

# Utilizing Machine Learning Techniques for Learning Analytics: A Case Study of Moodle LMS Activity Log Analysis

I Dewa Made Bayu Atmaja Darmawan

Informatic Department, Universitas Udayana,  
Jimbaran, Indonesia  
dewabayu@unud.ac.id

## Abstract

*Learning analytics collects data, analyzes, and interprets the learning process that has taken place. The output of this method can be used to improve the quality of teaching or learning. Moodle is a popular learning management system (LMS) used for online learning. Various learning activities carried out by students are recorded in the activity log. This paper shows the potential of using machine learning methods to analyze activity logs taken from Moodle LMS. The sample used in this study refers to implementing the Digital Society course, which students from different fields of science attend. This paper describes using supervised and unsupervised learning on activity log data taken from the Moodle LMS. The variables used as datasets include the frequency of activity reading pdf material, scores, videos, forums, quizzes, and graduation status. The supervised learning model that was built succeeded in obtaining an accuracy of 100% in the application of logistic regression and Naïve Bayes Classification. Unsupervised learning clustered all the data and showed the cluster related to the frequency of online learning activities and students' assessment success status.*

**Keywords:** machine learning, log activity, learning analytic, moodle, elearning

## 1. Introduction

Data analytics methods are used in learning analytics to examine and comprehend educational and learning processes. It entails gathering and examining information produced by students' interactions with educational resources such as online courses, learning management systems, and other software. Learning analytics aims to gain knowledge about how students engage with educational resources, which aims to use this knowledge to enhance teaching and learning results. Learning analytics can be used in a variety of ways, including: monitoring learner progress, personalizing learning, improving course design, predicting student outcomes, and conducting research [1]. Learning analytics can potentially revolutionize how we understand and approach education. By leveraging data analytics techniques to gain insights into learning processes and outcomes, we can better understand how to support learners and improve educational outcomes.

Many free and paid software programs have been proposed for educational uses in the form of Learning Management Systems (LMS), where students can actively engage in social learning activities while interacting with teachers and peers. For example, Moodle is a free interactive learning environment where students engage with didactic resources, peers, and tutors to learn. Many public institutions utilize this LMS to teach people for lifelong learning initiatives. Online learning allows each learner's activity on the LMS in a course to be recorded in the activity log. The learning activities included obedience in collecting assignments, frequency of accessing the course material, activeness in discussion forums, and up assessments in the form of summative or formative assessments.

Machine learning (ML) uses computer algorithms to learn patterns and relationships in data without being explicitly programmed. In other words, machine learning algorithms can "learn" from data and improve their performance over time without being explicitly programmed to do so. The

field of machine learning has been overgrown in recent years, driven by the explosion of available data and advances in computing power. Machine learning algorithms are used in various applications, including natural language processing, image recognition, fraud detection, personalized recommendations, and predictive analytics.

The machine learning training stage risks producing misinterpretations due to the data used [2]. As explained further, there are two types of logs: human-readable and database logs. Human-readable logs are the result of extraction from database logs. Because system administrators only own database access, this research uses human-readable logs. Other research uses machine learning to perform learning analytics. Several studies have been conducted regarding using machine learning as learning analytics or educational data mining. Graduation prediction using an artificial neural network (ANN) obtained an accuracy value of 73.21%[3]. Apart from predicting graduation, accurate predictions of students' length of study have been carried out using Random Forest and Gradient Boosting [4]. The accuracy of predicting the duration of the study using the method used succeeded in getting an accuracy value of 82.64%.

This paper examines the utilization of machine learning with supervised learning and unsupervised learning types on activity logs extracted from Moodle. The function of machine learning in data analysis in learning analytics jobs will be described in depth. As sample data, this paper uses the activity log of the Digital Society course, which was attended by 1,000 students from several non-IT study programs.

