第 48 卷 第 9 期
2021 年 9 月

湖南大学学报（自然科学版）
Journal of Hunan University（Natural Sciences）

Vol. 48. No. 9.
September 2021

Early Prediction of Students' Academic Achievement: Categorical Data from Fully Online Learning on Machine-Learning Classification Algorithms

Tuti Purwoningsih[1,2]*, Harry B. Santoso[1], Kristanti A. Puspitasari[2], Zainal A. Hasibuan[3]

[1] Faculty of Computer Science, Universitas Indonesia, Depok, West Java, Indonesia

[2] Faculty of Teacher Training and Education, Universitas Terbuka, South Tangerang, Banten, Indonesia

[3] Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Central Java, Indonesia

Abstract: Several challenges related to predicting students' academic achievement in fully online learning are defining the dataset used as a predictor. Accordingly, in this study, we define the dataset as categorical data from student demographic profile data, activities, and learning habits of Fully Online Learning students at the Universitas Terbuka (UT). This study's main objective is to predict early academic achievement of fully online learning students using category data as features and to identify relevant important features/predictors. We apply several machine learning (ML) classification algorithms to make early predictions of student academic achievement. This study uses 75,136,349 UT-LMS log data, combined with the demographic profile of 101,617 undergraduate students in fully online learning. Datasets were converted into categorical data to minimize noise arising from large datasets. This study found that the influence factors to student's academic achievement are online learning activities related to access day, study time, and student profession profile. Most students were accessing the UT-LMS on Monday, and the time was in the evening. The evaluations and experiments showed that the random forest algorithm could achieve 85.03% accuracy for the balancing dataset with SMOTE, encoding ordinal data with a label encoder and nominal data with a one-hot encoder. The findings can assist lecturers in designing instructional strategies to improve the student's academic achievement success. Furthermore, the principal novel contribution of this study is how to explore the UT-LMS log data and student demographic data to define it as a categorical data set in the machine-learning classification algorithms. The process of categorizing datasets in this study is more of an art than a science, but this research can form the basis for similar research with other scientific principles analysis. So that similar research after this produces a more optimal accuracy.

Keywords: learning management system, fully online learning, academic achievement, machine learning.

体能对女体育教师职业倦怠及心理健康的影响

摘要：與預測學生在完全在線學習中的學業成績相關的幾個挑戰是定義用作預測器的數據集。因此，在本研究中，我們將數據集定義為來自特布卡大學完全在線學習學生的學生人口統計資料、活動和學習習慣的分類數據。本研究的主要目標是使用類別數據作為特徵來預測完全在線學習的學生的早期學業成績，並確定相關的重要特徵/預測因素。我們應用了幾種機器學習分類算法來對學生的學業成績進行早期預測。本研究使用 75,136,349 特布卡大學-學習管理系統日誌數據，結合 101,617 名完全在線學習的本科生的人口統計資料。數據集被轉換為分類數據，以最大限度地減少大型數據集產生的噪音。本研究發現，影響學生學業成績的因素是與訪問天數、學習時間和學生職業概況相關的在線學習活動。大多數學生在周一訪問特布卡大學-學習管理系統，時間是晚上。評估和實驗表明，隨機森林算法對於使用合成少數過採樣技術的平衡數據集可以達到 85.03% 的準確率，使用標籤編碼器編碼序數數據，

Received: June 1, 2021 / Revised: June 6, 2021 / Accepted: August 11, 2021 / Published: September 30, 2021

About the authors: Tuti Purwoningsih, Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia; Faculty of Teacher Training and Education, Universitas Terbuka, South Tangerang, Indonesia; Harry B. Santoso, Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia; Kristanti A. Puspitasari, Faculty of Teacher Training and Education, Universitas Terbuka, South Tangerang, Indonesia; Zainal A. Hasibuan, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

Corresponding author Tuti Purwoningsih, tutipurwoningsih@gmail.com

使用單熱編碼器編碼標稱數據。研究結果可以幫助講師設計教學策略，以提高學生的學業成就。此外，本研究的主要新貢獻是如何探索特布卡大學-學習管理系統日誌數據和學生人口統計數據，以將其定義為機器學習分類算法中的分類數據集。本研究中對數據集進行分類的過程與其說是科學，不如說是一門藝術，但這項研究可以與其他科學原理分析形成類似研究的基礎。因此，在此之後的類似研究會產生更佳的準確性。

关键词：學習管理系統，完全在線學習，學術成就，機器學習。

1. Introduction

Open and Distance Learning (ODL) provides alternative learning and educational opportunities that citizens can access without geographical, physical, social, and economic constraints. Along with the development of Information and Communication Technology (ICT), ODL can apply technology that allows students to learn across time and space according to students' flexibility [1]. The application of ICT in education is known as online learning, which is usually via the internet, so that the characteristics of students in online learning are very heterogeneous. This can be seen from the diversity of students participating in online learning based on their demographic profiles. Online learning is a system that includes applying several ICTs to benefit students' learning and education anytime and anywhere. It is important to understand how students learn to determine the appropriate learning strategies through online learning in the knowledge construction process.

The online learning system provides students with more interactivity and flexibility to use online devices at any time and anywhere. On the one hand, teachers in online learning, especially learning fully online (FO), do not have complete information about the characteristics, habits, and activities of learning, as well as the progress of student academic achievement like that of teachers in the face to face (F2F) learning environment. At F2F, teachers can immediately see how students learn and can directly adjust the instructional strategies used if they feel that many students have experienced failures in the learning process. On the other hand, teachers in FO socialize students virtually, so they cannot directly adjust their instructional strategy [2].

Learning management systems (LMS) are widely used in online learning, both for blended learning and fully online learning. The LMS records all interactions the user makes on the system in a log file. Student's activity information in log files can be useful to predict the success of student's academic achievement. However, in online learning systems, teachers sometimes have difficulties measuring student engagement compared with traditional learning modes (e.g., value metrics, class attendance, and participation in discussions) because many variables are not directly available in online learning systems. Thus, investigating e-Leaning student activity becomes a challenging task.

The objective of this study is to explore the profile, learning habits, and learning activities in online learning to predict the success of student academic achievement in fully online learning. Using the ODL system, higher education institutions can plan the best instructional strategy to increase students' academic achievement. In this study, the success of student's academic achievement was measured based on the Grade Point Average (GPA) obtained by students [3]. Previous studies have shown that instructional strategies positively predict GPA [4]. On the one hand, instructional strategy training and motivation did produce a higher GPA of students and positively affected the learning outcomes of ODL students [5].

