وتحلیل مؤلفه اصلی[1] همراه با الگوریتم‌های یادگیری ماشین و با استفاده از تکنیک‌های انتخاب ویژگی[2] و الگوریتم‌های بهینه‌سازی[3] استفاده می‌کنیم. در این تحقیق، دقت پیش‌بینی با استفاده از روش اعتبارسنجی متقاطع ۱۰ برابری برای همه الگوریتم‌ها مورد ارزیابی قرار گرفت. همچنین الگوریتم‌های مختلف با استفاده از روش توصیفی تحلیلی و بر اساس معیارهای ارزیابی مقایسه شدند و بهترین مدل پیش‌بینی در این تحقیق معرفی شد. متغیر هدف مورد پیش‌بینی در این تحقیق معدل نهایی کلیه واحدهای

1. PCA
2. Feature selection
3. Optimization algorithms

کسب‌شده در زمان فارغ‌التحصیلی دانشجویان می‌باشد که همان‌طور که در بخش مبانی نظری در پیشینه تحقیق به آن اشاره شد شاخص مناسبی برای ارزیابی و پیش‌بینی وضعیت تحصیلی دانشجویان می‌باشد. در نمونه تحقیق حاضر تعدادی از دانشجویان همچنان در حال تحصیل هستند و معدل فارغ‌التحصیلی آن‌ها در مجموعه داده نامشخص می‌باشد؛ بنابراین در مرحله پردازش داده‌ها، فیلدهای بدون نمره مربوط به معدل نهایی را با میانگین معدل‌های دانشجویان جایگزین می‌کنیم و بنابراین هیچ دانشجویی از تحقیق حذف نمی‌شود. در فایل اکسل اولیه دانشجویان، مقادیر معدل دیپلم، معدل ترم اول، معدل ترم دوم و معدل نهایی فارغ‌التحصیلی از ۰ تا ۲۰ می‌باشد و این ویژگی‌ها بر اساس مدل گسسته سازی طبقه‌بندی شده است. فارغ‌التحصیلان بر اساس معدل نهایی در سطوح عالی، خوب، متوسط و ضعیف طبقه‌بندی و سپس به‌صورت عددی و کمی کدگذاری شدند.

فرآیند CRISP-DM

فرآیند داده‌کاوی باید توسط افرادی با سابقه داده‌کاوی کمی قابل‌اعتماد و تکرار باشد و برای این منظور استانداردهایی ایجاد شده است: KDD SEMMA, CRISP DM، (Oreski et al., 2017). در این تحقیق از CRISP DM که مخفف فرایند استاندارد بین صنعتی برای داده‌کاوی است استفاده شده است (Wirth & Hipp, 2000).

- درک تجاری

نگاهی به معضل عملکرد ضعیف دانشجویان مقطع کارشناسی دانشگاه شاهد; نتایج تحصیلی جمع‌آوری‌شده توسط دانشجویان و گفتگو با مدیران دانشگاه نشان از وجود مشکل تحصیلی دانشجویان در دانشگاه شاهد دارد.

- درک داده‌ها

در بررسی ادبیات تحقیق، ویژگی‌های دانشجویان در رابطه با عملکرد تحصیلی پایین، معدل پایین و ترک تحصیل بررسی شد. محاسبه‌گر حجم نمونه نشان می‌دهد که برای

جمعیت ۷۰۰۰ نفری (تعداد کل دانشجویان مقطع کارشناسی دانشگاه شاهد)، سطح اطمینان ۹۵ درصد و حاشیه خطای ۵ درصد، حجم نمونه ۳۶۵ و بالاتر می‌تواند نتایج به‌دست‌آمده را تعمیم دهد (Raosoft, 2004).

- **جمع‌آوری داده‌ها**

داده‌ها از سامانه اطلاعات دانشجویی آموزشی ناد گرفته شده است. در این مقاله از فاصله[1] اقلیدسی برای حذف نقاط پرت[2] در نرم‌افزار رپیدماینر استفاده شده است. همچنین با نرمال‌سازی داده‌ها[3] که یکی از رویکردهای پیش‌پردازش است داده‌ها به یک مقیاس در می‌آیند و یا تبدیل[4] می‌شوند تا سهم یکسانی از هر ویژگی داشته باشند. موفقیت الگوریتم‌های یادگیری ماشین به کیفیت داده‌ها برای به دست آوردن یک مدل پیش‌بینی تعمیم‌یافته[5] از مسئله طبقه‌بندی بستگی دارد (Singh & Singh, 2020).

- **آماده‌سازی داده‌ها**[6]

در مرحله انتخاب داده‌ها[7] با استفاده از عملگر append در برنامه Rapidminer9.9، دو مجموعه داده معدل تحصیلی دانشجویان و داده‌های هنگام ثبت‌نام‌شان را با هم ادغام کردیم.

در پاک‌سازی داده‌ها به دنبال داده‌های غیرضروری یا مقادیر ازدست‌رفته[8] یا ناقص می‌گردیم و این داده‌ها را حذف یا دست‌کاری می‌کنیم. در مورد متغیرهایی که با یکدیگر همبستگی دارند، متغیرهای غیرضروری و کم تأثیر را حذف کردیم. متغیرهای با واریانس کم نیز حذف شدند. معدل‌هایی که مقادیر ازدست‌رفته داشتند با میانگین معدل جایگزین

1. Detect Outlier Operator Based on Distance
2. Outliers
3. Data Normalization
4. Scaled or Transformed
5. Generalized
6. Data Preparation
7. Data Selection
8. Missing Values

شدند. در این تحقیق از PCA[1] برای پاک‌سازی داده‌ها استفاده شد. تحلیل مؤلفه اصلی یک روش ریاضی است که با استفاده از تبدیل متعامد[2]، مجموعه‌ای از مشاهدات احتمالاً همبسته را به مجموعه‌ای از مقادیر مشاهدات غیر همبسته تبدیل می‌کند که مؤلفه‌های اصلی نامیده می‌شوند. هر مؤلفه‌ای که واریانس بالایی داشته باشد را می‌توان با سایر مؤلفه‌ها نامرتبط دانست و ویژگی‌ها را بر اساس واریانس آن‌ها مرتب کرد. در این روش، با نشان دادن داده‌ها به‌صورت چند بردار متعامد[3]، زیرمجموعه‌ای از ویژگی‌های بهینه[4] به دست می‌آید (Jolliffe & Cadima, 2016).PCA با کاهش ابعاد داده‌ها بدون از دست دادن ارزش، تجسم داده‌ها و عملکرد سریع‌تر الگوریتم‌های یادگیری ماشین را فراهم می‌کند (Dabhade et al., 2021). در مرحله تبدیل داده‌ها[5] تمامی داده‌های کیفی و اسمی به داده‌های عددی و کمی تبدیل شد مانند ویژگی جنسیت که از مرد به ۱ و از زن به ۲ تبدیل شد. در مرحله کاهش داده‌ها[6] چندین رکورد بدون داده یا ناقص و با مقدار ازدست‌رفته بودند. مرحله انتخاب ویژگی[7] در افزایش دقت پیش‌بینی، افزایش کارایی یادگیری و کاهش پیچیدگی نتایج آموخته‌شده مؤثر است (Ramaswami & Bhaskaran, 2010). ویژگی‌های نامربوط در مجموعه داده، نتایج پیش‌بینی را کاهش می‌دهد و زمان پردازش مدل را افزایش می‌دهد(Batool et al., 2023). انتخاب ویژگی امکان شناسایی مجموعه‌ای از ویژگی‌های ضروری مرتبط با هدف تحقیق و یک راه‌حل بهینه را بدون افزایش پیچیدگی مدل‌سازی فراهم می‌کند(Mwadulo, 2016). روش‌های فیلترینگ ویژگی‌های مرتبط را با استفاده از معیارهایی مانند اطلاعات، فاصله، همبستگی و سازگاری[8] ارزیابی می‌کنند (Sasikala et al., 2016). در این تحقیق از روش‌های فیلتر با چهار تکنیک نسبت

1. Principal Component Analysis Algorithm
2. Orthogonal Transformation
3. Orthogonal Vectors
4. An Optimal Subset of Features
5. Data Transformation
6. Data Reduction
7. Feature Selection
8. Information, Distance, Correlation, and Consistency

بهره[1]، همبستگی[2]، Relief و بهره اطلاعات[3] استفاده گردید (Mwadulo, 2016). در مدل‌های یادگیری ماشین، تمام داده‌ها باید متغیرهای عددی باشند؛ بنابراین در مرحله کدگذاری ویژگی‌ها[4]، از تکنیک‌های مختلفی از جمله LabelEncoder و One-Hot Encoding استفاده کردیم. LabelEncoder به ما MAPE کمتری داد که در مجموعه داده خود استفاده کردیم. متغیرهای مورداستفاده در تحقیق و مقادیر ممکن در جدول ۳ آورده شده است.