## **2. Research Methods**

### **2.1. Machine Learning Concept**

Machine learning is a multidisciplinary field that includes probability theory, statistics, approximation theory, convex analysis, and computational complexity theory, among others [5]. Machine learning is the study of how to enhance a system's performance by utilizing intelligent computing and experience. Experience generates the corresponding algorithm model, and the process of algorithm model generation is essentially the process of machine automatic learning [6].

The three primary types of machine learning algorithms are supervised learning, unsupervised learning, and semisupervised learning. Supervised learning is commonly categorized into two distinct types of algorithms: regression and classification. Regression analysis is a statistical approach that involves fitting a mathematical function to a set of input and output variables, where the function is continuous. Classification refers to the process of associating input variables with distinct categories. Unsupervised learning refers to a type of machine learning where the output is not predetermined or known beforehand. Clustering can be utilized to extract a distinct structure from the data. Unsupervised learning involves the absence of multiple labels or the presence of a single label. Semisupervised learning is a machine learning approach that integrates supervised learning with unsupervised learning. Within machine learning, two distinct types of data exist: labeled data and unlabeled data. The utilization of semisupervised learning has the potential to enhance the efficacy and precision of the learning process.

In this paper, we limit the evaluation of algorithms for supervised learning using logistic regression and Naive Bayes, while for supervised learning, we use K-Mean Clustering.

### **2.2. Learning Analytics**

Learning analytics is becoming increasingly popular in the educational community, including universities. There are four essential elements of the learning analytics process as shown in Figure 1 [7]. Learning Analytics is the measurement, accumulation, analysis, and reporting of data about learners and their contexts for the purpose of optimizing learning and its environment [8]. Learning Analytics is a multidisciplinary discipline at the intersection of business intelligence, web analytics, educational data mining, and recommender/recommendation systems [9]. In addition to formal and informal education, LA is also applicable to non-formal learning.



**Figure 1.** The basic elements of learning analytics [7],

### 2.2.1. Data

Data is considered the fundamental asset for analytics and serves as the basis for generating analytical insights. In the context of education, data encompasses information that is typically collected during the learning process and pertains to various aspects such as the learners, learning environment, learning interactions, and learning outcomes. The data in this study was obtained from activity logs recorded by LMS Moodle.

### 2.2.2. Analysis

The process of analysis involves the conversion of gathered data into useful information through the application of mathematical and statistical algorithms and techniques. This involves the cleansing, transformation, and modeling of data with the aim of uncovering significant insights that can aid in decision-making and action. Before data modelling, this research carried out feature selection using Pearson Correlation. With Pearson Correlation, the correlation value of features with the final value can be seen and becomes the basis for selecting features as model input.

### 2.2.3. Report

The report serves the purpose of summarizing the analysis of the gathered data pertaining to learning and presenting it in a meaningful manner. It involves a series of procedures that facilitate the organization and presentation of the outcomes of the analysis of learners' and learning data in the form of charts and tables. The act of presenting information regarding learners' and learning data can offer valuable perspectives on the learners' conditions during the learning process. Analyzing these perspectives can facilitate data-informed decision-making to determine appropriate courses of action.

### 2.2.4. Action

The primary objective of any learning analytics process is to facilitate action. This entails a series of informed decisions and practical interventions that educational stakeholders will implement. The success or failure of analytical endeavors hinges on the outcomes of subsequent measures taken. The implementation of learning analytics is deemed valuable solely if it leads to a consequential course of action.

## 2.3. Moodle Log Activity

The online learning environment utilizes the Learning Management System (LMS) to manage the learning process. MOOCs (Massive Open Online Courses) are online courses designed for distance learning characterized by a large number of students. Examples are Coursera, Cisco Academy, and Moodle LMS, which various universities use to organize online learning. Moodle is an example of a popular LMS used to manage online learning [10]. The Moodle LMS owns several essential features, including Assignment, Attendance, Choice, Lesson, Page, Quiz, URL, Workshop, Folder, File, Glossary, SCORM Package, Feedback, and Database.