Modeling and predicting the academic achievement success of online learning students effectively based on LMS activity log data using machine learning classification algorithms are challenging tasks because different classifications will provide different predictive results in different contexts. Accordingly, we constructed a data set in this study by considering a broad exploratory data analysis on various mathematical and statistical techniques. The data set construction in this study used demographic profile data, academic data, student learning habits data, and activities related to interactions in LMS. The collected dataset is big data with quite large noise, so it needs exploratory data analysis techniques to minimize the noise. The prediction model in this study used a machine-learning classification algorithm because the type of class data was discrete. In this case, to analyze the effectiveness of the prediction model, ensemble methods for machine learning algorithms (Random Forest (RF), Decision Tree (DT), Gradient Boosting (GB), and Adaptive Boosting (AB) [6].

2. Related Works

The ODL system allows students to learn flexibly, which is not bound by time and space. UT implements the ODL system in the learning process. One of the learning modes provides by UT is fully online learning,

which in the process uses an LMS. UT student participants have a heterogeneous demographic profile because learning with the ODL system can be done anytime and anywhere. Thus, fully online participant students have a variety of learning contexts. The learning context represents the factors of the learning environment that can give meaning to the messages they receive. Considering the learning context of students will produce a systemic and systematic instructional design. The description of the context of student learning in participating in online learning can be analyzed based on the data stored in the LMS log file.

Modular Object-Oriented Dynamic Learning Environment (Moodle) is an open-source LMS platform that has been used in 251 countries by the end of 2020. Three indicators as the main function of Moodle are 1) Login activity, 2) Forum Activity, and 3) Assessment Activity. Students may use different LMS features in different ways; therefore, it is difficult to find a series of variables that consistently predict student performance in learning [7]. The researcher must use a meaningful log file size, which follows the learning theory [8].

According to the theory of self-regulation learning, the measure of time spent by students in the learning process is useful in modeling student's performance in several studies. Most of the research was carried out in the context of blended learning [7], [9], [10], [11]; therefore, student interactions with LMS features in that context are different from fully online students. The main requirement in planning fully online learning is to correctly predict early academic performance to address student weaknesses [12]. Factors predicting ODL students' success include high motivation, age, and study habits [13]. In addition, in the LMS, the interaction factor of students with the LMS, which varies with demographic factors, can affect students' performance [14]. However, the limitations of these studies are the small amount of data collected and in the context of blended learning. So, for fully online learning, a different data collection method was needed.

One approach that can be used is to predict academic achievement at the end of the semester using student's log data from the LMS. This study discusses how to construct data sets for this purpose. The collected raw data must go through preprocessing before it is ready to be used in the prediction model for the success classification of student's academic achievement. The prediction model in this study uses machine learning algorithms that have proven their ability to predict learning data [15], [16], [17].

Much literature focuses on predicting student performance in solving problems or completing courses [18]. Many machine learning techniques, such as artificial neural networks, decision trees, and probabilistic graphic models, are applied to develop predictive algorithms. Research aimed at predicting student academic performance using various performance metrics uses machine learning algorithms [19], [20], [21]. However, it is not clear which model is among the various models. Machine learning accurately predicts student performance because various authors present conflicting results regarding the accuracy of model predictions.

Overall, although the current literature provides interesting predictions in online learning, it is limited to data methods derived from the results of filling out student or teacher questionnaires on blended learning. So, the main purpose of this study is to use machine learning algorithms as a classification model in predicting academic achievement of online learning students in fully online learning based on student demographic profile data, student learning habits data, and student activity in e-Learning recorded in the LMS system.

3. Material and Method

In this study, a Jupyter Notebook was used with the Python programming language to conduct experiments because it is easy to understand and has an open-source that can develop insights on data analysis. We use various machine learning algorithms, which were applied to predict the academic achievement of online learning students based on student demographics, student learning habits, and learning activities in the LMS system. The mathematical and statistical techniques selected are suitable for attributes to the domain and categorical education. The main steps in this research use a data science approach, as shown in Fig. 1.

3.1. Dataset

Instructional at UT is based on the principle of self-regulated learning, which is an instructional process that demands students' initiative. Students can learn by studying teaching materials, studying through study groups, or by attending tutorials. The instructional mode can be done face to face (F2F), blended learning (BL), or fully online (FO). In FO mode, instructional is delivered in the form of e-Learning which can use LMS.

Online learning at UT is provided in the form of an online tutorial using the Moodle LMS platform. An online tutorial is a learning service provided by UT, held in 8 sessions for eight consecutive weeks. To participate in the online tutorial, UT students must activate the UT-LMS and fill out a form available to participate in the online tutorial. The online tutorial assessment consists of attendance scores, discussions, and assignments, where the assessment is all done online. The assessment contributes 30% to the course's final grade if the final semester exam score reaches

30% of the maximum score.

This study uses data from students who took part in the UT-LMS in 2019/2020.1. Respondents of UT-LMS participants in this study came from various regions, ages, professions, highest education, and gender, as well as various academic profiles (faculty, study program, and semester). We use student profile data from the Student Academic Information System-UT and student log data from UT-LMS.

