ASIA-AFRICA JOURNAL OF RECENT SCIENTIFIC RESEARCH (AAJRSR) VOL. 2, 2022
A PUBLICATION OF IAPAAR IN COLLABORATION WITH MEDI-CAPS UNIVERSITY, INDIA
ISSN: 2814-0397 Available online at https://journals.iapaar.com/index.php/AAJRSR

MODELING AND PREDICTING THE ACADEMIC ACHIEVEMENT SUCCESS OF NATIONAL OPEN UNIVERSITY OF NIGERIA STUDENTS USING MACHINE LEARNING CLASSIFICATION ALGORITHMS

[1]Nseabasi Essien, *Phd*

[2]Daniel Billy

&
[3]Emmanuel Philip Ododo
[1,2,3] Department of Computer and Robotics Education
University of Uyo, Nigeria

Abstract
Machine learning (ML) is transforming education and fundamentally changing teaching, learning and research. The ML technique helps the institution to utilize the resources in better waysand produces results in the best possible effective manner. The learning combines various processes like data preparation, classification, association, building models, training, clustering, prediction etc. toimprove performance of students. The main objective of the study was to predict early academic achievement of fully online learning studentsusing 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 NOUN-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 NOUN-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 NOUN-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 thana 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: machine learning, artificial intelligence, prediction, academic achievement

Introduction

National Open University of Nigeria (NOUN) provides alternative learning and educational opportunities that citizens can access without geographical, physical, social, and economic constraints. The National Open University of Nigeria and its library was established as an instrument for National Development and was reflected in the National Policy of Education (FRN 2013) thus, the legal instrument establishing the Open University Act of July 22, 1983 which was suspended in 1984. Its actualization started in 2001 by its resuscitation given birthto the National Open University of Nigeria. The primary objective of establishing national Open University of Nigeria is to give equal

opportunity to those who could not go to conventional University, thus, helps the individual to develop from this point onward to attain what he or she had hither to thought was unattainable. The major difference between studentsin National Open University of Nigeria and others in the conventional universities thereforeis the mode of instruction. The nature of studies in National Open University of Nigeria arenot the same with conventional University in Nigeria which run full time and residence campus studies, unlike National Open University that run part time –weekend studies, off campus studies. Their students are mainly civil servants, traders and matured old citizens who desired to read, the motto is "work and learn".

Along with the development of Information and Communication Technology (ICT), NOUN apply technology that allows students to learn across time and space according to students' flexibility. The application of ICT in education is known as distance learning, which is usually via the internet, so that the characteristics of students in distance learning are very heterogeneous. This can be seen from the diversity of students participating in online learning based on their demographic profiles. Distance learning is a system that includes applying several ICTs to benefit students' learning and education anytime and anywhere. It isimportant to understand how students learn to determine the appropriate learning strategies through online learning in the knowledge construction process.

The Distance learning system provides students with more interactivity and flexibility to use online devicesat anytime and anywhere. On the one hand, teachers in distance learning, especially learning fully online (FO), do not have complete information about the characteristics, habits, and activities of learning, as wellas 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 (Javier, Sonia and Sonia, 2021).

Learning management systems (LMS) are widely used in distance 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, teacherssometimes 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 (Travis, Charles and Susan, 2015). Previous studies have shown that instructionalstrategies positively predict GPA (Rannveig, Tove, Tore and Oddgeir, 2017). On the one hand, instructional strategy training and motivation did produce a higher GPA of students and positively affected the learning outcomes of Open and Distance Learning students (Pengfei, Shengquan and Dan, 2018).

Modeling and predicting the academic achievement success of distance 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, weconstructed 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) (Juan, Juan and Arturo, 2020). Machine learning accurately predicts student performance because various authors present conflicting results regarding theaccuracy 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 NOUN students on the student demographic profile data, student learning habits data, and student activity in the learning management system (LMS).

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-sourcethat can develop insights on data analysis. We use various machine learning algorithms, which wereapplied 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 statisticaltechniques selected are suitable for attributes to thedomain and categorical education. The main steps in this research use a data science approach, as shown in Fig. 1.