As students use the LMS, they leave behind lots of tracks which are then stored in log files generated by the platform. The Moodle log files record when users access the system, how long they spend on each page, which resources or activities they access, and which assessments they complete. In addition, log files can record information about system errors and problems, such as logon failures and server errors.

Moodle log files contain information that can be used for various purposes, including monitoring pupil progress, evaluating the effectiveness of instructional design, identifying problems or issues

within the system, and providing data for research and analysis. Users can filter and search log data to locate specific information using Moodle's user-friendly interface for accessing and viewing log files. Contains how data is collected, data sources and ways of data analysis.

### 3. Result and Discussion

This section will expound on the utilization of machine learning techniques to conduct analytical learning on Moodle activity logs. The technique for acquiring log data on Moodle, as well as the use of supervised learning and unsupervised learning, will serve as the basis for the discussion.

#### 3.1. Data Collection and Preparation

Moodle log data captures user activity at different levels : teaching level, participating level, system level [2]. The log data in this paper is at the participating level, which shows actions related to learning activities such as module views, posting to forums, or attempting a quiz. Figure 2 shows the process of taking a log of participation in a course in the Moodle LMS. The course participation log shows the number of participation carried out for each activity, such as participation in forums, quizzes, or viewing learning materials.

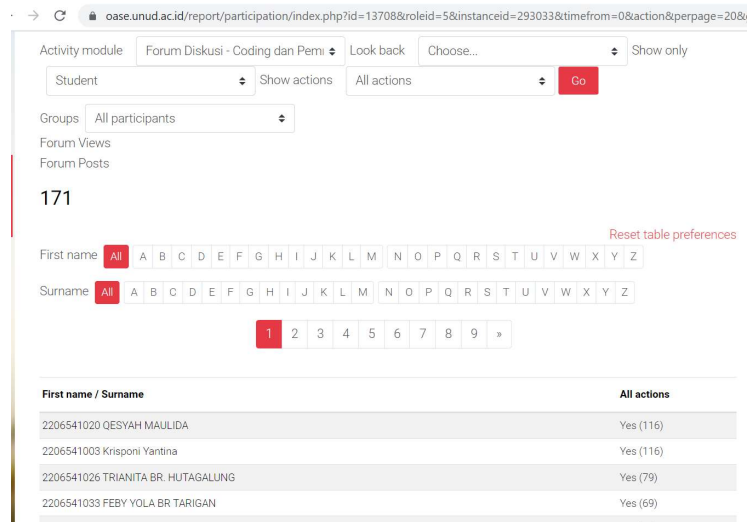
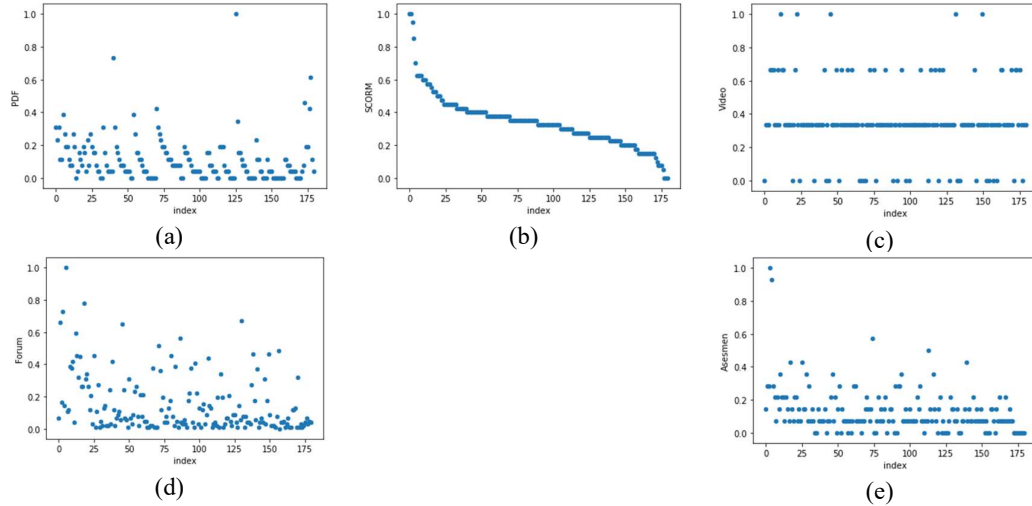