جدول ۳. ویژگی‌های مؤثر بر پیش‌بینی عملکرد تحصیلی دانشجویان

	ویژگی‌ها	مقادیر کدگذاری شده	کیفی - اسمی
۱	جنسیت	مرد: ۱ زن: ۲	کیفی
۲	سن	اعداد	اسمی
۳	رشته-گرایش	علوم کامپیوتر: ۱ مدیریت بازرگانی: ۲ مدیریت دولتی: ۳ مدیریت صنعتی: ۴ مهندسی برق: ۵ مهندسی برق-کنترل: ۵ مهندسی برق-الکترونیک: ۵ مهندسی برق-قدرت: ۵ رشته برق-ارتباطات: ۵ مهندسی پزشکی: ۶ مهندسی پزشکی - بالینی: ۶ مهندسی عمران: ۷ مهندسی عمران: ۷ مهندسی کامپیوتر: ۸ مهندسی کامپیوتر (سخت‌افزار):۸ مهندسی کامپیوتر (معماری سیستم‌های کامپیوتر):۸	کیفی - اسمی
۴	سال ورود به دانشگاه	سال ورود ۱۴۰۰ ۱ سال ورود ۱۳۹۹ ۲ سال ورود ۱۳۹۸ ۳ سال ورود ۱۳۹۰ تا ۱۳۹۷ ۴	کیفی - اسمی
۵	نحوه پذیرش	آزمون سازمان سنجش: ۱ انتقال از دانشگاه‌های دیگر با تغییر رشته: ۲ انتقال از دانشگاه‌های دیگر بدون تغییر رشته: ۲ انتقال درون دانشگاهی با تغییر رشته: ۳ مهمان: ۴ تغییر رشته در دانشگاه: ۳ دوره دوم داخل دانشگاه: ۴ مهمان‌ترم کامل از سایر دانشگاه‌ها: ۳ میهمانان یک دوره از دانشگاه‌های دیگر: ۳	کیفی - اسمی
۶	محل زندگی	بومی: ۱ غیربومی: ۲	کیفی - اسمی

1. Gain Ratio
2. Correlation
3. Information Gain
4. Features Encoding

	ویژگی‌ها	مقادیر کدگذاری شده	کیفی - اسمی
۷	وضعیت سربازی	دانشجویان سال آخر متوسطه یا پیش‌دانشگاهی: ۱ دارای مجوز تحصیلی از نظام‌وظیفه: ۲ معافیت برای دختران (شامل نمی‌شود): ۳ فارغ‌التحصیل کارشناسی ارشد تا یک سال پیش: ۴ معاف دائم: ۵ معاف تحت نظارت: ۵ سایر موارد: ۶	کیفی - اسمی
۸	وضعیت سهمیه	منطقه یک ۱ ایثارگر ۵ درصد ظرفیت ۲ منطقه ۳ آزاد ۴ منطقه ۲ ۵ ایثارگران ۲۵ درصد ظرفیت ۶ فرزندان جانباز ۷ شهید ۸ شاهد ۹ فرزند جانباز آزاده ۱۰ فرزند هیئت‌علمی ۱۱ رزمنده ۱۲ فرزندان آزاده ۱۰ فرزندان جانبازان منطقه ۳ ۷	کیفی - اسمی
۹	گروه آموزشی	فنی و مهندسی: ۱ علوم کامپیوتر: ۲ مدیریت: ۳	کیفی - اسمی
۱۰	دانشکده	دانشکده فنی و مهندسی: ۱ دانشکده علوم پایه: ۲ دانشکده علوم انسانی: ۳	کیفی - اسمی
۱۱	وضعیت تأهل	مجرد: ۱ متأهل: ۲ ناشناس: ۳	کیفی - اسمی
۱۲	دوره	روزانه: ۱ نوبت دوم: ۲ شبانه: ۳	کیفی - اسمی
۱۳	معدل دیپلم	اعداد	کیفی
۱۴	تعداد واحد افتاده	اعداد	اسمی
۱۵	تعداد ترم‌های ممتازی	۰ ترم: ۱ ۱ ترم: ۲ ۲ ترم: ۳ بیشتر از ۳ ترم: ۴	کیفی - اسمی
۱۶	تعداد نیمسال‌های مشروطی	۰ ترم: ۱ ۱ ترم: ۲ ۲ ترم: ۳ بیشتر از ۳ ترم: ۴	کیفی - اسمی
۱۷	معدل ترم ۱	طبقه ۱: شامل دانشجویان با عملکرد عالی (۲۰ > معدل نهایی >=۱۷) طبقه ۲: شامل دانشجویان با عملکرد خوب (۱۷ > معدل نهایی >=۱۵) طبقه ۳: شامل دانشجویان با عملکرد متوسط (۱۴ = معدل نهایی) طبقه ۴: شامل دانشجویان با عملکرد ضعیف (۱۴ > معدل نهایی)	کیفی - اسمی
۱۸	معدل ترم ۲	طبقه ۱: شامل دانشجویان با عملکرد عالی (۲۰ > معدل نهایی >=۱۷) طبقه ۲: شامل دانشجویان با عملکرد خوب (۱۷ > معدل نهایی >=۱۵) طبقه ۳: شامل دانشجویان با عملکرد متوسط (۱۴ = معدل نهایی) طبقه ۴: شامل دانشجویان با عملکرد ضعیف (۱۴ > معدل نهایی)	کیفی - اسمی

	ویژگی‌ها	مقادیر کدگذاری شده	کیفی - اسمی
۱۹	معدل نهایی	طبقه ۱: شامل دانشجویان با عملکرد عالی (۲۰ > معدل نهایی >=۱۷) طبقه ۲: شامل دانشجویان با عملکرد خوب (۱۷ > معدل نهایی >=۱۵) طبقه ۳: شامل دانشجویان با عملکرد متوسط (۱۴ = معدل نهایی (طبقه ۴: شامل دانشجویان با عملکرد ضعیف (۱۴ > معدل نهایی(	کیفی - اسمی

- مدل‌سازی

روش طبقه‌بندی، یک مورد خاص از پیش‌بینی است که در آن یک کلاس (برچسب یا مقدار گسسته[1]) با استفاده از یک طبقه‌بندی پیش‌بینی می‌شود. یک طبقه‌بندی کننده با استفاده از داده‌های آموزشی[2] یک مدل طبقه‌بندی ایجاد می‌کند. این مدل باید به‌خوبی با داده‌های آموزشی مطابقت داشته باشد و کلاس داده‌های مجهول (داده‌های آزمون[3]) را به‌خوبی پیش‌بینی کند (Han & Kamber, 2006)

- درختان تصمیم[4]

الگوریتم‌های درخت تصمیم در داده‌کاوی آموزشی به‌عنوان یک روش قدرتمند برای طبقه‌بندی دانشجویان بر اساس ایستگاه یادگیری آن‌ها در حال ظهور هستند (Chen & Lin, 2023). درختان تصمیم با محاسبه شاخص جینی، وزن‌های خاصی به ویژگی‌های یک مجموعه داده می‌دهد و با توجه به قوانین تولیدشده، ویژگی‌های مؤثر پیش‌بینی می‌شود (Alsalman et al., 2019). آن‌ها برای متغیرهای عددی بازه‌ای مناسب هستند، جایی که هر گره شاخه[5] نشان‌دهنده انتخابی از میان تعدادی گزینه و هر گره برگ[6] نشان‌دهنده یک تصمیم است(Quinlan, 1986). الگوریتم درخت تصمیم باید بهینه‌ترین

1. Label or Discrete Value
2. Training Data
3. Test Data
4. Decision Trees
5. Branch Node
6. Leaf Node

درخت را پیدا کند. درخت این الگوریتم از درجه ناخالصی گره‌های فرزند[1] برای تعیین بهترین تقسیم استفاده می‌کند(Tan et al., 2013) . مقدار آنتروپی بالا نشانه توزیع کلاس همگن‌تر است(Yulianto et al., 2020) .

- درخت تصمیم ID3

ID3 یک درخت تصمیم‌گیری ساده است که توسط رأس کوینلان[2] توسعه یافته است (Ogunde & Ajibade, 2014; Surjeet & Saurabh, 2012; Yadav et al., 2012)

الگوریتم ID3 ویژگی‌های نمونه‌های آموزشی[3] را جستجو می‌کند و ویژگی‌ای را استخراج می‌کند که بهترین نمونه‌ها را از هم جدا می‌کند. این الگوریتم از جستجوی حریصانه[4] استفاده می‌کند، یعنی به عقب نگاه نمی‌کند و بهترین ویژگی را انتخاب می‌کند. ID3 در صورتی متوقف می‌شود که به طور کامل ویژگی مجموعه‌های آموزشی را طبقه‌بندی کند. در غیر این صورت، به‌صورت بازگشتی بر روی زیرمجموعه تقسیم‌بندی شده n(که در آن n = تعداد مقادیر ممکن یک ویژگی) عمل می‌کند تا «بهترین» ویژگی خود را پیدا کند(Ogunde & Ajibade, 2014).

- درخت تصمیم C4.5

الگوریتم C4.5 بهبود الگوریتم IDE3 توسعه یافته است (Ross Quinlan, 1993). به C4.5 الگوریتم J48 نیز گفته می‌شود و بهتر از ID3 است مانند:

- می‌تواند داده‌های طبقه‌بندی‌شده و گسسته را اداره کند.
- الگوریتم درخت تصمیم C 4.5 می‌تواند مقادیر ازدست‌رفته را مدیریت کند. مقادیر ویژگی ازدست‌رفته به‌سادگی در محاسبات بهره (gain) و آنتروپی استفاده نمی‌شود.
- C4.5 درختان را با بازگشت به درخت هرس[5] می‌کند. با جایگزین کردن گره‌های

1.He Impurity Degree of Child Nodes
2. Ross Quinlan
3. Training Samples
4. Greedy Search
5. Prunes

داخلی با گره‌های برگ، شاخه‌هایی[1] را که کمکی نمی‌کنند حذف می‌کند (Nijhawan et al., 2017).