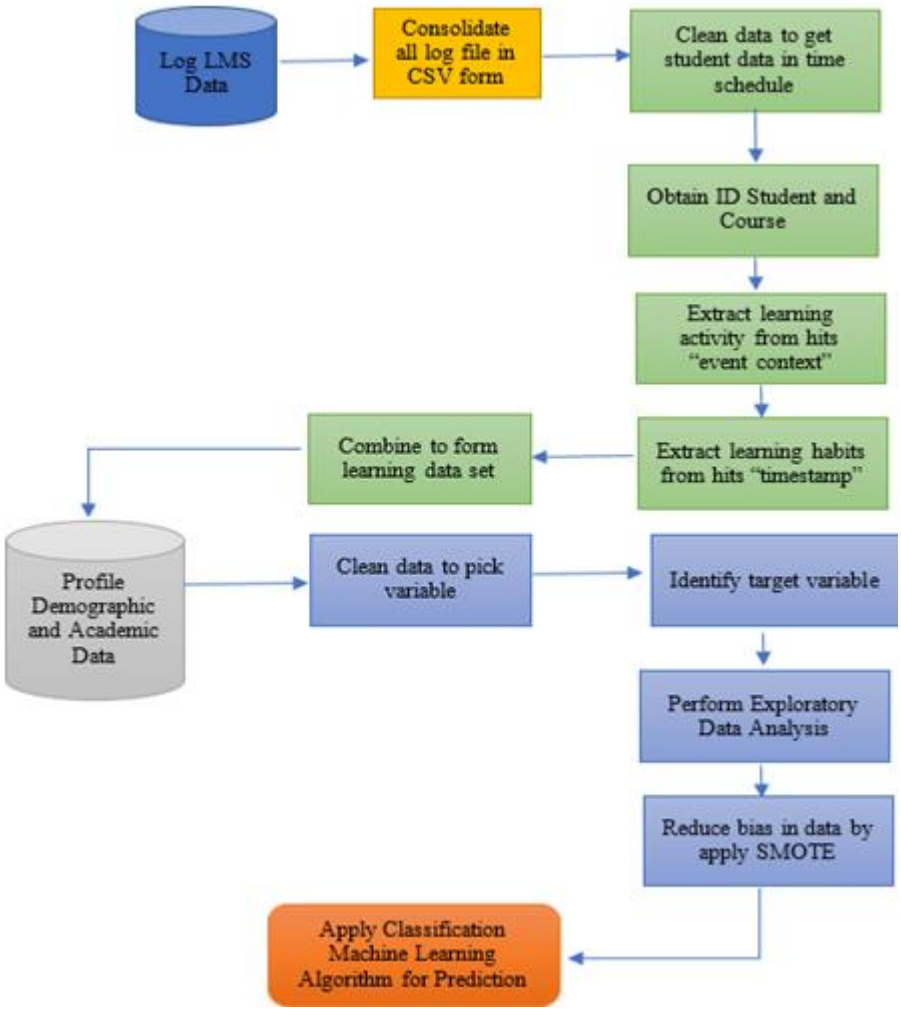

Fig. 1: Data science approach used in this study

Dataset

Instruction at NOUN is based on the principle of self- regulated learning, which is an instructional processthat 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.

Distance learning at NOUN is provided in the form of an online tutorial using the Moodle LMS platform. An online tutorial is a learning service provided by NOUN, held in 8 sessions for eight consecutive weeks. To participate in the online tutorial, NOUN students must activate the LMS and fill out a form available to participate in the online tutorial. The online tutorialassessment 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 LMS in 2019/2020. Respondents of 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). The student profile data was from the Student Academic Information System and student log data from LMS.

The Moodle log file as a 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 ismore commonly used in learning (Rianne, Chris, Ad and Uwe, 2017). 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 Online Distance Learning system.

Referring to the Online Tutorial Guide, online learning courses at NOUN have the same structure, namely: initiation materials, discussion activities, and assignments. 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.