Figure 2. Data collection form log course participation in Moodle

The present investigation employed a set of features comprising the count of material views in pdf format, the count of accesses to self-directed learning content in the form of scorms, the count of video views, the count of forum posts, and the count of quiz attempts. The objective is to attain a passing status, which is assigned a value of 1 when the mean score of the quizzes is 65 or higher. The data as mentioned earlier pertains to an activity within the Coding and Programming domain of the Digital Society curriculum. Following the data collection stage, the subsequent step involves data preparation, which includes normalization through the min-max technique and eliminating outliers. The data preparation process reduced the sample size used in the study to 180, with each subset consisting of 90 data points categorized as either pass or fail. The distribution of data for the utilized features is depicted in Figure 3. Figure 3 memperlihatkan perilaku student terhadap course berbeda setiap student. Figure 3 shows student behaviour towards different courses for each student. Figures 3(a) and 3(e) appear to have almost the same pattern, where most values are below 0.5. This can be explained by the fact that students generally download PDF materials, so only a little activity on PDF resources is recorded. Likewise, for assessment activities, even though more than one attempt was given, the average number of attempts from the sample of students used was almost the same. Figure 3(b) shows that activity on SCORM resources appears to have an even distribution, indicating students' need to understand different content. SCORM is a resource that provides material content and a self-assessment that guides students to repeat until the passing grade from the

self-assessment is met. Figure 3(c) shows student activities regarding video watching activities, which can be grouped into four groups. Student participation in forum activities varies, with the majority of students not being actively involved in the forum; this is shown in Figure 3(d), where most of the values are in the range of 0.0 to 0.2. After understanding the data used, the next stage is to see the correlation between each feature and the final result (passing status).



**Figure 3.** Features Value Distribution, a) PDF, b) SCORM, c) Video, d) Forum, and e) Quiz

In the process of data preparation, a correlation analysis is conducted to identify features that exhibit no correlation with the target variable. The Pearson method is employed in the correlation test, utilizing the Pandas library. Table 1 displays the correlation matrix that pertains to the relationship between the features and targets derived from the utilized data. The findings of the correlation analysis indicate that there exists a negligible correlation (0.047949) between the success status of students in the final assessment and the PDF feature, which pertains to the count of students accessing course material in PDF format. This phenomenon may occur due to the ability of students to procure PDF materials, thereby rendering the precise count of activities involved in reading said materials inaccurately reflected in the activity log. On the other hand, the level of engagement exhibited by students in the forum displays the most robust correlation when compared to other factors in determining their success in fulfilling the assessment requirements. The utilization of the PDF feature is not employed in the construction of machine learning models.

**Table 1.** Pearson Correlation Test Results

	PDF	SCORM	Video	Forum	Quiz	Status
PDF	1.000000	0.140537	0.073035	0.095730	0.004948	<b>0.047949</b>
SCORM	0.140537	1.000000	0.066451	0.393399	0.431531	<b>0.399205</b>
Video	0.073035	0.066451	1.000000	0.225371	0.174714	<b>0.247052</b>
Forum	0.095730	0.393399	0.225371	1.000000	0.400634	<b>0.644597</b>
Quiz	0.004948	0.431531	0.174714	0.400634	1.000000	<b>0.427085</b>
Status	0.047949	0.399205	0.247052	0.644597	0.427085	<b>1.000000</b>