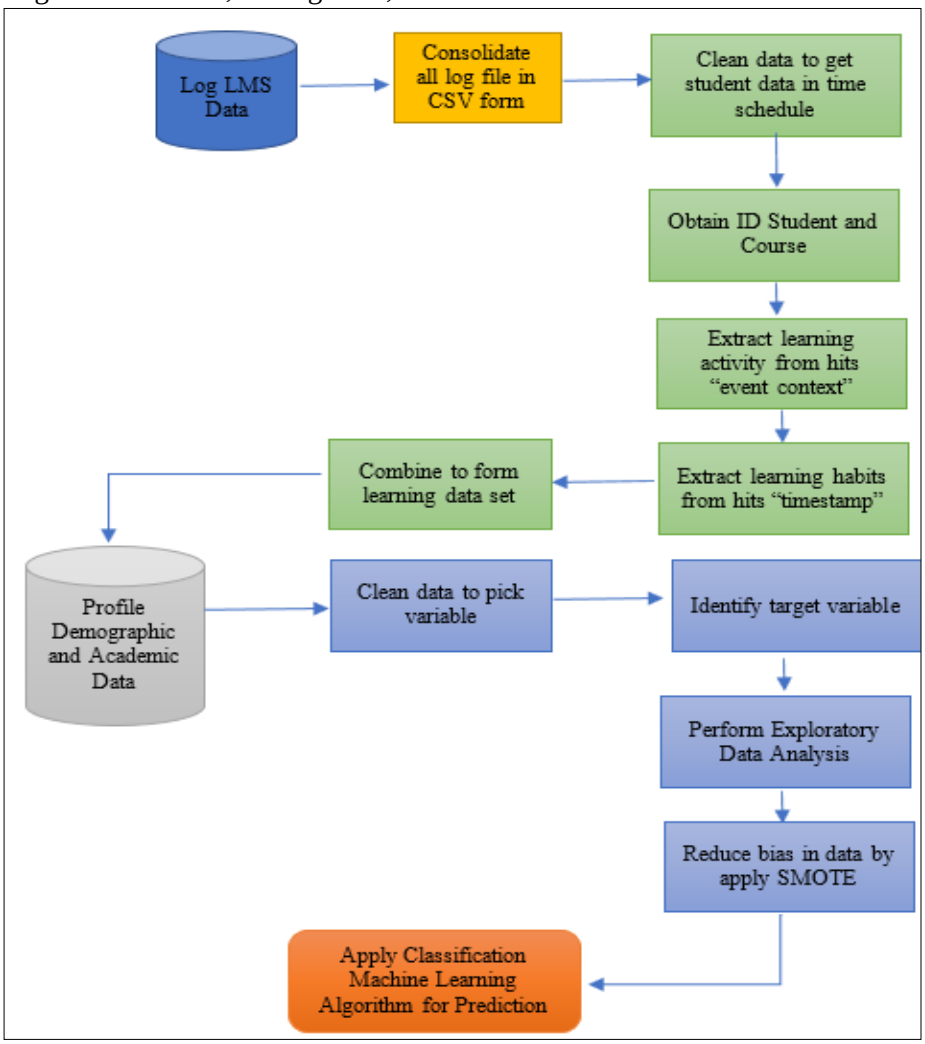

Fig. 1 Workflow for prediction of student academic achievement success with data science approach

The Moodle log file as a UT-LMS platform contains records of student activities in online learning, which are still in raw data. This raw data has not concretely demonstrated a theoretical framework that is more commonly used in learning [7]. This study seeks to generalize LMS data so that analysis can be carried out accurately, especially online learning, which is carried out fully online at universities using the ODL system.

Referring to the UT-Online Tutorial Guide, online learning courses at UT have the same structure, namely: 8 initiation materials, eight discussion activities, and three assignments (on weeks 3, 5, and 7). Students carry out online tutorial activities asynchronously so that their activities and access times to online tutorials vary widely. In general, log data can show each student's learning habits and activities in an online learning class. Statistic descriptions of the UT-LMS features used in the study are summarized in Table 1 below.

Table 1 Statistic description of log data UT-LMS 2019/2020.1

Feature	Count	Unique	Description
Time	75,136,349	4,442,260	The timestamp when the activity was recorded
User full name	75,136,349	119,107	Student's full name and ID
Affected user	75,136,349	116,334	Full name of the affected user
Event context	75,136,349	164,248	Activity Context to which the activity is subject
Component	75,136,349	27	The component of the section to which the activity is subject
Course	75,136,349	13,080	Information about Course
Event name	75,136,349	82	The name of the activity is according to the type and class of activity

Feature	Count	Unique	Description
Description	75,136,349	48,543,727	A description of the activity that describes the activity and user in Moodle
Origin	75,136,349	3	Log Origin (Client/Web Server)
IP address	75,136,349	546,750	IP Address of the device that the user uses to log into the system
Source	75,136,349	1,022	File data

The data described in Table 1 is the raw LMS log data obtained by downloading from the UT server system. The log data consists of 1,022 files that are aggregated using the glob () function in Python. The data is from 13,080 classes (class courses and UT community forum classes). The log data is unstructured because of the considerable diversity of each column. The LMS log data is extracted into features of learning habits and learning activities according to analytical needs, interpreted in a structured format as output. Student learning habits data is obtained by extracting time information from the "time" column, while student learning activity data is obtained by extracting information from the "event name" column in the raw LMS log data.

This study uses data on learning habits and activities that were relatively strong predictors in previous research [7] and adapted to the online tutorial structure consisting of material, discussion forums, and assignments. The data extraction results are stored in a file with CSV format, which is then merged with the profile data using Student_ID and Course_ID as keys. The data collected is data with large and inconsistent transactions, so certain concepts and methodologies are needed to change the data structure. Data munging is a set of concepts and methodologies for taking data from unusable and faulty forms to the structure and quality required in analytics.

The raw data collected comes from several sources and is large in number, so there needs to be a specific technique in gathering and reading this data. This study uses Microsoft Excel to manage data sets in different formats and forms. As for preprocessing, this study uses the Jupyter Notebook with the Python 3.6 programming language, Pandas, NumPy, and Matplotlib. The preprocessing data for learning habits, activity learning, and profile produce a dataset ready to be entered into a prediction model using a machine learning algorithm.

Preprocessing data to be numeric into categorical data varies between features depending on the characteristics of the data. The results of converting numeric data into categorical data produce a new dataset labeled as predictor and target attributes with detailed descriptions shown in Table 2 below.

Table 2 List and description of predictor and target of attributes

Attribute	Description
Predictor Learning Activity	
N_module_viewed	The frequency of student viewed resources (learning material) which is categorized into "low," "moderate," or "high"
N_discussion_viewed	The frequency with which students discuss the forums which are categorized as: "once," or "at no time"
N_discussion_created	The number of discussions a student creates on the forum is categorized as: "low," "moderate," or "high"
N_assignment_viewed	The frequency of students viewed the status of tasks on assignments which are categorized into "low," "moderate," or "high"
N_assignment_upload	The number of tasks that students submitted or uploaded to assignments which categorized as: "zero," "one task," " two tasks," or "three tasks"
N_assignment_created	The number of tasks that students made on assignments which categorized into "at no time," "1-3 time," " >3 time," "zero," "one task," "two tasks," or "three tasks"
N_hits_T1	The number of student hits in week 1 which categorized into "low," "moderate," or "high"
N_hits_T2	The number of student hits in week 2 which categorized into "low," "moderate," or "high"
N_hits_T3	The number of student hits in week 3 which categorized into "low," "moderate," or "high"
N_hits_T4	The number of student hits in week 4 which categorized into "low," "moderate," or "high"
N_hits_T5	The number of student hits in week 5 which categorized into "low," "moderate," or "high"
N_hits_T6	The number of student hits in week 6 which categorized into "low," "moderate," or "high"
N_hits_T7	The number of student hits in week 7 which categorized into "low," "moderate," or "high"
N_hits_T8	The number of student hits in week 8 which categorized into "low," "moderate," or "high"
N_time	The number of student hits on online learning which categorized into 'low', 'medium', or 'high'
Predictor Learning Habit	
N date	The number of days for students access to online learning which categorized into "low," "moderate," or "high"
Mode access days	Most of a weekday that student access to online learning (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday)
Mode Study Time	Most of the time that student access to online learning (morning, afternoon, evening, or night)
Predictor Student Profile	
Gender	Gender of Student (female, or male)