The data described in Table 1 is the raw LMS log data obtained by downloading from the NOUN 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 NOUN 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, whilestudent 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 (Rianne, Chris, Ad and Uwe, 2017) 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 areneeded 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 studyuses Microsoft Excel to manage data sets in different formats and forms. As for preprocessing, this studyuses the Jupyter Notebook with the Python 3.6programming language, Pandas, NumPy, andMatplotlib. 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 1 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)
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"

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.

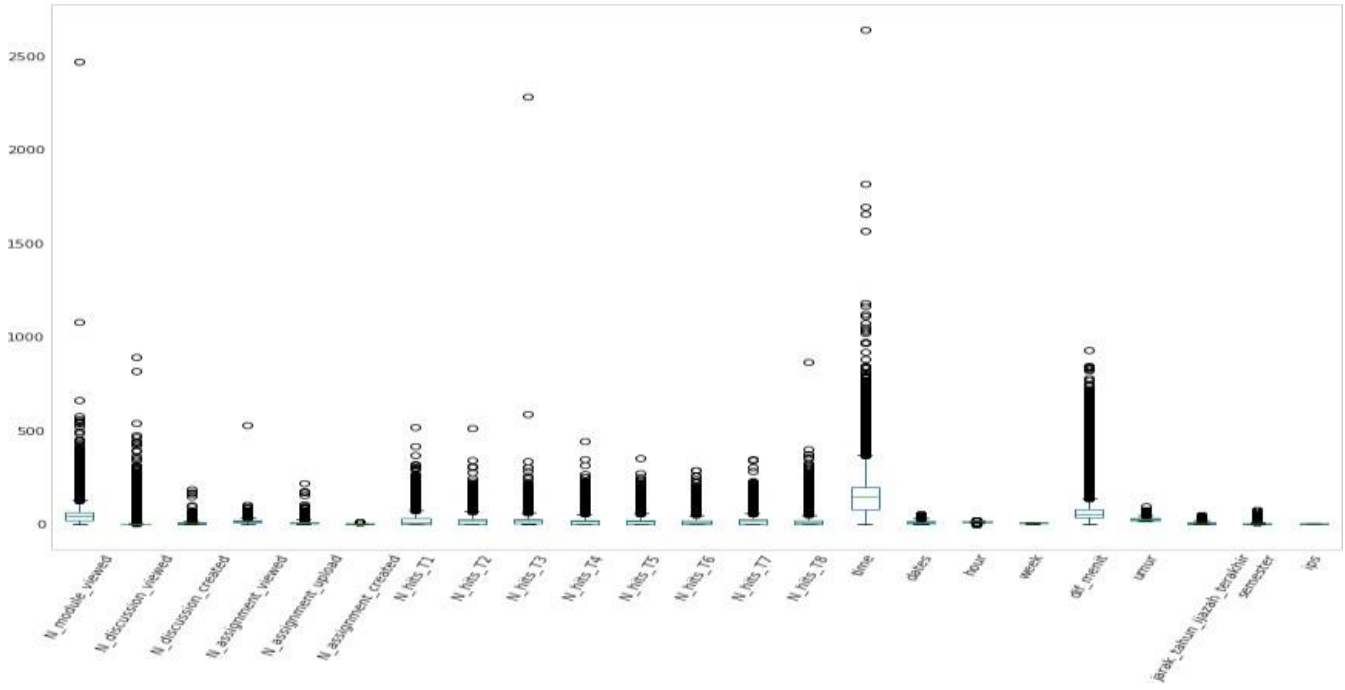

Fig 2: visualizing outliers data set with boxplot

Categorical data encoding has an important effect on the performance of machine learning algorithms (Ivan, Edwin, Alejandro, Hiram, Victor, and Saul, 2020). 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.

Prediction Model

The prediction model in this study is a supervised machine learning method and uses a machine-learning classification algorithm. Supervised classificationtechniques are used to determine the best predictive model that fits the requirements to provide optimal results. The four machine learning algorithms chosen inthis study are Random Forest (RF), Decision Tree (DT), Gradient Boosting (GB), and Adaptive Boosting (AB) because the prediction results in this study are thediscrete class, namely: poor achievement, moderate achievement, and good achievement. The fouralgorithms 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 (Evandro, Baldoino, Marcelo,

Fabrísia, Araújo, and Joilson, 2017).