### 3.2. Supervised Learning

In supervised learning models are built using data that is equipped with labels as the target variable. The data used for building has several features and targets with binary type. The utilization of supervised learning in analyzing log data about the frequency of learning activities in e-learning facilitates the development of predictive models. These models enable the comprehension of user behavior, forecasting specific events or outcomes, and enhancing decision-making processes in online learning. The following are instances of executing supervised learning:

1. Regression: In certain instances, it may be desirable to make projections regarding a continuous variable using the learning log information. In this scenario, regression algorithms such as Linear Regression, Logistic Regression, Polynomial Regression, or Neural Networks can be employed. An illustrative approach involves constructing a predictive model that estimates the duration of a student's engagement with a learning module, utilizing pre-existing characteristics.
2. Classification: Classification algorithms, including Naïve Bayes, Decision Trees, Random Forests, and Support Vector Machines (SVM), can be utilized to construct models that can forecast specific categories or labels by analyzing the features present in the learning log data. An illustrative approach involves constructing a predictive model that can determine the likelihood of a student's course completion based on their activity patterns.

### 3.3. Regression

The regression technique is employed to construct a predictive model for determining the passing status (target) by utilizing the features utilized in the model's construction. The employed statistical technique is Logistic Regression. The logistic regression approach was chosen due to the binary nature of the target variable and the mutual independence of the features, as reported in reference [11]. The process of constructing a regression model involves several key stages, namely data loading, feature selection, data partitioning, model building and prediction, and model assessment. In this study, 75% of the total data was allocated for training purposes, while the remaining 25% was reserved for data testing. The process of assessing the performance of a model is accomplished through the utilization of a confusion matrix. The results depicted in Figure 5 demonstrate that logistic regression achieved a 100% accuracy rate in its predictive capabilities.

```
#Model Development and Prediction
# import the class
from sklearn.linear_model import LogisticRegression

# instantiate the model (using the default parameters)
logreg = LogisticRegression(random_state=16)

# fit the model with data
logreg.fit(X_train, y_train)

y_pred = logreg.predict(X_test)
```

Figure 4. Source Code Logistic Regression

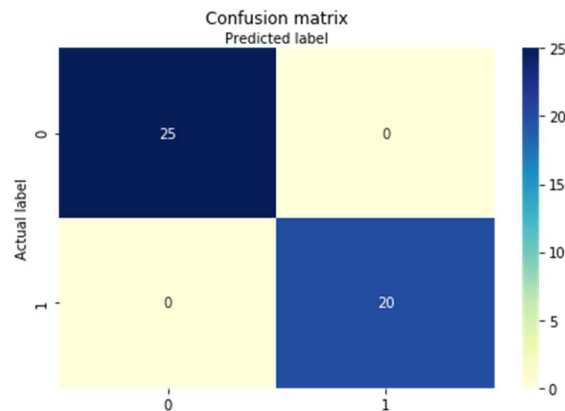


Figure 5. Evaluation of the Logistic Regression Model with the Confusion Matrix

### 3.4. Classification

The objective of the classification conducted in the realm of learning analytics, as per the aim of this investigation, is to anticipate the likelihood of students' success in fulfilling academic evaluations. Various classification techniques, such as Naïve Bayes, Decision Tree, Random Forest, Support Vector Machines (SVM), and Neural Networks, can be employed. The present study aims to demonstrate the utilization of the Naïve Bayes algorithm in forecasting student graduation. Specifically, the Bayes theorem will be employed to estimate the likelihood of a target class, given the available features. The Naive Bayes algorithm is widely recognized as a highly efficient and expeditious classification technique [12]. The model is recognized for its efficacy and ease of use in training and prediction stages.