- الگوریتم جنگل تصادفی

جنگل تصادفی یک الگوریتم یادگیری نظارت شده است که توسط لئو بریمن در سال ۲۰۰۱ توسعه یافته است(Breiman, 2001).این یک شکل از روش یادگیری ماشین است که از رگرسیون و طبقه‌بندی بر اساس تجمیع تعداد زیادی درخت تصمیم استفاده می‌کند (Akar & Güngör, 2012). هر رگرسیون درخت تصمیم یک عدد را به‌عنوان خروجی برای یک مجموعه داده معین پیش‌بینی می‌کند. رگرسیون RF میانگین آن پیش‌بینی‌ها را به‌عنوان خروجی "نهایی" خود می‌گیرد (Kulkarni, 2014) . جنگل تصادفی از طریق رأی‌گیری بهترین راه‌حل را انتخاب می‌کند. طبقه‌بندی کننده جنگل تصادفی از شاخص جینی به‌عنوان معیار انتخاب ویژگی استفاده می‌کند که ناخالصی یک ویژگی را با توجه به کلاس‌ها اندازه‌گیری می‌کند (Ahmed & Sadiq, 2018; Pal, 2005).

- الگوریتم شبکه عصبی[2]

مطالعه مک کالوچ در مورد شبیه‌سازی یک سیستم عصبی بیولوژیکی منجر به توسعه روش ANN در دهه ۱۹۴۰ شد. NN ها یکی از رایج‌ترین و کارآمدترین سیستم‌های یادگیری هستند. شبکه‌های عصبی مصنوعی به همان شیوه مغز انسان‌ها، از خود یاد می‌گیرند، خود را آموزش و تغییر می‌دهند(Russell & Norvig, 2016). شبکه‌های عصبی مصنوعی رویدادهای مرتبط را تعمیم می‌دهند، اطلاعات را جمع‌آوری می‌کنند و در مورد رویدادهای جدیدی که با آن‌ها مواجه می‌شوند تصمیم می‌گیرند(Bahadir, 2016).

1. Branches
2. Artificial Neural Network

- **الگوریتم K نزدیک‌ترین همسایه**[1]

روش KNN برای اولین بار در اوایل دهه ۱۹۵۰ توصیف شد که یک الگوریتم یادگیری ماشینی نظارت شده است. KNN کل مجموعه آموزشی را حفظ می‌کند و تمام تلاش‌های تعمیم استقرایی[2] را تا انجام رگرسیون به تعویق می‌اندازد (Han et al., 2011). در مرحله آموزش، این الگوریتم تنها بردارهای ویژگی را ذخیره می‌کند و داده‌های نمونه آموزشی را طبقه‌بندی می‌کند. سپس در مرحله طبقه‌بندی، همان ویژگی‌ها برای تست داده‌ها (با طبقه‌بندی ناشناخته) محاسبه می‌شود. فاصله این بردار جدید تا همه بردارهای نمونه آموزشی محاسبه شده و نزدیک‌ترین عدد k گرفته می‌شود (Yulianto et al., 2020).

- **الگوریتم درخت تقویت‌شده گرادیان**[3]

این تکنیک یک پیش‌بینی را در قالب مجموعه‌ای از مدل‌های پیش‌بینی ضعیف درخت تصمیم انجام می‌دهد. برای پیش‌بینی، GB از یک یادگیرنده قوی[4] استفاده می‌کند که طبقه‌بندی‌کننده‌ای است که به طور دلخواه همبستگی خوبی با طبقه‌بندی واقعی دارد و از ترکیبی از موارد مختلف ساخته می‌شود. یادگیرندگان ضعیف طبقه‌بندی کننده‌هایی هستند که فقط کمی با طبقه‌بندی کننده واقعی همبستگی دارند. در GB، این رویکرد به‌صورت تکراری اعمال می‌شود (Machado et al., 2019). این طبقه‌بندی کننده با استفاده از تکنیک‌های هسته[5] می‌تواند قبل از پارتیشن‌بندی غیرخطی را به خطی تبدیل کند (Adejo & Connolly, 2018).

- **طبقه‌بندی کننده نایو بیز**[6]

این طبقه‌بندی بر اساس قضیه بیز است. فرض ساده استقلال شرطی کلاس[7] اغلب برای

1. K Nearest Neighbor Algorithm
2. Inductive Generalization
3. Gradient Boosted Tree Algorithm
4. Robust Learner
5. Kernel Techniques
6. Naïve Bayes Classifier
7.He Naive Assumption Of Class Conditional Independence

کاهش هزینه محاسباتی[1] ساخته می‌شود (Leung, 2007). الگوریتم Naive Bayes در برابر ویژگی‌های نامرتبط قوی عمل می‌کند. بااین‌حال، ویژگی‌های مرتبط عملکرد آن را کاهش می‌دهد (Tan et al., 2013). الگوریتم Naive Bayes یک طبقه‌بندی کننده احتمالی ساده است که مجموعه‌ای از احتمالات را با جمع‌کردن فراوانی‌ها و ترکیب مقادیر مجموعه داده‌ها محاسبه می‌کند. این الگوریتم احتمال عضویت هر کلاس را پیش‌بینی می‌کند و محتمل‌ترین کلاس در نظر گرفته می‌شود. وجود یا عدم وجود یک ویژگی بر وجود یا عدم وجود هیچ ویژگی دیگری تأثیر نمی‌گذارد(Karo et al., 2022).

- مرحله آزمون و ارزیابی[2]

در مرحله ارزیابی مدل، یافتن مدلی با بهترین ویژگی‌ها که قابلیت تعمیم اطلاعات را داشته باشد، انجام می‌شود. اعتبارسنجی متقاطع ۱۰ برابری برای تأیید و اعتبارسنجی نتایج الگوریتم‌های تحقیق برای اندازه‌گیری‌های دقیق و دقت بالا استفاده شد (Anguita et al., 2009; Arlot & Celisse, 2010).

- اعتبار سنجی متقاطع[3]

اعتبار سنجی متقاطع یکی از تکنیک‌های ارزیابی/ اعتبارسنجی دقت یک مدل ساخته شده است. اعتبار سنجی متقاطع ۱۰ برابری به طور تصادفی مجموعه داده‌ها را به ۱۰ قسمت تقسیم می‌کند و آموزش و آزمایش را انجام می‌دهد و مدل را ۱۰ بار می‌سازد. این فرآیند شامل ۹۰ درصد از مجموعه داده برای آموزش و ۱۰ درصد برای آزمایش در هر تکرار است. نتایج آزمون داده‌شده در پایان فرآیند، ماتریس سردرگمی[4] است. ماتریس سردرگمی عملکرد طبقه‌بندی یک مدل را با توجه به داده‌های تجربی خلاصه می‌کند(Ting, 2017)..

1. Reduce Computational Cost
2. Test And Evaluation Stage
3. Cross Validation
4. The Confusion Matrix

- معیارهای ارزیابی عملکرد در RapidMiner

دقت[1] یک معیار بسیار مهم است، اما برای ارزیابی عملکرد مدل کافی نیست. هنگام ارزیابی، اغلب از میزان خطا[2] و دقت استفاده می‌کنیم؛ اما آنچه مهم است این است که مدل ما چقدر قابل‌اعتماد[3] است و چگونه برای مجموعه داده دیگری (تعمیم‌پذیری[4]) کار می‌کند و چقدر انعطاف‌پذیر است(Romero & Ventura, 2010; Shaikh et al., 2015) و بنابراین، به دلیل اینکه هر معیار جنبه‌های مختلفی از عملکرد کلی طبقه‌بندی می‌باشد، عملکرد طبقه‌بندی باید با در نظر گرفتن همه آن‌ها به طور هم‌زمان ارزیابی شود (Ballabio et al., 2018).

- ماتریس سردرگمی

ماتریس سردرگمی یک روش رایج برای محاسبه دقت در مفهوم داده‌کاوی است (Yulianto et al., 2020). ماتریس سردرگمی وضعیت فعلی و تعداد پیش‌بینی‌های صحیح/نادرست مدل را نشان می‌دهد. جدول ۴ ماتریس سردرگمی را نشان می‌دهد. عملکرد مدل با تعداد نمونه‌های طبقه‌بندی صحیح و تعداد نمونه‌های طبقه‌بندی اشتباه محاسبه می‌شود. سطرها تعداد واقعی نمونه‌ها را در مجموعه تست و ستون‌ها تخمین مدل را نشان می‌دهند(Costa et al., 2007).

جدول ٤. **Confusion matrix**

Data class	Classified as *pos*	Classified as *neg*
pos	true positive (*tp*)	false negative (*fn*)
neg	false positive (*fp*)	true negative (*tn*)

$$\begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix}$$

1. Accuracy
2. Error Rate
3. Reliable
4. Generalizability

دقت[1]

اندازه‌گیری دقت (Acc) اثربخشی طبقه‌بندی کننده را با درصد پیش‌بینی‌های صحیح آن ارزیابی می‌کند(Costa et al., 2007) . CA نسبت پیش‌بینی‌های صحیح (TP + TN) به تعداد کل نمونه‌ها (FN + TP + TN + FP) است (Sokolova & Lapalme, 2009).

معادله ۱,۳

(Costa et al., 2007)

$$Acc = \frac{|TN| + |TP|}{|FN| + |FP| + |TN| + |TP|}$$

صحت، درستی[2]

Precisionاحتمال درست بودن یک پیش‌بینی مثبت را تخمین می‌زند (Costa et al., 2007). Precision، نسبت مثبت‌های واقعی که به‌درستی پیش‌بینی می‌شود به تمام مقادیر مثبت پیش‌بینی‌شده توسط مدل است و مقداری در محدوده [۰,۱] دریافت می‌کند (Nisbet et al., 2009).