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 toselect 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.

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 2, statistical descriptions of categorical data show each attribute's uniqueness and highest frequency in this study.

Table 2: Statistical description of categorical data as predictors

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 categoricaldata in Table 2 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 (1).

$Y = (X)$ (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, $X_2, X_3, \dots, X_i \dots, X_r$) in this study. Therefore, the resulting prediction model output is a sequence ($Y_1, Y_2, Y_3, \dots\dots, Y_i \dots\dots, Y_r$), 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

academicachievement in the coming semester using activity dataand student learning habits in the previous semester.

This allows teachers to determine instructional strategies that are appropriate to the context of the learners.

Prediction and Evaluation of the Optimal Model

This paper only used accuracy values due to the limitedspace 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 (2).

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Negatives + True\ Negatives + False\ Positives}$$

(2).

Table 3 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 3 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	85,03	51,14	49,28	48,35

column)

The imbalanced data set used in the classificationmodel 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 2, 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 learningstudents in this study. A balanced dataset with SMOTE using one-hot encoding techniques for nominal dataand labels encoding techniques for ordinal data shows an accuracy of 85.03%. These results align with (Cédric and Jeffrey, 2019), 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 regressioncase (Cédric and Jeffrey, 2019).

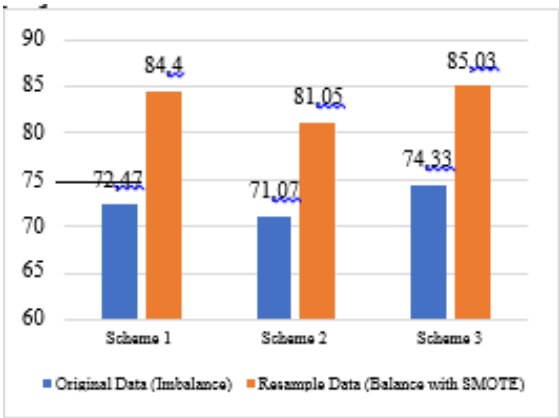

Fig 2: accuracy chart of schemas of original data (imbalance) and Resample data (balance with SMOTE)