Attribute	Description
Age	The student age in years which categorized into <25, 25-35, 36-45, 46-55, or > 55
Region	The region where students live are categorized based on the islands in Indonesia, namely: Sumatra, Java, Borneo, Sulawesi, Bali & Nusa Tenggara, Maluku & Papua", or "Overseas"
Profession	Category of student work (teacher, police/army, civil servants, private, entrepreneur, works (no name), or "does not work)
Highest Education	Student's highest education (High school, diploma, bachelor, or'Postgraduate)
Range years of the highest education	Range years of the highest education which categorized into <5 years, 6-10 years, or > 10 years
Faculty	Faculty of Student (FE, FHISIP, FST, or FKIP)
Study Program	Student Study Programs are categorized into "science" and "social"
Semester	Semester students when taking online learning are categorized into: 1-2 smt, 3-4 smt, or >4 smt
Target	
Academic Achievement	Students' academic achievement based on semester-GPA (S-GPA) which categorized into "Poor," "Moderate," or "Good"

3.2. Exploratory Data Analysis

The dataset used in machine learning algorithms should not have missing values and outliers for maximum results. The results of presenting numerical data in the boxplot show that they are not included in the observation box located near the quartile, which are outliers (Fig. 2). Furthermore, the Exploration Data Analysis technique needs to be carried out before the data is included in the prediction model using a machine learning classification algorithm because the existing dataset has missing values and outliers.

Machine learning algorithms will produce better performance when their numerical input is at a standard scale. Based on this, we use exploratory data analysis as a set of techniques in engineering for data before applying it to machine learning algorithms.

Based on the exploratory data analysis technique, this study compares the accuracy of the machine learning algorithm between the two techniques for converting categorical values into numerical values. There are many techniques for converting categorical values to numeric values with different trade-offs and impacts on the dataset's features. This research focuses on Label-Encoder and One-hot Encoder techniques using the SciKit-Learn library in Python, which are expected to model and perform better.

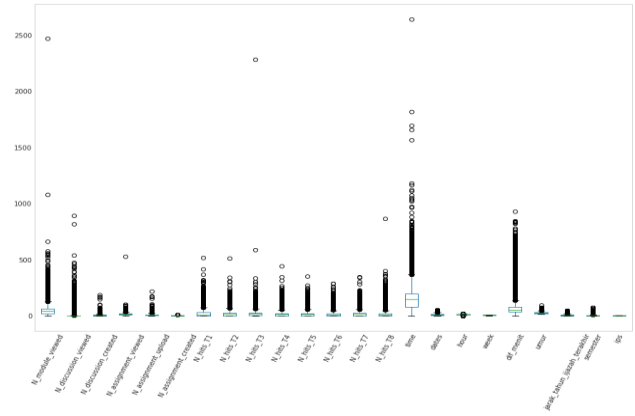

Fig. 2 Visual outlier data set with boxplot

Categorical data encoding has an important effect on the performance of machine learning algorithms [22]. This study compared the accuracy of the dataset with different encoding techniques. There are 3 (three) dataset schemes for encoding techniques, namely: scheme 1: all categorical data attributes (ordinal and nominal) are converted to numbers using encoding labels; scheme 2: all categorical attributes (ordinal and nominal) are converted to numbers using one-hot encoding; and scheme 3: ordinal categorical data attributes are converted to numbers using label encoding while nominal ones use a one-hot encoding.

3.3. Prediction Model

The prediction model in this study is a supervised machine learning method and uses a machine-learning classification algorithm. Supervised classification techniques are used to determine the best predictive model that fits the requirements to provide optimal results. The four machine learning algorithms chosen in this study are Random Forest (RF), Decision Tree (DT), Gradient Boosting (GB), and Adaptive Boosting (AB) because the prediction results in this study are the discrete class, namely: poor achievement, moderate achievement, and good achievement. The four algorithms include the Ensemble Learning algorithm, which combines several individual prediction models (called estimators) in an ensemble to improve the quality of predictions. These algorithms work extensively in Learning Analytics research and can work well with a missing value [23]–[28].

4. Result and Discussion

In this section, we perform an extensive experimental analysis of various machine learning algorithm classification models using profile data and student activity log data. The analysis was carried out on students who participated in online learning fully online at the Universitas Terbuka. Preprocessing research data uses Exploratory Data Analysis techniques to obtain data that can maximize the efficiency and effectiveness of machine learning algorithms.

This study used a supervised learning approach by classifying student academic achievement based on GPA data at the end of the semester (S-GPA). The classification of student academic achievement was grouped into 3 (three) based on the S-GPA obtained,

namely: 1) "poor" if the S-GPA is between 0 and 2.00; 2) "moderate" if the S-GPA is between 2.01 and 3.00, and 3) "good" if the S-GPA is between 3.01 and 4.00. Ehe RF, DT, GB, and AB algorithms were compared to select the most suitable and robust algorithm for this study. Algorithms vary depending on the dataset, efficiency, and performance of the tool library used. The machine-learning algorithm uses training data and test data in this study using 70% training data and 30% test data. The discussion of the results of this study is divided into description analysis, prediction and evaluation, and feature analysis. These are now presented and discussed.

4.1. Description Analysis

The dataset used in this study has gone through a data preprocessing process, which takes quite a lot of time, among other process stages. In the Exploratory Data Analysis technique, the user must experience a try-error so that the resulting data set follows the learning theory in general. At the end of the preprocessing activity, it is obtained 27 attributes with 373,732 instances can be used in the prediction model of this study. As shown in Table 4, statistical descriptions of categorical data show each attribute's uniqueness and highest frequency in this study.

Table 3 Statistical description of categorical data as predictors and targets

Attribute	Count	Unique Value	Symbol
Predictor Learning Activity			
N_module_viewed	373.732	3	X_1
N_discussion_viewed	373.732	2	X_2
N_discussion_created	373.732	3	X_3
N_assignment_viewed	373.732	3	X_4
N_assignment_upload	373.732	4	X_5
N_assignment_created	373.732	3	X_6
N_hits_T1	373.732	3	X_7
N_hits_T2	373.732	3	X_8
N_hits_T3	373.732	3	X_9
N_hits_T4	373.732	3	X_{10}
N_hits_T5	373.732	3	X_{11}
N_hits_T6	373.732	3	X_{12}
N_hits_T7	373.732	3	X_{13}
N_hits_T8	373.732	3	X_{14}
N_time	373.732	3	X_{15}
Predictor Learning Habit			
N_date	373.732	3	X_{16}
Mode_access_days	373.732	7	X_{17}
Mode_Study_Time	373.732	4	X_{18}
Predictor Student Profile			
Gender	373.732	2	X_{19}
Age	373.732	4	X_{20}
Region	373.732	8	X_{21}
Profession	373.732	7	X_{22}
Highest Education	373.732	4	X_{23}
Range years of the highest education	373.732	3	X_{24}
Faculty	373.732	4	X_{25}
Study Program	373.732	2	X_{26}
Semester	373.732	3	X_{27}
Target			
Academic Achievement	373.732	3	Y

Most machine learning algorithms are better off with numeric input, so the features from the categorical data in Table 3 are converted into numeric data. Furthermore, the prediction model is used to determine which target category of the predictors is as input. The machine learning algorithm produces a function $f: \mathbb{R}^n \rightarrow \{1,2,3\}$ to accomplish this task. The model can be written as the equation (4.1).

$$Y = f(X) \quad (4.1)$$

The model provides the input described by the vector X with the target category identified by the numeric code Y.