معادله ۲,۳

$$P = \frac{|TP|}{|TP| + |FP|}$$

یادآوری یا حساسیت[3]

اندازه‌گیری یادآوری (R) اثربخشی یک طبقه‌بندی‌کننده برای هر کلاس را در یک مسئله باینری ارزیابی می‌کند. Recall نسبت نمونه‌های متعلق به کلاس مثبت است که به‌درستی به‌عنوان مثبت پیش‌بینی می‌شوند (Costa et al., 2007). یادآوری که حساسیت یا sensitivity نیز نامیده می‌شود، توانایی یک مدل را برای طبقه‌بندی کارآمدی موارد مثبت اندازه‌گیری می‌کند (Sokolova & Lapalme, 2009).

معادله ۳,۳

$$R = \frac{|TP|}{|TP| + |FN|}$$

1. Accuracy
2. Precision
3. Recall Or Sensitivity

معادله ۴,۳

$$Spe = \frac{|TN|}{|FP| + |TN|}$$

(Sokolova & Lapalme, 2009) (Costa et al., 2007)

خطای طبقه‌بندی (نرخ خطا)[1]

Error rate و accuracy رایج‌ترین معیارهای عملکرد در مسائل طبقه‌بندی، از جمله طبقه‌بندی باینری و طبقه‌بندی چندگانه (چند کلاسه) هستند. میزان خطا طبقه‌بندی‌کننده را بر اساس درصد پیش‌بینی‌های نادرست آن ارزیابی می‌کند. میزان خطا نسبت نمونه‌های طبقه‌بندی اشتباه به همه نمونه‌ها است. درحالی‌که دقت نسبت نمونه‌هایی است که به‌درستی طبقه‌بندی شده‌اند. نرخ خطا به شرح زیر است

معادله ۵,۳

$$Err = \frac{|FN| + |FP|}{|FN| + |FP| + |TN| + |TP|} = 1 - Acc$$

(Costa et al., 2007)

- مرحله استقرار

کلیه پیاده‌سازی و پردازش داده‌کاوی در این مطالعه با استفاده از RapidMiner 9.9 و Excel انجام شده است.

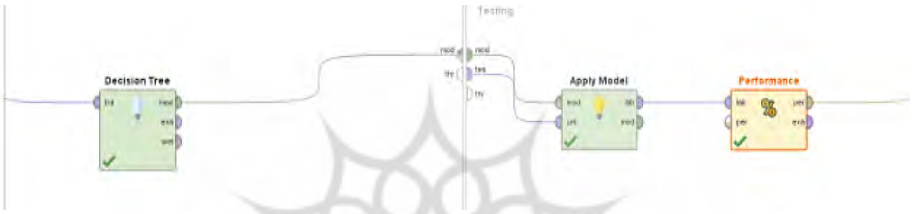
یافته‌ها

این مطالعه از ابزار مدل‌سازی RapidMiner 9.9 استفاده می‌کند که دارای چندین الگوریتم طبقه‌بندی برای مدل‌سازی داده‌ها است. بهترین مدل بهترین مقادیر را برای معیارهای اندازه‌گیری عملکرد انتخابی دارد(Lever et al., 2016). طبقه‌بندی شامل پیش‌بینی ارزش یک ویژگی (کلاس) بر اساس مقادیر سایر ویژگی‌ها (ویژگی‌های پیش‌بینی کننده) است. مدل‌های طبقه‌بندی در نظر گرفته‌شده برای اهداف این تحقیق عبارت‌اند از الگوریتم‌های یادگیری ماشین درخت تصمیم C4.5، K،ID3 نزدیک‌ترین

1. Error Classification (Error Rate)

همسایه ، Naive Bayes، NBK، شبکه عصبی، جنگل تصادفی و درختان تقویت‌شده گرادیان. تنظیمات پارامتر برای الگوریتم‌ها در جدول ۵ ارائه شده است. طبقه‌بندی‌کننده‌های C4.5 با استفاده از ابزار مدل‌سازی RapidMiner 9.9 ساخته می‌شوند و اعتبارسنجی متقاطع ۱۰ برابری برای آزمایش عملکرد مدل انجام می‌شود(شکل ۱). سایر الگوریتم‌های تحقیقاتی نیز با همین روش پیاده‌سازی و ارزیابی می‌شوند.

شکل ۱. مدل Decision tree C4.5 model

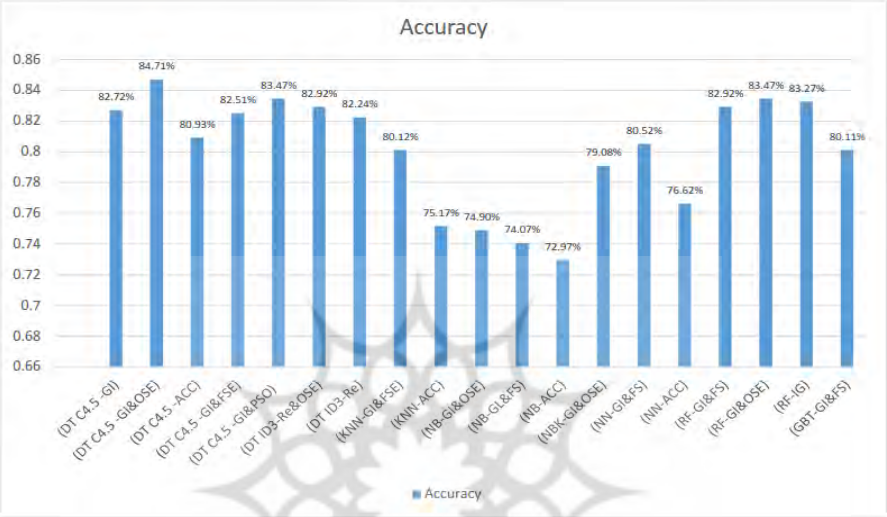


جدول ۵. تنظیم پارامترها در الگوریتم‌های طبقه‌بندی

الگوریتم	پارامترهای تنظیم
RF	Number of Tree = 20 Criterion =gain ratio maximal depth = 10 Voting strategy =Confidence vote
DT C4.5	Splitting criterion = information gain ratio Minimal size of split = Minimal leaf size= 1 Minimal gain = 0.01 Maximal depth = 20 Confidence = 0.1
DT ID3	Split criterion = the ratio of obtaining information the minimum split size = 4 Minimum leaf size = 2 Minimum gain = 0.01
ANN	training cycles = 500 Learning Rate = 0.3 momentum = 0.2 Error epsilon = 1.0E-5 The neural network has two hidden layers with 5 nodes. Parameter settings of Weight by Gini index operator in modeling with neural network weight relation =top p% p = 0.5
KNN	The number of K = 5 Measure Types = Mixed Measures Mixed Measure = Mixed Euclidean Distance
NB	Naive Bayes operator was set based on Laplace correlation. Weight relation= top p% P=0.5
GBT	Number of trees =50 maximal depth = 5 Min Rows= 10 min split improvement = 1.0E-5 Number of bins = 20 Learning Rate = 0.01 Sample rate = 1.0

شکل ۲ نموداری است که دقت الگوریتم‌های مورداستفاده در این تحقیق را مقایسه می‌کند.

شکل۲. نمودار مقایسه‌ای دقت الگوریتم‌های مورداستفاده در تحقیق



جدول ۶ مقادیر معیارهای ارزیابی به‌دست‌آمده از مدل‌سازی را برای تمامی الگوریتم‌های مورداستفاده در تحقیق حاضر در حالت چهار کلاسه نشان می‌دهد. با مقایسه این نتایج، بهترین الگوریتم برای پیش‌بینی عملکرد تحصیلی دانشجویان تعیین شد.

جدول ٦. مقایسه معیارهای ارزیابی مدل‌ها در طبقه‌بندی چهار کلاسه

	مدل پیش‌بینی کننده	دقت	صحت	یادآوری	خطای طبقه‌بندی
DT	Decision Tree C4.5 with Gini Index (DT C4.5 -GI)	82.72%	82.09%	82.07%	17.28%
	Decision Tree C4.5 with Gini Index &with Optimize selection (Evolutionary) (DT C4.5 -GI&OSE)	84.71%	84.17%	83.42%	15.29%
	Decision Tree C4.5 with Accuracy (DT C4.5 -ACC)	80.93%	81.48%	79.67%	19.07%

خطای طبقه‌بندی	یادآوری	صحت	دقت	مدل پیش‌بینی کننده	
17.49%	81.67%	83.08%	82.51%	Decision Tree C4.5 with Gini Index & Forward selection (Evolutionary) (DT C4.5 -GI&FSE)	
16.53%	82.60%	83.14%	83.47%	Decision Tree C4.5 with Gini Index & Optimize Weights (PSO) (DT C4.5 -GI&PSO)	
17.08%	82.76%	81.53%	82.92%	Decision Tree ID3 with Weight by Relief& Optimize selection (Evolutionary) (DT ID3-Re&OSE)	
17.76%	82.02%	80.83%	82.24%	Decision Tree ID3 with Weight by Relief (DT ID3-Re)	
19.88%	78.70%	81.67%	80.12%	k-Nearest neighbor(kNN) with Gini Index & Forward selection (Evolutionary) (KNN-GI&FSE)	KNN
24.83%	71.71%	74.86%	75.17%	k-Nearest neighbor with Accuracy (KNN-ACC)	
25.10%	73.63%	75.55%	74.90%	Naive Bayes with Gini Index & with Optimize Selection (Evolutionary)(NB-GI&OSE)	NB
25.93%	73.62%	73.76%	74.07%	Naive Bayes with Gini Index & with Forward Selection (NB-GL&FS)	
27.03%	72.65%	73.41%	72.97%	Naive Bayes with Accuracy (NB-ACC)	
20.92%	80.22%	79.72%	79.08%	Naive Bayes Kernel with Gini Index & with Optimize Selection Evolutionary) (NBK-GI&OSE)	
19.48%	79.07%	81.54%	80.52%	Neural Network with Gini Index &Forward selection (NN-GI&FS)	NN
23.38%	75.66%	74.42%	76.62%	Neural Network with Accuracy (NN-ACC)	
17.08%	81.48%	83.79	82.92%	Random Forest with Gini Index &Forward selection (RF-GI&FS)	RF
16.53%	82.19%	83.61%	83.47%	Random Forest with Gini Index &Optimize Selection (Evolutionary)(RF-GI&OSE)	