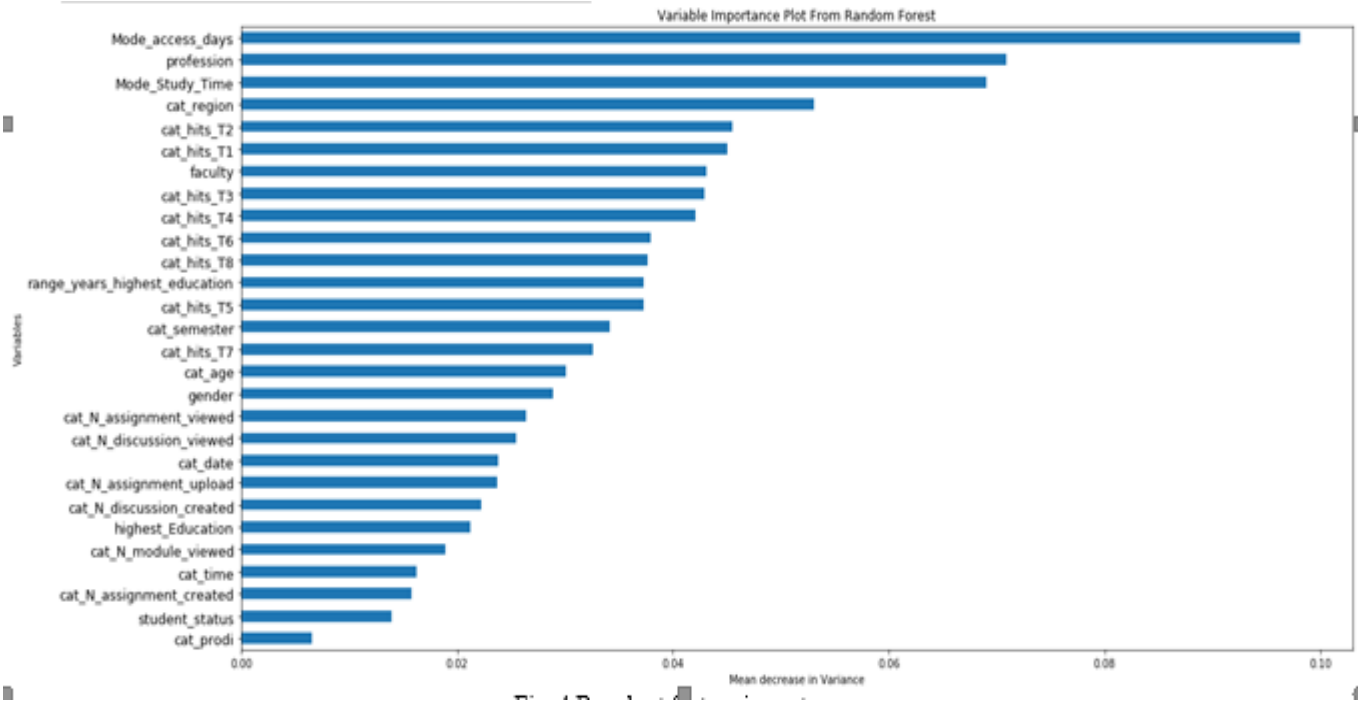

Fig. 3: Bar chart feature importance

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%. Ofcourse, the difference in these numbers is very significant in an accuracy value in a prediction model.

Feature Analysis

Each feature predictor influences the resultingprediction. To determine the influencing features, we determine the importance feature score. The RFalgorithm can measure the relative importance of each feature on the predictions. Python's Sklearn library provides a tool that measures important features bylooking 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. 3, features that play a role inpredicting 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 toLMS. This result is in line with research (Javier, Sonia, and Sonia, 2021) which states that four factors determine a major contribution to predicting student academic performance, profession, study time, and region.

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, noteasy to explain or replicate. However, this research can be a basis for similar research with other scientificprinciples 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 studentacademic achievement.

This study collected demographic profile data and 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 LMS, combined with student demographic profiles and S-GPA. The firstexperiment 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 3 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 access days, student profession, and mode of study time of student access to LMS. Most students in fully online learning access LMS on Mondays and at night.

References

Cédric B. and Jeffrey S. R. (2019) Predicting University Students' Academic Success and Major Using Random Forests. *Research in Higher Education*, 60(7): 1048– 1064.

Evandro B. C, Baldoino F., Marcelo A. S., Fabrísia F., Araújo, and Joilson R. (2017) Evaluating the effectiveness of educational data mining techniques for earlyprediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*,73: 247–256.

Ivan L., Edwin A., Alejandro M., Hiram G., Victor M., and Saul G. (2020) A memory-efficient encoding method for processing mixed-type data on machine learning. *Entropy*, 22(12): 1391. https://doi.org/10.3390/e22121391.

Javier. B-A, Sonia J. R., and Sonia P. (2021) Early prediction of undergraduate Student's academic performancein 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.

Juan L. R, Juan A. G., and Arturo D. (2020) Analyzing and predicting students' performance by

means of machine learning: A review. *Applied Sciences (Switzerland),* 10(3): 1-16, https://doi.org/10.3390/app10031042.

Pengfei W. U., Shengquan Y. U. and Dan W. (2018) Using a Learner-Topic Model for Mining Learner Interestsin Open Learning Environments. *Journal of Educational Technology & Society*, 21(2): 192–204, [Online].

Rannveig. G., Tove I. D., Tore S., and Oddgeir, F. (2017) Relationships between learning approach, procrastination and academic achievement amongst first-year university students. *Higher Education*, 74(5): 757–774,

Rianne C., Chris S., Ad K., and Uwe M. (2017) 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*, 10(1): 17–29.

Travis T. Y., Charles G, and Susan R. (2015) Defining and measuring academic success. *Practical Assessment, Research and Evaluation*, 2015, 20(5): 1–20.