Students' learning activities and habits were captured with the input sequence $(X_1, X_2, X_3, \ldots, X_i \ \ldots, X_r)$ in this study. Therefore, the resulting prediction model output is a sequence $(Y_1, Y_2, Y_3, \ldots., Y_i \ldots., Y_r)$, with Y_i representing the category of student academic achievement in semester $X_i$according to the input sequence. Thus, the prediction model predicts the category of student academic achievement in the coming semester using activity data and student learning habits in the previous semester. This allows teachers to determine instructional strategies that are appropriate to the context of the learners.

4.2. Prediction and Evaluation of the Optimal Model

The key performance indicators described in this paper only used accuracy values due to the limited space in this paper. In the case of classification, accuracy is the most used evaluation metric in machine learning. Accuracy is the ratio between the number of true positive and true negative results of the comprehensive test data. The accuracy formula using a confusion matrix is shown in equation (4.2).

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Negatives + True\ Negatives + False\ Positives} \quad (4.2)$$

Table 4 presents the accuracy of scheme 1, scheme 2, and scheme 3 for the imbalanced data set and the balanced dataset using SMOTE. The machine learning classification algorithms used are RF, DT, GB, and AB.

Table 4 Comparison for accuracy of scheme and classifier algorithm (N = 373,732)

Scheme	Accuracy (%)			
	RF	DT	GB	AB
Original Data (Imbalance)				
Scheme 1 (28 columns)	72,47	60,20	59,706	58,730
Scheme 2 (104 column)	71,07	60,45	59,544	58,494
Scheme 3 (54 column)	74,33	60,32	59,712	58,728
Resample Data (Balance with SMOTE)				
Scheme 1 (28 columns)	84,40	52,02	56,61	52,59
Continuation of Table 4				
Scheme 2 (104 column)	81,05	50,69	50,41	49,32
Scheme 3 (54 column)	85,03	51,14	49,28	48,35

The imbalanced data set used in the classification model tends to show less accuracy in predicting minor

classes because classifiers tend to ignore minor class misclassifications. The number of attributes is insignificant with the accuracy achieved. Based on Table 5, the accuracy of the classification of student academic achievement in the three schemes between the Imbalance and Balance data has a different pattern in terms of the highest accuracy.

RF has the highest average accuracy for classifying the successful academic achievement of online learning students in this study. A balanced dataset with SMOTE using one-hot encoding techniques for nominal data and labels encoding techniques for ordinal data shows an accuracy of 85.03%. These results align with [29], which states that RF in many empirical studies has high predictive accuracy with good tolerance for abnormal values and noise. RF is a combination algorithm proposed by Breimanin 2001. If the prediction result is a discrete value, then the classification case, and if the prediction result is a constant value, then the regression case [29].

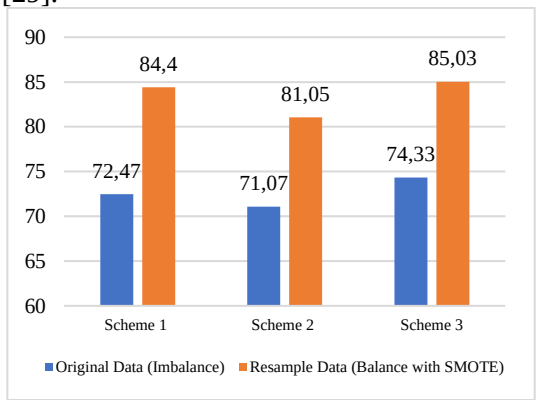

Fig. 3 Accuracy (%) of classification using the random forest algorithm

Fig. 3 compares the RF accuracy of each scheme, where the accuracy with the balanced dataset has higher accuracy than the imbalance data. The difference in accuracy between the imbalanced dataset and the balance is between 9.98% and 11.93%. Of course, the difference in these numbers is very significant in an accuracy value in a prediction model.

4.3. Feature Analysis

Each feature predictor influences the resulting prediction. To determine the influencing features, we determine the importance feature score. The RF algorithm can measure the relative importance of each feature on the predictions. Python's Sklearn library provides a tool that measures important features by looking at how many nodes are using those features. The core idea is to calculate the degree of reduction in RF prediction accuracy by adding noise to each feature. Fig. 4 shows the importance of the dataset's features using the RF and Sklearn classification algorithms.

According to Fig. 4, features that play a role in predicting academic achievement in this study are those related to the mode of days to access, student profession, and mode of study time of student access to UT-LMS. This result is in line with research [2] which states that four factors determine a major contribution to predicting student academic performance, profession, study time, and region.