	مدل پیش‌بینی کننده	دقت	صحت	یادآوری	خطای طبقه‌بندی
	Random Forest Trees with Information Gain (RF-IG)	83.27%	83.74%	81.46%	16.73%
GBT	Gradient Boosted Trees with Gini Index &Forward selection (GBT-GI&FS)	80.11%	79.65%	79.35%	19.89%

نتایج مدل‌سازی در حالت چهار کلاسه

نتایج جدول ۶ نشان می‌دهد که در حالت چهار کلاسه، الگوریتم درخت تصمیم C4.5 (با اعمال شاخص جینی و با الگوریتم انتخاب بهینه (تکاملی)) بالاترین دقت را با ۸۴,۷۱ درصد نسبت به سایر الگوریتم‌ها دارد؛ بنابراین طبقه‌بندی‌کننده DT C4.5-GI&OSE به‌عنوان مدل مهم و اصلی در پیش‌بینی در این تحقیق می‌باشد. این مدل علاوه بر دقت بالاتر نسبت به مدل‌های دیگر، از نظر سایر معیارهای ارزیابی مانند درستی، یادآوری، خطای طبقه‌بندی، همبستگی، کندال و اسپیرمن جایگاه بالاتری نسبت به سایر مدل‌ها دارد. با توجه به نتایج حاصل از ماتریس سردرگمی مدل‌های تحقیق، مدل درخت تصمیم DT ID3-Re&OSE با پیش‌بینی ۸۲,۶۹٪ بهترین روش برای پیش‌بینی دانشجویان عالی می‌باشد. مدل درخت تصمیم DTC4.5-OSE&GIبهترین روش برای پیش‌بینی دانشجویان خوب با پیش‌بینی ۸۵/۲۶ درصد، مدل DT C4.5-GI&PSO بهترین روش برای پیش‌بینی دانشجویان متوسط با پیش‌بینی ۸۵,۷۱ درصد و الگوریتم RF-OS&GI بهترین تشخیص را برای پیش‌بینی دانشجویان ضعیف و ناموفق (۹۰٪) در مقایسه با سایر الگوریتم‌ها دارند.

جدول ۷. ماتریس سردرگمی مدل تحقیق DT C4.5-GI&OSE در حالت ٤ کلاسه

صحت	دانشجویان با عملکرد ضعیف	دانشجویان با عملکرد متوسط	دانشجویان با عملکرد خوب	دانشجویان با عملکرد عالی	
٪۷۸,۶۴	۰	۰	۲۲	۸۱	پیش‌بینی ۱
٪۷۸,۶۷	۹	۴۹	۲۹۵	۲۲	پیش‌بینی ۲
٪۸۶,۴۶	۵۰	۴۹۸	۲۷	۱	پیش‌بینی ۳

صحت	دانشجویان با عملکرد ضعیف	دانشجویان با عملکرد متوسط	دانشجویان با عملکرد خوب	دانشجویان با عملکرد عالی	
٪۸۹,۳۶	۳۶۱	۴۱	۲	۰	پیش‌بینی ۴
	٪۸۵,۹۵	٪۸۴,۶۹	٪۸۵,۲۶	٪۷۷,۸۸	یادآوری

با توجه به نتایج ماتریس سردرگمی DT C4.5-GI&OSE در حالت چهار کلاسه در جدول ۷، الگوریتم DTC4.5 قادر به پیش‌بینی کلاس ۱۲۳۵ شی از ۱۴۵۸ است که مقدار دقت ۸۴,۷۱ درصد را به آن می‌دهد. تمام پیش‌بینی‌های صحیح در قطرهای هر ماتریس قرار دارند. در ستون اول (کلاس ۱ از ۴ کلاس، دانشجویان عالی) از ۱۰۴ دانشجوی این کلاس، طبقه‌بندی‌کننده تعداد ۸۱ دانشجوی عالی را به‌درستی پیش‌بینی کرده است. فراخوان ٪۷۷,۸۸ (یعنی ۸۱/۱۰۴) و صحت برای این کلاس ٪۷۸,۶۴ است. در ستون چهارم (کلاس ۴ از ۴ کلاس، دانشجویان ضعیف) از ۴۲۰ دانشجوی این کلاس، طبقه‌بندی‌کننده به‌درستی تعداد ۳۶۱ دانشجو را به‌عنوان دانشجویان ضعیف پیش‌بینی کرده است. فراخوان ٪۸۵,۹۵ (یعنی ۳۶۱/۴۲۰) و صحت برای این کلاس ٪۸۹,۳۶ است؛ بنابراین مدل DTC4.5-GI&OSE در پیش‌بینی عملکرد دانشجویان ضعیف نسبت به دانشجویان موفق در حالت چهار کلاسه بهتر عمل کرده است. بر اساس ماتریس سردرگمی میزان تشخیص دانشجویان عالی ٪۷۷/۸۸، خوب ٪۸۵/۲۶، متوسط ٪۸۴/۵۹ و ضعیف ٪۸۵/۹۵ است.

ویژگی‌های مهم در حالت چهار کلاسه

اولویت‌بندی متغیرهای پیش‌بین بر اساس وزن آن‌ها در حالت چهار کلاسه در الگوریتم درخت تصمیم 5۴C. (با شاخص جینی و با انتخاب بهینه (تکاملی)) به‌صورت زیر می‌باشد:

معدل دیپلم: ۰,۲۶۲

معدل ترم ۱: ۰,۲۰۱

معدل ترم ۲: ۰,۱۹۷

تعداد ترم ممتازی:۰,۱۲۲

تعداد مشروطی: ۰,۱۱۴

سال ورود: ۰,۱۰۴

چنانچه مشاهده می‌شود معدل دیپلم بیشترین تأثیر را در روند پیش‌بینی دارد. در این مدل تمامی ویژگی‌های جمعیت شناختی تأثیر اندکی در پیش‌بینی عملکرد داشتند.

نتایج اجرای مدل پیش‌بینی عملکرد دانشجویان در حالت چهار کلاسه

نتایج اجرای مدل پیش‌بینی در جدول ۸ نشان می‌دهد که درخت تصمیم قادر به پیش‌بینی دقیق عملکرد تحصیلی دانشجویان بوده است.

جدول ۸. نتایج اجرای مدل **DT C4.5-GI&OSE** در حالت چهار کلاسه

Row No.	شماره دانشجویی	معدل کل طبقه بند...	prediction(مع...	confidence(1)	confidence(2)	confidence(3)	confidence(4)
1	932106010	3	3	0	0	1	0
2	982174021	3	4	0	0	0	1
3	972106010	1	1	1	0	0	0
4	902161006	3	3	0	0	1	0
5	962155025	2	3	0	0.455	0.545	0
6	992151020	3	3	0	0.071	0.857	0.071
7	992151006	2	2	0	1	0	0
8	902161013	3	3	0	0.059	0.941	0

بحث و نتیجه‌گیری

هدف این پژوهش بررسی عوامل مؤثر در پیش‌بینی عملکرد دانشجویان مقطع کارشناسی بر اساس مدل‌های طبقه‌بندی می‌باشد. عملکرد دانشجویان با معدل نهایی آن‌ها سنجیده شده‌اند و در پایان تحصیل قابل پیش‌بینی است. نتایج به‌دست‌آمده نشان می‌دهد که بین ویژگی‌های اجتماعی و تحصیلی دانشجویان با عملکرد تحصیلی آن‌ها رابطه وجود دارد.

بهترین الگوریتم در حالت چهار کلاسه، Decision Tree C4.5-GI&OSE با دقت پیش‌بینی ۸۴,۷۱ است. این مدل صحت ۸۴,۱۷ درصد و حساسیت ۸۳,۴۲ درصد را نشان داد. تکنیک DT C4.5-GI&OSE فارغ‌التحصیلی ۷۷,۸۸ درصد از دانشجویان عالی و ۸۵,۲۶ درصد از دانشجویان خوب و ۸۴,۶۹ درصد از دانشجویان متوسط و ۸۵,۹۶ درصد از دانشجویان ضعیف را به‌درستی پیش‌بینی کرد.