The Naïve Bayes model uses the same data sharing as this paper's other machine learning models. Based on the tests conducted on 45 data, accuracy and an F1-score of 100% can be obtained. The implementation code for classification with naïve Bayes is shown in Figure 5.

```
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import (
    accuracy_score,
    confusion_matrix,
    ConfusionMatrixDisplay,
    f1_score,
    classification_report,
)

model = GaussianNB()
model.fit(X_train, y_train);

y_pred = model.predict(X_test)
accuracy = accuracy_score(y_pred, y_test)
f1 = f1_score(y_pred, y_test, average="weighted")
print("Accuracy:", accuracy)
print("F1 Score:", f1)
```

Figure 6. Source Code Naïve Bayes

### 3.5. Unsupervised Learning

The utilisation of unsupervised learning in the analysis of learning activity frequency log data has the potential to facilitate comprehension of user behaviour, enable grouping of users exhibiting similar patterns, facilitate identification of anomalous behaviour, and uncover latent relationships or patterns within the data. Unsupervised learning is a technique that enables the identification of concealed patterns or structures within data without the need for prior labelling or categorization. The following are instances of applying unsupervised learning techniques:

1. Clustering: By employing clustering techniques such as K-Means or Hierarchical Clustering, it is possible to categorise students into clusters based on their comparable activity patterns. One possible approach is to categorise students according to their degree of engagement with educational resources or comparable usage tendencies.
2. Anomaly Detection: Anomaly detection algorithms, such as One-Class SVM or DBSCAN, can be employed on learning log data to detect atypical patterns of behaviour. In the event that users exhibit access patterns that deviate significantly from the norm, an anomaly detection algorithm can be employed to facilitate their identification.
3. Dimensionality Reduction: In cases where the data in the learning log exhibits high dimensionality, it is possible to employ dimensionality reduction techniques such as Principal Component Analysis (PCA) or t-SNE to effectively reduce the dimensionality of the data while retaining crucial information. This approach has the potential to simplify data visualisation and streamline subsequent analytical processes.

4. Association Rule Mining: Algorithms such as Apriori or FP-Growth can be utilised to detect relationships or associations among items in learning log data. An instance of this would be the identification of correlations among educational resources that users frequently access in conjunction.

### 3.6. K-Means Clustering

The K-means model was built using 180 data sets without using the Status variable, the target variable in the supervised learning method. The number of K is determined using the Elbow method. In the Elbow method, K values are varied from 2 to 8. Figure 7 shows the results of the Silhouette score to determine the best K value. Based on the Elbow method, the selected K value is 5.

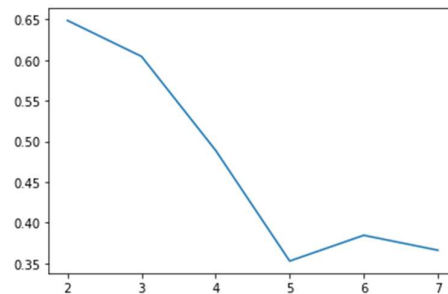


Figure 7. Choosing the Best Number of Clusters (K)

After finding the best K value, model testing was conducted to produce five clusters using the same features as in the previous model development. Cluster information for each index can then be added to the dataframe to show the relationship between the cluster and the target variable. Because the K-Means model is built using multiple features, to be able to represent it in two dimensions, dimension reduction must be carried out using PCA. Figure 9 shows the results of K-Means clustering with two dimensions.

```
cluster_df = pd.DataFrame(newdf,
                          columns=['SCORM', 'Video', 'Forum', 'Asesmen', 'Status'])
#develop model
kmeans = KMeans(n_clusters=5)
#clustering
y = kmeans.fit_predict(cluster_df[['SCORM', 'Video', 'Forum', 'Asesmen']])
#add Cluster column to dataframe
cluster_df['Cluster'] = y

### reduce the dimensions
pca_num_components = 2

reduced_data = PCA(n_components=pca_num_components).fit_transform(cluster_df)
results = pd.DataFrame(reduced_data, columns=['pca1', 'pca2'])
#plotting
sns.scatterplot(x="pca1", y="pca2",
                hue=cluster_df['Cluster'], data=results, palette="deep")
plt.show()
```