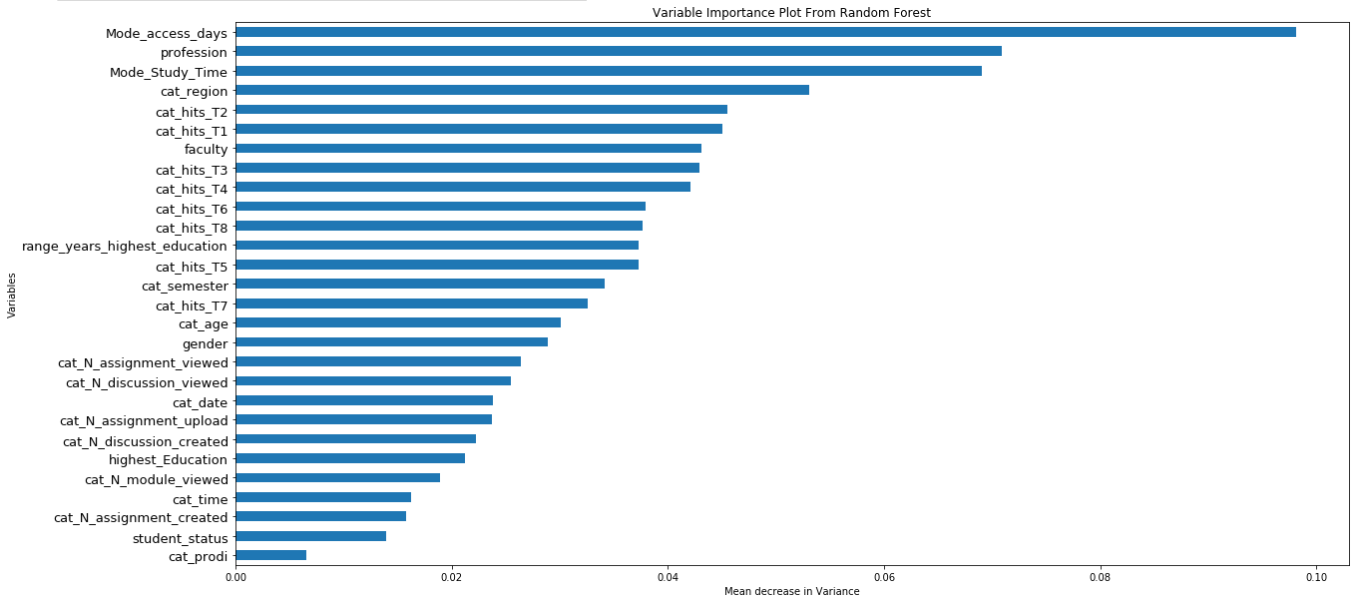

Fig. 4 Bar chart feature importance

5. Conclusion

The early prediction of student academic achievement in this study uses a machine-learning classification algorithm. Classifying student academic achievement can be done at the beginning of learning based on student profile data, activities, habit learning, and the previous semester's S-GPA. This study aims to first use categorical data as predictors and targets, and then the early prediction of student academic achievement with the selected model, and identify important features/predictors that are relevant. The categorical process of the dataset in this study is more

of an art than a science because categorizing each feature, both predictor and target, is subjective, not easy to explain or replicate. However, this research can be a basis for similar research with other scientific principles analysis. So that similar research after this produces a more optimal accuracy. The use of data categories as predictors and verification of model accuracy by testing datasets can be carried out as a routine procedure at the beginning of each semester. More accurate prediction models and specific critical features are used for further analysis. The empirical results of this study will provide knowledge for teachers/tutors in developing practical and realistic instructional strategies through making the right decisions and focusing on maximizing student academic achievement.

This study collected demographic profile data and UT-LMS log data to build predictive models of student academic achievement in fully online learning. EDA is carried out to define the dataset precisely and format the dataset for the ML classification model. Data categorization has been carried out to reduce noise caused by the distribution of the dataset. Several ML classification algorithms were applied to the dataset of this study. The ML classification model utilizes student learning activity data recorded in UT-LMS, combined with student demographic profiles and S-GPA. The first experiment results showed that the RF algorithm is the best algorithm with an accuracy of 85.03% on the imbalance data technique using SMOTE, the categorical data conversion technique using one-hot encoding technique for nominal data, and the label encoding technique for ordinal data. Table 4 reveals that the accuracy with the RF algorithm is higher than the accuracy with the DT, GB, and AB algorithms. The results of the second experiment show that the most important variables to predict academic achievement of fully online class students are the mode of days to access, student profession, and mode of study time of student access to UT-LMS. Most students in fully online learning access UT-LMS on Mondays and at night.

However, some limitations should be noted. First, students who are respondents in this article are limited to fully online class participants at UT in one semester. This could be improved by analysis for students from other universities over a longer semester span. Second, the course content factor is not included as a predictor in the dataset because the content collection for each available course is carried out in this study. The dataset collection in this study uses a data lake with a post-hoc approach, where metadata is generated after the data set is created, without the help of the dataset owner [1]. In this study, the dataset owner is a university that has different policies regarding student data. So that researchers have their challenges in the data acquisition process. In future work, we plan to use data spanning more semesters and then use the predicted results as the basis for recommending appropriate instructional strategies for fully online learning. This approach will help students achieve higher academic achievement at the end of the semester.

Acknowledgment

This research was supported by Direktorat Risetdan Pengembangan (Risbang) Universitas Indonesia through Hibah Publikasi Terindeks Internasional (PUTI) Q2 2020 (Number: NKB-4061/UN2.RST/HKP.05.00/2020) and Tokopedia-UI AI Centre of Excellence. The authors also deliver special thanks to the Vice-Rector for Academic Affairs at the Universitas Terbuka for his support for data collection and insight.

References

[1] VIMBI P. M. *The Good, the Bad, and the Ugly of Distance Learning in Higher Education. Trends in E-learning*. 2018: 17–29. [Online]. Available: http://dx.doi.org/10.5772/intechopen.75702.

[2] JAVIER. B-A, SONIA J. R., and SONIA P. Early prediction of undergraduate Student's academic performance in completely online learning: A five-year study. *Computers in Human Behavior*, 2021, 115(02): 106595. http://dx.doi.org/10.1016/j.chb.2020.106595.

[3] TRAVIS T. Y., CHARLES G, and SUSAN R. Defining and measuring academic success. *Practical Assessment, Research and Evaluation*, 2015, 20(5): 1–20. https://doi.org/10.7275/hz5x-tx03.

[4] RANNVEIG. G., TOVE I. D., TORE S., and ODDGEIR F. Relationships between learning approach, procrastination and academic achievement amongst first-year university students. *Higher Education*, 2017, 74(5): 757–774, https://doi.org/10.1007/s10734-016-0075-z.

[5] PENGFEI WU, SHENGQUAN YU and DAN WANG. Using a Learner-Topic Model for Mining Learner Interests in Open Learning Environments. *Journal of Educational Technology & Society*, 2018, 21(2): 192–204, [Online]. Available: http://www.jstor.org/stable/26388396.

[6] JUAN L. R, JUAN A. G., and ARTURO D. Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences (Switzerland)*, 2020. 10(3): 1-16, https://doi.org/10.3390/app10031042.

[7] RIANNE C., CHRIS S., AD K., and UWE M. Predicting student performance from LMS data: A Comparison of 17 Blended Courses Using Moodle LMS. *Institute of Electrical and Electronics Engineers Transactions on Learning Technologies*, 2017, 10(1): 17–29. https://doi.org/10.1109/TLT.2016.2616312.

[8] DRAGAN G., SHANE D., TIM R., and DANIJELA G. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *Internet and Higher Education*, 2016, 28: 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002.

[9] JACLYN B. Comparing online and blended learner's self-regulated learning strategies and academic performance. *The Internet and Higher Education*, 2017, 33: 24-32.https://doi.org/10.1016/j.iheduc.2017.01.004.

[10] MARÍA L. S, ÁNGEL F., and GUSTAVO A.