در روش اصلی تحقیق یعنی DT C4.5-GI&OSE مشاهده می‌شود که معدل دیپلم بیشترین تأثیر را بر روند پیش‌بینی عملکرد دانشجو دارد. در این تحقیق کل داده‌ها به روش اعتبار سنجی متقاطع K-fold به مجموعه داده‌های آموزشی و تجربی تقسیم می‌شوند. تخمین دقت الگوریتم‌ها در این روش بسیار بیشتر از روش‌های دیگر است. در این تحقیق مقدار k برابر با ۱۰ در نظر گرفته شده است. در پژوهش حاضر، مشخص شد که عامل تحصیلی تأثیر بیشتری بر عملکرد تحصیلی دانشجویان دارد، درحالی‌که عوامل جمعیت‌شناختی تأثیر کمتری داشتند. این نتیجه با نتایج مطالعات بوتو (Bhutto et al., 2020) مطابقت دارد. در پژوهش وحید (۲۰۲۰)، اطلاعات جمعیت شناختی دانشجویان بر عملکرد آن‌ها تأثیر بسزایی دارد(Waheed et al., 2020). این داده‌ها همیشه ایده درستی از پیشگیری از شکست را ارائه نمی‌دهند (Bernacki et al., 2020) .(Yağcı, 2022). دابهد و همکاران (۲۰۲۱) همچنین نشان دادند که بین ویژگی‌های اجتماعی و تحصیلی دانشجویان و عملکرد تحصیلی رابطه وجود دارد (Dabhade et al., 2021). در تحقیق کروز و همکاران (۲۰۲۰) پیش‌بینی نمرات نهایی دانشجویان با پارامترهای نمره نیم ترم، هیئت‌علمی و گروه آموزشی انجام شد (Cruz-Jesus et al., 2020). کاستا مندز و همکاران (۲۰۲۱) و کروز و همکاران (۲۰۲۰) عملکرد دانشجو را بر اساس درآمد، سن، اشتغال، سطح فرهنگی، محل سکونت و اطلاعات اجتماعی-اقتصادی پیش‌بینی کردند (Costa-Mendes et al., 2021; Cruz-Jesus et al., 2020). در مدل تحقیق حاضر معدل دیپلم بیشترین تأثیر را در روند پیش‌بینی دارد. ویژگی‌های معدل دیپلم، معدل ترم ۱ و معدل ترم ۲ ،تعداد ترم ممتازی،تعداد مشروطی و سال ورود به ترتیب در رتبه بعدی برای پیش‌بینی عملکرد دانشجو قرار دارند. در پژوهش خسانه (۱۳۹۶)، نیز نتایج نشان داد که معدل ترم اول در تمامی روش‌های انتخاب ویژگی در سطح بالاتری قرار دارد (Khasanah, 2017). در تحقیق حاضر الگوریتم‌های DT و RF نتایج دقیق‌تری در پیش‌بینی نمرات پیشرفت تحصیلی دانشجویان دارند و الگوریتم NB کمترین دقت طبقه‌بندی را دارد. به طور مشابه در تحقیق نگه و پیتر(۲۰۰۷)، عملکرد درخت تصمیم دقیق‌تر از شبکه‌های بیزی

بود که برای شناسایی دانشجویان ضعیف و دانشجویان خوب مفید بود (Nghe et al., 2007). در مقاله پریام و همکاران (۲۰۱۳)، درخت تصمیم C4.5 نسبت به الگوریتم‌های CART و ID3 برای مجموعه داده‌های کوچک دقت و عملکرد بهتری داشت (Priyam et al., 2013). رحمان و همکاران (۲۰۱۷) همچون تحقیق حاضر تکنیک‌های انتخاب ویژگی را برای بهبود دقت طبقه‌بندی عملکرد تحصیلی دانشجویان پیشنهاد می‌کند (Rahman et al., 2017). در پژوهش حاضر، معیار سنجش عملکرد تحصیلی دانشجویان، معدل نهایی آن‌ها در مقطع کارشناسی است. موسو و همکاران (۲۰۲۰) و پانلومجیک و راچبور(۲۰۱۵) نیز وضعیت تحصیلی آینده دانشجویان را بر اساس معدل نهایی در زمان فارغ‌التحصیلی پیش‌بینی کرده‌اند(Musso et al., 2020) (Punlumjeak & Rachburee, 2015) . استفاده از مدل‌های یادگیری ماشین به‌عنوان یک ابزار پشتیبانی تصمیم‌گیری، سطح علمی دانشجویان را بهبود می‌بخشد. با استفاده از این الگوها، مدیران آموزشی این امکان را دارند که با اعمال سیاست‌های آموزشی جدید و ارائه توصیه‌های لازم به دانشجویان از رسیدن آن‌ها به وضعیت بحرانی و افت تحصیلی جلوگیری کنند و از تعداد دانشجویان بالقوه ناموفق و ترک تحصیل بکاهند. محدودیت‌هایی همچون جمع‌آوری اطلاعات، وجود داده‌های پرت و احتیاط در تعمیم یافته‌ها در اجرای روش تحقیق پیشنهادی وجود دارد. این مطالعه در مقطع کارشناسی انجام شده است که می‌تواند در تحقیقات آتی برای مقطع کارشناسی ارشد و دکتری مورداستفاده قرار گیرد. همچنین پیشنهاد می‌شود پژوهشگران آینده با استفاده از نتایج ارزیابی‌های اساتید به بررسی و تجزیه‌وتحلیل این‌گونه پژوهش‌ها بپردازند تا داده‌های علمی و اخلاقی جدید مرتبط با هیئت‌علمی مانند شاخص‌های شایستگی، صلاحیت‌های آموزشی و معیارهای استخدامی را کشف کنند. در تحقیقات آتی، می‌توان به بررسی و کشف سایر عوامل تعیین‌کننده مانند علاقه به رشته تحصیلی؛ مهارت‌ها؛ سابقه کار یا اشتغال؛ اهداف شغلی؛ وضعیت فرهنگی؛ سرگرمی‌ها؛ درآمد خانواده؛ فضای علمی؛ معدل پیش‌دانشگاهی؛ زمان صرف شده در رسانه‌های اجتماعی؛ رفتارهای یادگیرنده؛ مشارکت رفتاری دانشجویان در فعالیت‌های تحصیلی، تعامل با

معلمان، مشارکت فوق‌برنامه دانشجویان پرداخت. همچنین در تحقیقات آتی، با گسترش مدل‌های یادگیری ماشین و استفاده از پارامترهای جدید، می‌توان به نتایج مطمئن‌تر و دقیق‌تری دست یافت.

سپاسگزاری

مقاله حاضر مستخرج از پایان‌نامه دکترای مدیریت فناوری اطلاعات بوده و در اینجا فرصتی است که از داور محترم و مشاور محترم پایان‌نامه مربوطه و همچنین داوران گرامی مجله مطالعات مدیریت کسب‌وکار هوشمند که با بیان نظرات ارزشمند خود به هر چه بهتر شدن این پژوهش کمک شایانی نمودند، تشکر و قدردانی گردد.

تعارض منافع

تعارض منافع ندارم.

ORCID

Mozhdeh Salari **https://orcid.org/0009-0005-3249-9752**

Reza Radfar **https://orcid.org/0000-0002-3951-9905**

Mahdi Faghihi **https://orcid.org/0009-0006-7473-1991**

منابع

نجفی محمود.، افضلی، مهدی.، مرادی، محمود. (۱۴۰۰). کاربرد داده‌کاوی آموزشی جهت شناسایی عوامل مؤثر بر افت تحصیلی دانش آموزان. *فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند.*

سلطانی، ستاره.، جاودانی گندمانی، تقی. (۱۴۰۰). مقایسه تحلیلی عملکرد الگوریتم‌های داده‌کاوی در پیش‌بینی پیشرفت تحصیلی دانشجویان. دومین کنفرانس ملی آخرین دستاوردهای مهندسی داده و دانش و محاسبات نرم

رئیسی وانانی، سینا.، رئیسی وانانی، ایمان.، تقوی فرد، محمدتقی. (۱۳۹۹). مدلی برای بخش بندی یادگیرندگان و بهبود عملکرد آموزشی با استفاده از الگوریتم‌های داده‌کاوی. نشریه علمی *مطالعات مدیریت کسب‌وکار هوشمند*، سال نهم، شماره ۳۳ -۳۸- ۵

فدوی رودسری، آزاده.، صالحی، کیوان.، خدایی، ابراهیم.، مقدم زاده، علی.، جوادیپور، محمّد. (۱۳۹۸). مدل شبکه بیزی عوامل مرتبط با افت تحصیلی دانشجویان دانشگاه تهران، *مجله علوم روان‌شناختی*، ۱۸ (۷۶): ۴۲۹-۴۱۷

خسروی، هادی .، شفیعی، ریحانه.(۱۳۹۶). پیش بینی عملکرد دانش آموزان با استفاده از داده کاوی، دانشکده مهندسی کامپیوتر.

شاملو، رسول.، امید، منوچهر.، امین فر، فائزه. (۱۳۹۳). بررسی پیش‌بینی رفتار آموزشی دانشجویان با رویکرد داده‌کاوی در مؤسسات آموزش عالی (مطالعه موردی دانشگاه آزاد واحد بویین‌زهرا). گروه صنایع، دانشگاه آزاد اسلامی، واحد قزوین، دانشکده صنایع و مکانیک.

اسدی ورمله، پرویز.، احمدی، هادی.، حسنی پیرمحمدی، حشمت الله. (۱۳۹۳). بررسی علل افت تحصیلی دانش آموزان سال اول دبیرستان با استفاده از تکنیکهای داده کاوی"، دومین کنفرانس بین المللی دستاوردهای نوین در علوم مهندسی و پایه، اردبیل.

ایرجی، اعظم.، مینایی، بهروز.، شکورنیاز ونوس. (۱۳۹۲). به‌کارگیری فن‌آوری داده‌کاوی به‌منظور آسیب‌شناسی افت تحصیلی هنرجویان هنرستانی و استخراج نمایه‌ساز توصیفی در ارائه تمایز دانش آموزان ضعیف و ممتاز تهران، دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران.

یقینی مسعود.، اکبری، امین.، شریفی، سید محمدمهدی. (۱۳۸۷). پیش‌بینی وضعیت تحصیلی

دانشجویان با استفاده از تکنیک‌های داده‌کاوی، دومین کنفرانس داده‌کاوی ایران، تهران.

References

Abdelmagid, A., & Qahmash, A. (2023). Utilizing the Educational Data Mining Techniques" Orange Technology" for Detecting Patterns and Predicting Academic Performance of University Students. *Inf. Sci. Lett*, *12*, 1415-1431 .