Figure 8. Source Code K-Means



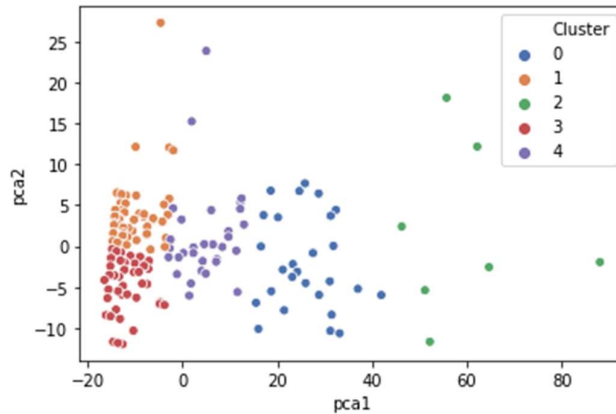


Figure 9. K-means Clustering with Two Dimensions

In order to better understand the information in Figure 9, an analysis of the relationship between clusters and status is carried out, as shown in Figure 10. Figure 10 shows the similarities between clusters 0, 2, and 4, which have status 1 (Pass). In contrast, clusters 1 and 3 consist of students who mostly have status 0 (fail). Furthermore, Table 3 shows the average frequency of activities carried out for each feature of each cluster. Table 2 shows that clusters 1 and 3 with status = 0 have an average value of student frequency in accessing videos, forums, and quizzes that is smaller than clusters 0, 2, and 4. This can mean the frequency of video activities, quizzes, and especially activities in discussion forums can influence student success in completing assessments. This is in line with the correlation test results, where the Forum has the strongest correlation compared to the other variables, which is equal to 0.64. Nonetheless, it should be noted that a higher frequency of forum activity does not necessarily correspond to a higher status pass value. A student's achievement in completing an assessment is not solely determined by a single variable but rather by a combination of various factors. Cluster 3 exhibits a significantly higher frequency value in the Forum feature than the other clusters. However, the number of status passes in Cluster 3 is not higher than those in Clusters 0 and 4.

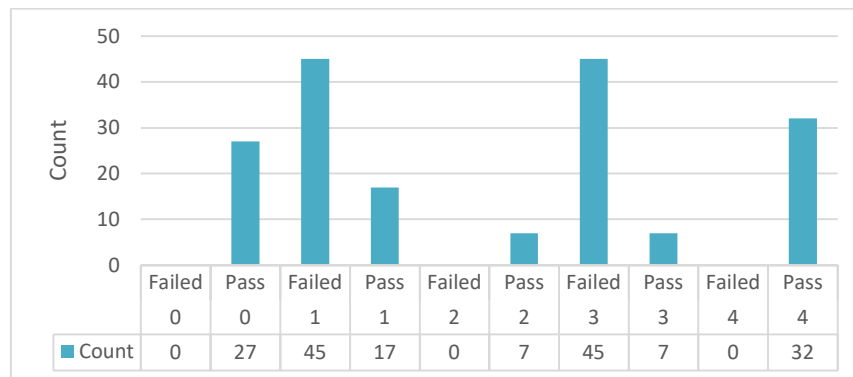


Figure 10. Correlation between Cluster and Target Variable (Status)

Table 2. The Average Value of the Frequency of Online Learning Activities

Cluster	SCORM	Video	Forum	Quiz
0	15.407407	1.185185	41.518519	2.259259
1	15.822581	0.919355	4.790323	1.322581
2	24.142857	1.571429	74.714286	4.571429
3	7.480769	1.019231	4.346154	0.903846
4	15.468750	1.125000	20.218750	3.218750