Technology behaviors in education innovation. *Computers in Human Behavior*, 2017, 72: 596–598. https://doi.org/10.1016/j.chb.2016.11.049.
[11] ANALÍA C., EDUARDO V., ABELARDO P., VIKTORIA P., ANGELA F., CARLA B., and STEFANIE L. Finding traces of self-Regulated learning in activity streams. *LAK'18: International Conference on Learning Analytics and Knowledge*, 2018: 191–200. https://dl.acm.org/doi/pdf/10.1145/3170358.3170381.
[12] LU T. H. OWEN, HUANG Q. Y. ANNA, HUANG H. C. JEFF, LIN Q. J. ALBERT, OGATA HIROAKI, and YANG H. J. STEPHEN. Applying learning analytics for the early prediction of students' academic performance in blended learning.*Educational Technology and Society*, 2018, 21(2): 220–232. https://www.jstor.org/stable/26388400.
[13] KRISTANTI A. P., and BOEDHI O. Successful Students in an Open and Distance Learning System. *Turkish Online Journal of Distance Education*, 2018, 19(2): 189–200. https://dergipark.org.tr/tr/download/article-file/458690.
[14] SANYAM B., SAI S., and ABHAY B. Application of learning analytics using clustering data Mining for Students' disposition analysis. *Education and Information Technologies*, 2017, 23: 957–984. https://doi.org/10.1007/s10639-017-9645-7.
[15] SADIQ H., MEHMET A., JOSAN DT, and ALEEZA S. Big Data and Learning Analytics Model. *International Journal of Computer Sciences and Engineering*, 2018, 6(7): 654–3. https://doi.org/10.26438/ijcse/v6i7.654663.
[16] ILYA M., STANISLAV P., and KSENIA T. Predictors of academic achievement in blended learning: The case of data science minor. *International Journal of Emerging Technologies in Learning*, 2019, 14(5): 64–74. https://doi.org/10.3991/ijet.v14i05.9512.
[17] PRATYA N., WONGPANYA N. DIREK T., KANAKARN P., and SITTICHAI B. Prediction Model of Student Achievement in Business Computer Disciplines. *International Journal of Emerging Technologies in Learning*, 2020, 15(20): 160–181. https://doi.org/10.3991/ijet.v15i20.15273.
[18] SAGARDEEP R., and SHAILENDRA N. S. Emerging trends in applications of big data in educational data mining and learning analytics. The 7th International Conference Confluence 2017 on Cloud Computing, Data Science and Engineering, 2017: 193-198. https://doi.org/10.1109/CONFLUENCE.2017.7943148.
[19] MOHAMED E. Predict Network, Application Performance Using Machine Learning and Predictive Analytics. *Department of Electrical, Computer & Telecom Engineering Technology, Rochester Institute of Technology*, 2019: 1-49. https://www.proquest.com/dissertations-theses/predict-network-application-performance-using/docview/2231085357/se-2?accountid=17242.
[20] RADHIKA R. H. Application of Machine Learning algorithms for betterment in education system, in International Conference on Automatic Control and Dynamic Optimization Techniques. *International Conference on Automatic Control and Dynamic Optimization Techniques, ICACDOT*, 2017: 1110–1114. https://doi.org/10.1109/ICACDOT.2016.7877759.
[21] AHMED A. M., CAO Han, and ZHANG Weizhen. Prediction of Students' Early Dropout Based on Their Interaction Logs in Online Learning Environment. *Interactive Learning Environments*, 2019: 1–20. https://doi.org/10.1080/10494820.2020.1727529.
[22] IVAN L., EDWIN A., ALEJANDRO M., HIRAM G., VICTOR M., and SAUL G. A memory-efficient encoding method for processing mixed-type data on machine learning. *Entropy*, 2020, 22(12): 1391. https://doi.org/10.3390/e22121391.
[23] EVANDRO BC, BALDOINO F., MARCELO A. S., FABRÍSIA F., ARAÚJO, and JOILSON R. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, 2017, 73: 247–256. https://doi.org/10.1016/j.chb.2017.01.047.
[24] LIANG JIAJUN, YANG JIAN, WU YONGJI, LI CHAO, and ZHENG Li. Big data application in education: Dropout prediction in edx MOOCs. *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, 2016: 440-443. https://doi.org/10.1109/BigMM.2016.70.
[25] CHARLES E. D, and CHANG LIU. Data Analytics Education: A Longitudinal View. *International Journal of Information, Business and Management*, 2019, 11(4): 8–19. https://www.intechopen.com/chapters/60465.
[26] LIU LEI, NI YIZHAO , ZHANG NANHUA , and PRATAP J. Mining patient-specific and contextual data with machine learning technologies to predict cancellation of children's surgery. *International Journal of Medical Informatics*, 2019, 129: 234-241. https://doi.org/10.1016/j.ijmedinf.2019.06.007.
[27] IQBAL H. S., KAYES A. S. M., and PAUL W. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *Journal of Big Data*, 2019, 6(1) 57. https://doi.org/10.1186/s40537-019-0219-y.
[28] NAGESWARI S, and PALLAVI M. G. Comparison of classification techniques on data mining.*International Journal of Emerging Technology and Innovative Engineering*, 2019, 5(5): 267–272. https://ssrn.com/abstract=3375191.
[29] CÉDRIC B. and JEFFREY S. R. Predicting University Students' Academic Success and Major Using Random Forests. *Research in Higher Education*, 2019, 60(7): 1048–1064. https://doi.org/10.1007/s11162-019-09546-y.
[30] Yuji R., Geon H., Steven E. W. A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective. *Institute of Electrical and Electronics Engineers Transactions on Knowledge and Data Engineering*, 2021, 33(4): 1328–1347. https://doi.org/10.1109/TKDE.2019.2946162.

参考文:

[1] VIMBI P. M. 高等教育遠程學習的好、壞和醜。電子學習的趨勢。2018 年：17-29。[在線的]。可用：http://dx.doi.org/10.5772/intechopen.75702。
[2] 哈維爾。 B-A、SONIA J. R. 和 SONIA P. 完全在線學習中本科學生學業成績的早期預測：一項為期五年的研究。人類行為中的計算機。 2021，115（02）：106595。http://dx.doi.org/10.1016/j.chb.2020.106595。
[3] TRAVIS T. Y.、CHARLES G 和 SUSAN R. 定義和衡量學術成功。實踐評估，研究與評估，2015 年，20(5)：1–20。https://doi.org/10.7275/hz5x-tx03。