Adejo, O. W., & Connolly, T. (201 .(۸Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education*.

Ahmad, Z., & Shahzadi, E. (2018). Prediction of Students' Academic Performance Using Artificial Neural Network .*Bulletin of Education and Research*, *40*(3), 157-164.

Ahmed, N. S., & Sadiq, M. H. (2018). Clarify of the random forest algorithm in an educational field. 2018 international conference on advanced science and engineering (ICOASE),

Ajibade, S.-S. M., Ahmad, N. B., & Shamsuddin, S. M. (2019). An heuristic feature selection algorithm to evaluate academic performance of students. 2019 IEEE 10th Control and System Graduate Research Colloquium (ICSGRC),

Akar, Ö., & Güngör, O. (2012). Classification of multispectral images using Random Forest algorithm. *Journal of Geodesy and Geoinformation*, *1*(2), 105-112.

Al-Emran, M., Malik, S. I., & Al-Kabi, M. N. (2020). A survey of Internet of Things (IoT) in education: Opportunities and challenges. *Toward social internet of things (SIoT): enabling technologies, architectures and applications*, 197-209.

Alboaneen, D., Almelihi, M., Alsubaie, R., Alghamdi, R., Alshehri, L., & Alharthi, R. (2022). Development of a Web-Based Prediction System for Students' Academic Performance. *Data*, *7*(2), 21.

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, *37*, 13-49.

Alghamdi, A. S., & Rahman, A. (2023). Data mining approach to predict success of secondary school students: A Saudi Arabian case study. *Education Sciences*, *13*(3), 293.

Alsalman, Y. S., Halemah, N. K. A., AlNagi, E. S., & Salameh, W. (2019). Using decision tree and artificial neural network to predict students academic performance. 2019 10th International Conference on Information and Communication Systems (ICICS),

Alshanqiti, A., & Namoun, A. (2020). Predicting student performance and

its influential factors using hybrid regression and multi-label classification. *IEEE Access*, *8*, 203827-203844.

Ampadu, Y. B. (2023). Handling Big Data in Education: A Review of Educational Data Mining Techniques for Specific Educational Problems .*AI, Computer Science and Robotics Technology*.

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015). Preprocessing and analyzing educational data set using X-API for improving student's performance. 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT),

Anguita, D., Ghio, A., Ridella, S., & Sterpi, D. (2009). K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines. DMIN,

Arcinas, M. M., Sajja, G. S., Asif, S., Gour, S., Okoronkwo, E., & Naved , M. (2021). Role of Data Mining in Education for Improving Students Performance for Social Change. *Turkish Journal of Physiotherapy and Rehabilitation*, *32*(3), 6519-6526.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, *4*, 40-79.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177-194.

Bahadir, E. (2016). Using Neural Network and Logistic Regression Analysis to Predict Prospective Mathematics Teachers' Academic Success upon Entering Graduate Education. *Educational Sciences: Theory and Practice*, *16*(3), 943-964.

Ballabio, D., Grisoni, F., & Todeschini, R. (2018). Multivariate comparison of classification performance measures. *Chemometrics and Intelligent Laboratory Systems*, *174*, 33-44.

Banik, P., & Kumar, B. (2019). Impact of information literacy skill on students' academic performance in Bangladesh. *International Journal of European Studies*, *3*(1), 27-33.

Batirovna, S. B. (2023). EDUCATIONAL DATA MINING AND LEARNING ANALYTICS.

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.-Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance :A survey study. *Education and Information Technologies*, *28*(1), 905-971.

Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. *Computers & Education*, *158*, 103999.

Bhutto, E .S., Siddiqui, I. F., Arain, Q. A., & Anwar, M. (2020). Predicting students' academic performance through supervised machine learning.

2020 International Conference on Information Science and Communication Technology (ICISCT),

Breiman, L. (2001). Random forests Mach Learn 45 (1): 5–32. In.

Capuano, N., & Toti, D. (2019). Experimentation of a smart learning system for law based on knowledge discovery and cognitive computing. *Computers in Human Behavior*, *92*, 459-467.

Chamorro-Premuzic, T., & Furnham, A .(۲۰۰۹) .Mainly Openness: The relationship between the Big Five personality traits and learning approaches. *Learning and individual Differences*, *19*(4), 524-529.

Chen, S., & Lin, X. (2023). Application of Decision Tree Algorithm in Educational Data Mining .*Curriculum and Teaching Methodology*, *6*(8), 120-127.

Costa-Mendes, R., Oliveira, T., Castelli, M., & Cruz-Jesus, F. (2021). A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach. *Education and Information Technologies*, *26*(2), 1527-1547.

Costa, E., Lorena, A., Carvalho, A., & Freitas, A. (2007). A review of performance evaluation measures for hierarchical classifiers. Evaluation methods for machine learning II: Papers from the AAAI-2007 workshop,

Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. *Heliyon*, *6*(6), e04081.

Cuevas, R., Ntoumanis, N., Fernandez-Bustos, J. G., & Bartholomew, K. (2018). Does teacher evaluation based on student performance predict motivation, well-being, and ill-being? *Journal of school psychology*, *68*, 154-162.

Dabhade, P., Agarwal, R., Alameen, K., Fathima ,A., Sridharan, R., & Gopakumar, G. (2021). Educational data mining for predicting students' academic performance using machine learning algorithms. *Materials Today: Proceedings*, *47*, 5260-5267.

Debang, M., & Hassan, B. U. (2023). Educational Data Mining :Prospects and Applications.

Ding, W., Jing, X., Yan, Z., & Yang, L. T. (2019). A survey on data fusion in internet of things: Towards secure and privacy-preserving fusion. *Information Fusion*, *51*, 129-144.

El Aissaoui, O., El Alami El Madani, Y., Oughdir, L., Dakkak, A., & El Allioui, Y. (2019). A multiple linear regression-based approach to predict student performance. International Conference on Advanced Intelligent Systems for Sustainable Development,

Fernandes, E., Carvalho, R., Holanda, M., & Van Erven, G. (2017).

Educational data mining: Discovery standards of academic performance by students in public high schools in the federal district of brazil. World conference on information systems and technologies,

Galla, B. M., Wood, J. J., Tsukayama ,E., Har, K., Chiu, A. W., & Langer, D. A. (2014). A longitudinal multilevel model analysis of the within-person and between-person effect of effortful engagement and academic self-efficacy on academic performance. *Journal of School Psychology*, *52*(3), 295-308.

Garcia, J. D., & Skrita, A. (2019). Predicting academic performance based on students' family environment: evidence for Colombia using classification trees. *Psychology, Society & Education*, *11*(3), 299-311.

Han, J., & Kamber, M. (2006). Data mining :concepts and techniques, 2nd. *University of Illinois at Urbana Champaign: Morgan Kaufmann*.

Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, *5*(4), 83-124.

Harwati, A. (2014). AP, & Wulandari, FA (2014). Mapping student's performance based on data mining approach. The 2014 International Conference on Agro-industry (ICoA): Competitive and Sustainable Agroindustry for Human Welfare,

Hasani, A. A., & Bazrafshan ,M. (2018). Analyzing Students' Educational Information to Evaluate Their Success via Using Data Mining Method (Case Study: Faculty of Management and Industrial Engineering, Shahrood University of Technology). *Journal of Management and Planning In Educational System*, *11*(2), 187-208.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). *Predicting academic performance: a systematic literature review* Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, Larnaca, Cyprus. https://doi.org/10.1145/3293881.3295783

Hoffait, A.-S., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, *101*, 1-11. https://doi.org/https://doi.org/10.1016/j.dss.2017.05.003

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447-459.

Jayaprakash, S., Krishnan, S., & Jaiganesh, V. (2020). Predicting students academic performance using an improved random forest classifier. 2020 international conference on emerging smart computing and informatics (ESCI),

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, *13*(1), 61-72.

Karo, I., Fajari, M., Fadhilah, N & ,.Wardani, W. (2022). Benchmarking Naïve Bayes and ID3 Algorithm for Prediction Student Scholarship. IOP Conference Series: Materials Science and Engineering,

Kaur, K., & Dahiya, O. (2023). Role of Educational Data Mining and Learning Analytics Techniques Used for Predictive Modeling. 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM),

Khasanah, A. U. (2017). A comparative study to predict student's performance using educational data mining techniques. IOP Conference Series: Materials Science and Engineering,

Kulkarni, V. Y. (2014). Effective learning and classification using random forest algorithm.

Kumar, A. D., Selvam, R. P., & Kumar, K. S. (2018). Review on prediction algorithms in educational data mining. *International Journal of Pure and Applied Mathematics*, *118*(8), 531-537.

Kumar, M., & Salal, Y. K. (2019). Systematic review of predicting student's performance in academics. *Int. J. of Engineering and Advanced Technology*, *8*(3), 54-61.

Lei, C., & Li ,K. F. (2015). Academic performance predictors. 2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops,

Leung, K. M. (2007). Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, *2007*, 123-156.

Lever, J., Krzywinski, M., & Altman, N. (2016). Points of significance: model selection and overfitting. *Nature methods*, *13*(9), 703-705.

Machado, M. R., Karray, S., & de Sousa, I. T. (2019). LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry. 2019 14th International Conference on Computer Science & Education (ICCSE),

Manoharan, R., Stalin, M. S., & Loganathan, G. B. (2023). Ensemble Model for Educational Data Mining Based on Synthetic Minority Oversampling Technique.