#### 4. Conclusion

Machine learning can serve as a valuable instrument for conducting analytical learning. Performing analytic learning through machine learning involves multiple stages, including data collection, data preparation, model development, model evaluation, and analysis and interpretation of results. The Logistic Regression method was employed in supervised learning to accurately predict the success status of completing the assessment with a 100% accuracy rate. The phenomenon described is also evident in the Naïve Bayes classification approach, which achieved a perfect accuracy and F1-score of 100%. Moreover, the application of unsupervised learning enables the implementation of clustering techniques to identify distinct clusters of students exhibiting varying levels of activity frequency concerning their graduation status. The analysis of the cluster outcomes indicates a correlation between the frequency of participation in the discussion forum and the achievement of passing status as reflected in the assessment scores. This paper elucidates the function of machine learning in the domain of learning analytics. The forthcoming investigation will encompass a broader range of log data types in conjunction with the activity frequency data employed in the present study.

#### 5. Acknowledge

Acknowledgment addressed to Universitas Udayana as the main funder of this research that has been listed at contract No. B/767/un14.2.8.11/PT.01.03/2021 as Penelitian Unggulan Program Studi (PUPS).

#### References

- [1] C. M. Forsyth, C. Tenison, and B. Arslan, "The current trends and opportunities for machine learning in learning analytics," in *International Encyclopedia of Education (Fourth Edition)*, R. J. Tierney, F. Rizvi, and K. Ercikan, Eds., Oxford: Elsevier, 2023, pp. 404–416. doi: <https://doi.org/10.1016/B978-0-12-818630-5.10050-8>.
- [2] D. Rotelli, G. Fiorentino, and A. Monreale, "Making Sense of Moodle Log Data," *ArXiv*, Jun. 2021, doi: 10.48550/arXiv.2106.11071.
- [3] R. Pratama, R. Herdiana, R. Hamonangan, and S. Anwar, "Analisis Prediksi Kelulusan Mahasiswa Menggunakan Metode Artificial Neural Network," *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 1, pp. 687–693, 2024.
- [4] M. L. Mu'tashim, Z. Ati, and B. S. Yulistiawan, "Klasifikasi Ketepatan Lama Studi Mahasiswa Dengan Algoritma Random Forest Dan Gradient Boosting (Studi Kasus Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jakarta)," in *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, 2023, pp. 155–166.
- [5] S. Liu, X. Wang, M. Liu, and J. Zhu, "Towards better analysis of machine learning models: A visual analytics perspective," *Visual Informatics*, vol. 1, no. 1, pp. 48–56, Mar. 2017, doi: 10.1016/J.VISINF.2017.01.006.
- [6] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. 2007.
- [7] S. Mouggiakou, D. Vinatsella, D. Sampson, Z. Papamitsiou, M. Giannakos, and D. Ifenthaler, "Learning Analytics," in *Educational Data Analytics for Teachers and School Leaders*, S. Mouggiakou, D. Vinatsella, D. Sampson, Z. Papamitsiou, M. Giannakos, and D. Ifenthaler, Eds., Cham: Springer International Publishing, 2023, pp. 131–188. doi: 10.1007/978-3-031-15266-5\_3.
- [8] F. Sciarrone, "Machine Learning and Learning Analytics: Integrating Data with Learning," in *2018 17th International Conference on Information Technology Based Higher Education and Training (ITHET)*, 2018, pp. 1–5. doi: 10.1109/ITHET.2018.8424780.
- [9] R. Ferguson, "The state of learning analytics in 2012: A review and future challenges," *Knowledge Media Institute, Technical Report KMI-2012-01*, 2012.
- [10] N. Kadoić and D. Oreski, *Analysis of Student Behavior and Success Based on Logs in Moodle*. 2018. doi: 10.23919/MIPRO.2018.8400123.
- [11] A. Agresti, *An Introduction to Categorical Data Analysis*, Third edition. Hoboken, NJ: John Wiley & Sons, 2019. [Online]. Available: <http://www.wiley.com/go/wsp>
- [12] A. Wibawa *et al.*, "Naïve Bayes Classifier for Journal Quartile Classification," *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, vol. 7, p. 91, Jun. 2019, doi: 10.3991/ijes.v7i2.10659.