[4] 蘭維格。 G.、TOVE I. D.、TORE S. 和 ODDGEIR F. 大學一年級學生的學習方法、拖延和學業成績之間的關係。高等教育。 2017, 74(5): 757–774, https://doi.org/10.1007/s10734-016-0075-z。
[5] 吳鵬飛，餘盛泉，王丹，使用學習者主題模型挖掘開放學習環境中的學習者興趣。教育技術與社會雜誌，2018，21（2）：192-204，[在線]。可用：http://www.jstor.org/stable/26388396。
[6] JUAN L. R、JUAN A. G. 和 ARTURO D. 通過機器學習分析和預測學生的表現：綜述。應用科學（瑞士），2020 年，10(3)：1-16，https://doi.org/10.3390/app10031042。
[7] RIANNE C.、CHRIS S.、AD K. 和 UWE M。根據学习管理系统數據預測學生表現：使用表情包学习管理系统比較 17 門混合課程。IEEE 學習技術彙刊，2017 年，10(1)：17–29，https://doi.org/10.1109/TLT.2016.2616312。
[8] DRAGAN G.、SHANE D.、TIM R. 和 DANIJELA G. 學習分析不應提倡一刀切：教學條件對預測學業成功的影響。互聯網與高等教育，2016 年，28：68-84，https://doi.org/10.1016/j.iheduc.2015.10.002。
[9] JACLYN B. 比較在線和混合學習者的自我調節學習策略和學業成績。互聯網與高等教育，2017，33：24-32，https://doi.org/10.1016/j.iheduc.2017.01.004。
[10] MARÍA L. S、ÁNGEL F. 和 GUSTAVO A. 教育創新中的技術行為。人類行為中的計算機，2017，72：596-598，https://doi.org/10.1016/j.chb.2016.11.049。
[11] ANALÍA C.、EDUARDO V.、ABELARDO P.、VIKTORIA P.、ANGELA F.、CARLA B. 和 STEFANIE L. 在活動流中尋找自我調節學習的痕跡。拉克'18：學習分析和知識國際會議，2018：191-200，https://dl.acm.org/doi/pdf/10.1145/3170358.3170381。
[12] LU T. H. OWEN、HUANG Q. Y. ANNA、HUANG H. C. JEFF、LIN Q. J. ALBERT、OGATA HIROAKI 和 YANG H. J. STEPHEN。應用學習分析對學生在混合學習中的學業表現進行早期預測。教育技術與社會，2018 年，21(2)：220–232，https://www.jstor.org/stable/26388400。
[13] KRISANTI A. P. 和 BOEDHI O. 開放遠程學習系統中的成功學生。土耳其遠程教育在線雜誌，2018 年，19(2)：189–200，https://dergipark.org.tr/tr/download/article-file/458690。
[14] SANYAM B.、SAI S. 和 ABHAY B. 使用聚類數據挖掘進行學習分析的應用，用於學生的性格分析。教育和信息技術，2017，23：957-984，https://doi.org/10.1007/s10639-017-9645-7。
[15] SADIQ H.、MEHMET A.、JOSAN D.T 和 ALEEZA S. 大數據和學習分析模型。國際計算機科學與工程雜誌，2018，6(7)：654-3。https://doi.org/10.26438/ijcse/v6i7.654663。
[16] ILYA M.、STANISLAV P. 和 KSENIA T. 混合學習中學業成就的預測因素：以輔修數據科學為例。國際新興學習技術雜誌，2019 年，14(5)：64–74。https://doi.org/10.3991/ijet.v14i05.9512。
[17] PRATYA N.、WONGPANYA N. DIREK T.、KANAKARN P. 和 SITTICHAI B. 商務計算機學科學生成績預測模型。國際新興學習技術雜誌，2020，15(20)：160–181。https://doi.org/10.3991/ijet.v15i20.15273。
[18] SAGARDEEP R. 和 SHAILENDRA N. S. 大數據在教育數據挖掘和學習分析中的應用的新興趨勢。合流 2017 第七屆雲計算、數據科學與工程國際會議，2017: 193-198。https://doi.org/10.1109/CONFLUENCE.2017.7943148.
[19] MOHAMED E. 預測網絡，使用機器學習和預測分析的應用程序性能。羅徹斯特理工學院電氣、計算機與電信工程技術系，2019:1-49。https://www.proquest.com/dissertations-theses/predict-network-application-performance-using/docview/2231085357/se -2?accountid=17242。
[20] RADHIKA R. H. 機器學習算法在教育系統中的應用，在自動控制和動態優化技術國際會議上。自動控制和動態優化技術國際會議，2017：1110-1114。https://doi.org/10.1109/ICACDOT.2016.7877759。
[21] AHMED A. M., 曹涵, 張維珍.基於在線學習環境中學生互動日誌的學生早期輟學預測。互動學習環境，2019：1-20。https://doi.org/10.1080/10494820.2020.1727529。
[22] IVAN L.、EDWIN A.、ALEJANDRO M.、HIRAM G.、VICTOR M. 和 SAUL G. 一種在機器學習中處理混合類型數據的內存高效編碼方法。熵，2020，22（12）：1391。https://doi.org/10.3390/e22121391。
[23] EVANDRO B.C.、BALDOINO F.、MARCELO A.S.、FABRÍSIA F.、ARAÚJO 和 JOILSON R。評估教育數據挖掘技術的有效性，用於早期預測學生在入門編程課程中的學業失敗。人類行為中的計算機，2017，73：247–256。https://doi.org/10.1016/j.chb.2017.01.047。
[24] 梁家軍;楊健;吳永基;李超；還有鄭麗。教育中的大數據應用：大規模開放在線課程中的輟學預測。 2016 電氣和電子工程師學會第二屆多媒體大數據國際會議，2016：440-443。https://doi.org/10.1109/BigMM.2016.70。
[25 CHARLES E. D, 和 CHANG LIU。數據分析教育：縱向視圖。國際信息、商業和管理雜誌，2019 年，11(4)：8–19。https://www.intechopen.com/chapters/60465。
[26] LIU LEI, NI YIZHAO , ZHANG NANHUA 和 PRATAP J. 用機器學習技術挖掘患者特定和上下文數據以預測兒童手術取消。國際醫學信息學雜誌，2019 年，129: 234-241 頁。https://doi.org/10.1016/j.ijmedinf.2019.06.007。
[27] IQBAL H. S.、KAYES A. S. M. 和 PAUL W. 機器學習分類模型的有效性分析，用於預測個性化上下文感知智能手機的使用情況。大數據雜誌，2019，6(1) 57。https://doi.org/10.1186/s40537-019-0219-y。
[28] NAGESWARI S 和 PALLAVI M. G. 數據挖掘分類技術的比較。國際新興技術與創新工程雜誌，2019，5(5)：267–272。https://ssrn.com/abstract=3375191。
[29] CÉDRIC B. 和 JEFFREY S. R. 使用隨機森林預測大學生的學業成功和專業。高等教育研究，2019，60(7): 1048-1064。https://doi.org/10.1007/s11162-019-09546-y。
[30] YUJIR.、GEONH.、STEVEN E. W. 機器學習數據收集調查：大數據-人工智能集成視角。 IEEE 知識與數據工程彙刊，2021 年，33(4)：1328–1347。https://doi.org/10.1109/TKDE.2019.2946162。