Marjan, M. A., Uddin, M. P., & Ibn Afjal, M. (2023). An Educational Data Mining System For Predicting And Enhancing Tertiary Students' Programming Skill. *The Computer Journal*, *66* (5),1101-1083.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.

Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. *International journal of modern education & computer science*, *8*.(۱۱)

Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: a machine-learning approach. *Higher Education*, *80*(5), 875-894.

Mwadulo ,M. W. (2016). A review on feature selection methods for classification tasks.

Nghe, N. T., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. 2007 37th annual frontiers in education conference-global engineering: knowledge without borders, opportunities without passports,

Nijhawan, V. K., Madan, M., & Dave, M. (2017). The Analytical Comparison of ID3 and C4. 5 using WEKA. *International Journal of Computer Applications*, *167*(11), 1-4.

Nisbet, R ,.Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Academic press.

Ogunde, A. O., & Ajibade, D. A. (2014). A data mining system for predicting university students' graduation grades using ID3 decision tree algorithm. *Journal of Computer Science and Information Technology*, *2*(1), 21-46.

Oreski, D., Pihir, I., & Konecki, M. (2017). Crisp-DM process model in educational setting. *Economic and Social Development: Book of Proceedings*, 19-28.

Oskouei, R. J., & Askari ,M. (2014). Predicting academic performance with applying data mining techniques (generalizing the results of two different case studies). *Computer Engineering and Applications Journal*, *3*(2), 79-88.

Osmanbegović, E., Suljić, M., & Agić, H. (2014). Determining dominant factor for students performance prediction by using data mining classification algorithms. *Tranzicija*, *16*(34), 147-158.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1), 217-222.

Pandey, U. K., & Pal, S. (2011). Data Mining: A prediction of performer or underperformer using classification. *arXiv preprint arXiv:*1104.4163.

Phan, H. P., & Ngu, B. H. (2014). An empirical analysis of students' learning and achievements :A motivational approach. *Education Journal*, *3*(4), 203-216.

Priyam, A., Abhijeeta, G., Rathee, A., & Srivastava, S. (2013). Comparative analysis of decision tree classification algorithms. *International*

Journal of current engineering and technology, *3*(2), 334-337.

Punlumjeak, W., & Rachburee, N. (2015, 29-30 Oct. 2015). A comparative study of feature selection techniques for classify student performance. 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE),

Putpuek, N., Rojanaprasert, N., Atchariyachanvanich, K., & Thamrongthanyawong, T. (2018). Comparative study of prediction models for final GPA score: a case study of Rajabhat Rajanagarindra University. 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS),

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

Rahman, L., Setiawan, N. A., & Permanasari, A. E. (2017). Feature selection methods in improving accuracy of classifying students' academic performance. 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE),

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. *arXiv preprint arXiv:1002.1144*.

Raosoft. (2004). Raosoft, 2004. Sample Size Calculator. Available at: http://www.raosoft.com/samplesize.html.

Rebai, S., Yahia, F. B., & Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in Tunisia. *Socio-Economic Planning Sciences*, *70*, 100724.

Rizvi, S., Rienties, B., & Khoja, S. A. (2019). The role of demographics in online learning; A decision tree based approach. *Computers & Education*, *137*, 32-47.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601-618.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey [https://doi.org/10.1002/widm.1355]. *WIREs Data Mining and Knowledge Discovery*, *10*(3), e1355 . https://doi.org/https://doi.org/10.1002/widm.1355

Ross Quinlan, J. (1993). C4. 5: programs for machine learning. *Mach. Learn*, *16*(3), 235-240.

Rostami, M., Ayat, S., Saghari, F., & Yaghoobi, F. (2015). Applying fuzzy clustering to assess and anticipate students' educational progress in learning environments. *Technology of education*, 10(1), 23-36.

Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia. In: Pearson Education Limited London, UK.:

Saa, A. A. (2016). Educational data mining & students' performance prediction. *International Journal of Advanced Computer Science and Applications*, *7*(5), 212-220.

Salal, Y., & Abdullaev, S. (2019). Optimization of classifiers ensemble Construction: Case study of Educational data Mining. *Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника*, 19(4).

Sasikala, S., alias Balamurugan, S. A., & Geetha, S. (2016). Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set. *Applied Computing and Informatics*, *12*(2), 117-127.

Seif, A. (2016). Modern educational psychology: Psychology of learning and education. *Tehran: Doran Publishing. [In Persian]*.

Shaikh, A., Mahoto, N., Khuhawar, F & ,.Memon, M. (2015). Performance evaluation of classification methods for heart disease dataset. *Sindh University Research Journal-SURJ (Science Series), 47(3)*.

Simundic, A.-M. (2008). Confidence interval. *Biochemia Medica, 18*(2), 154-161.

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524.

Sokkhey, P., & Okazaki, T. (2020). Hybrid machine learning algorithms for predicting academic performance. *Int. J. Adv. Comput. Sci. Appl*, 11(1), 32-41.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, *45*(4), 427-437.

Surjeet, K., & Saurabh, P. (2012). Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification" WCSIT. In: ISSN.

Tan, P., Steinbach, M., & Kumar, V. (2013). Introduction to Data Mining: Pearson New International Edition (English Edition). In: Pearson Education Limited, Harlow ,ESX, UK.

Ting, K. M. (2017). Confusion matrix. *Encyclopedia of machine learning and data mining*, 260.

Turabieh, H. (2019, 9-11 Oct. 2019). Hybrid Machine Learning Classifiers to Predict Student Performance. 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS),

Umer, R., Susnjak, T., Mathrani, A., & Suriadi, S. (2017). On predicting academic performance with process mining in learning analytics. *Journal of Research in Innovative Teaching & Learning*, *10*(2), 160-176. https://doi.org/10.1108/JRIT-09-2017-0022

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98-110.

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human behavior*, 104, 106189.

Widyastuti, T., Kurniawan, A., & Chandra, N. P. (2017). Coping Strategies on Students After Experiencing Academic Failure: An Indigenous Study in Javanese Context. *Work Pap Ser*, *3*, 22-26.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining,

Wood, J. M. (2007). Understanding and computing Cohen's Kappa: A tutorial. *WebPsychEmpiricist. Retrieved October*, 3(2007), 145-160.

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, *98*, 166-173.

Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. *arXiv preprint arXiv:1202.4815*.

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 1-19.

Yehuala, M. A. (2015). Application of data mining techniques for student success and failure prediction (The case of Debre Markos university). *International journal of scientific & technology research*, 4(4), 91-94.

Yulianto, L. D., Triayudi, A., & Sholihati, I. D. (2020). Implementation Educational Data Mining For Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4. 5: Implementation Educational Data Mining For Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4. 5. *Jurnal Mantik*, 4(1), 441-451.

Zimmermann, J., Brodersen, K. H., Heinimann, H. R., & Buhmann, J. M. (2015). A Model-Based Approach to Predicting Graduate-Level

Performance Using Indicators of Undergraduate-Level Performance. *Journal of Educational Data Mining*, 7(3), 151-176.

Zulfiker, M. S., Kabir, N., Biswas, A., Chakraborty, P., & Rahman, M. M. (2020). Predicting students' performance of the private universities of Bangladesh using machine learning approaches. *International Journal of Advanced Computer Science and Applications*, 11(3), 672-679.

References [In Persian]

Masoud's certainty; Akbari, Amin; Sharifi, Seyyed Mohammad Mahdi (2007). Predicting the educational status of students using data mining techniques, *the second data mining conference of Iran*, Tehran.

Najafi Mahmoud; Afzali, Mehdi; Moradi, Mahmoud (2021). The use of educational data mining to identify the factors affecting the academic drop of students. *Quarterly journal of intelligent multimedia processing and communication systems*.

Soltani, Star; Javadani Gadmani, Taghi (2021). Analytical comparison of performance of data mining algorithms in predicting academic progress of students. *The second national conference on the latest achievements in data engineering, knowledge and soft computing*.

Ghodoosi, Mir Saeedi, kosha. Predicting and analyzing student performance using data mining techniques to improve academic performance. Department of Industrial Engineering, Ferdowsi University of Mashhad.

Fadavi Rudsari Azadeh, Salehi Keyvan, Khodayi Ebrahim, Moghadamzadeh Ali, Javadipour Mohammad (2018). Bayesian network model of factors related to academic failure of Tehran University students, *Journal of Psychological Sciences*; 18(76): 429-417.

Khosravi, Hadi; Shafii, Reyhane. (2016). Predicting student performance using data mining, Faculty of Computer Engineering.

M. Zahedi Fard, A. Attarzadeh, and H. Pazakhzadeh (2014). Prediction of high school students' performance with data mining techniques.

Shamlu, Rasul; Hope, Manouchehr; Amin Far, Faeze (2014). Investigating the prediction of students' educational behavior with a data mining approach in higher education institutions (case study of Azad University, Boyin Zahra Branch). Department of Industries, Islamic Azad University, Qazvin Branch, Faculty of Industries and Mechanics.

Asadi Vermele, Parviz; Ahmadi, Hadi; Hosni Pirmohammadi, Heshmatullah. (2013). Investigating the causes of academic failure of first year high school students using data mining techniques", *Second International*

Conference on New Achievements in Engineering and Basic Sciences, Ardabil.

استناد به این مقاله: سالاری، مژده.، رادفر، رضا.، فقیهی، مهدی. (۱۴۰۳). پیش‌بینی عملکرد دانشجویان با استفاده از الگوریتم‌های یادگیری ماشین و داده‌کاوی آموزشی (مطالعه موردی دانشگاه شاهد)، *مطالعات مدیریت کسب وکار هوشمند*، ۱۲(۴۷)، ۳۱۵-۳۶۶. DOI: 10.22054/ims.2023.75523.2375