

Exploring Learning Analytics In E-Learning: A Comprehensive Analysis of Student Characteristics and Behavior

Tuti Purwoningsih¹, Wahyu Inayanto², Muhammad Yunus¹

¹³The Faculty of Teacher Training and Education, Universitas Terbuka,
Tangerang Selatan, Indonesia

²The Learning Support Centre, Universitas Terbuka, Tangerang Selatan,
Indonesia

Corresponding email: tuti@ecampus.ut.ac.id

Article history:

Received: April 17th, 2023 Accepted: November 27th, 2023 Published:

December 29th, 2023

Abstract *This article aims to explore learning analytics in e-learning through a comprehensive analysis of student characteristics and behavior. E-learning has become increasingly significant in education, particularly due to the social situation influenced by the pandemic. The Learning Management System (LMS) has become a crucial tool for educators to track and record student learning in e-learning environments. Learning analytics can aid in understanding the context of students, ensuring that they receive a personalized learning experience aligned with learning objectives. However, educators often face challenges in conducting learning analytics for e-learning students, primarily due to the large number of students to analyze and limited data availability. This study seeks to provide a detailed description of e-learning students within the Open and Distance Education (ODE) system. ODE students exhibit high diversity in demographic profiles, learning behaviors, and competency backgrounds. To support this research, we utilize datasets containing student demographic profiles and learning activity data during e-learning sessions. The datasets are obtained from the academic system and LMS log data of Universitas Terbuka. The article employs Exploratory Data Analysis (EDA) and data science approaches as the foundation for predictive and prescriptive analytics of student learning outcomes. Relevant features are extracted from the dataset to build a robust predictive model. The analysis results present patterns and relationships between student characteristics, learning behaviors, and academic*

achievements. This research aims to provide valuable insights for the development of more effective and personalized e-learning strategies to enhance student learning outcomes in the context of distance education. Moreover, the analysis findings can serve as a basis for informed academic decision-making to improve the quality of e-learning environments.

Keywords E-learning, Exploratory Data Analysis (EDA), Learning Analytics, Learning Management System (LMS), the Open and Distance Education (ODE)

INTRODUCTION

In recent years, the adoption of e-learning systems has gained immense significance, particularly in response to the social disruptions caused by the global pandemic. Notably, Moodle, an e-learning platform, has emerged as a prominent choice in Indonesia, with approximately 5,758 institutions actively employing its features (Moodle Statistics, 2023). Like traditional face-to-face learning, e-learning relies on well-defined learning outcomes complemented by appropriate instructional designs. Effective e-learning design necessitates a comprehensive analysis of students and their learning context (Dick et al., 2015). While a small cohort of students might be manageable for such analysis, the landscape changes significantly when dealing with e-learning participants in Open and Distance Education (ODE) programs, characterized by a diverse range of demographics, learning behaviors, and competency backgrounds. Consequently, there arises a crucial need for a specialized approach to discern the characteristics of a substantial number of e-learning participants at ODE.

With advancements in computer-aided learning systems and educational data analysis technology, extensive efforts have been directed towards enhancing learning outcomes (Chatti et al., 2012). In 2011, Siemens introduced the concept of learning analytics (LA) as a powerful tool for unearthing hidden insights and patterns from the vast amount of raw data collected within educational environments (Siemens & Long, 2011). However, LA goes beyond mere data collection; it seeks to derive meaningful interpretations to enrich future learning experiences. Learning analytics employs various analytical methodologies, including descriptive analytics, diagnostic analytics, predictive

analytics, and prescriptive analytics. Among these, descriptive analytics forms a foundational aspect that can be effectively utilized in isolation. This article offers a comprehensive analytical portrayal of student characteristics and behaviors in e-learning settings, leveraging exploratory data analysis (EDA) and data science approaches. The research employs demographic profile data and learning behavior data as key components for driving predictive and prescriptive analytics to optimize student learning outcomes. The dataset utilized in this study encompasses historical records of students engaged in e-learning, comprising demographic profiles, academic backgrounds, and learning activity data meticulously recorded on the Learning Management System (LMS), specifically utilizing Moodle's activity log.

METHODOLOGY

This study explores learning analytics in the context of e-learning by conducting a comprehensive analysis of student characteristics and behavior. The research data used for this study consists of a subset of students enrolled in the Open University (UT) e-Learning class. It encompasses student demographic and academic profile data extracted from the Student Information System (SIM), along with student tutorial activity data obtained from the LMS activity log file. The research design adopted for this study serves as a comprehensive strategy for addressing the research problems, providing a structured approach to research procedures, data collection, and subsequent analysis carried out by the researchers (Leedy & Ormrod, 2015). Figure 1 outlines the systematic stages executed within the context of this research.

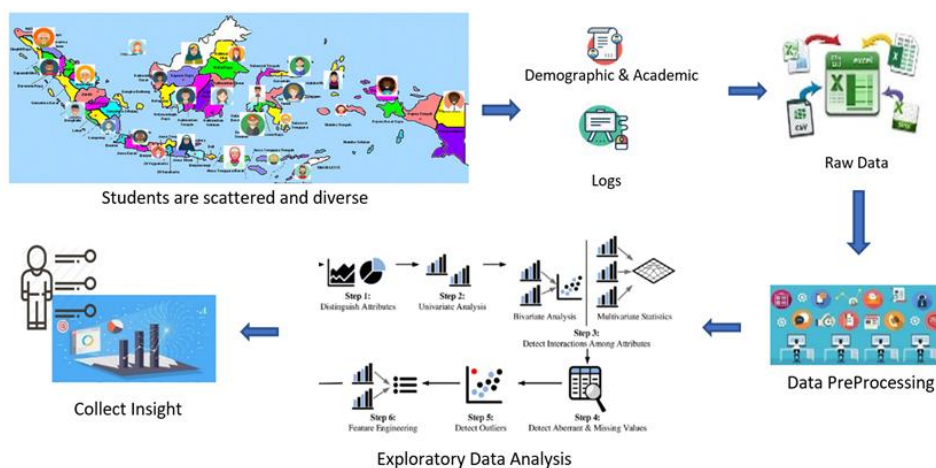


Figure 1. Research Design

The data processing and analysis were conducted using Jupyter Notebook with Python 3.6 programming language. Leveraging the rich dataset of student activity records on the LMS, we gained valuable insights into e-learning students' behaviors and study habits. The research questions primarily centered around providing an analytical description of the diverse profiles and activities of students engaged in e-learning, particularly in Open and Distance Education (PJJ) programs. The data set structure, as shown in Figure 2, facilitated organizing and managing the research data efficiently.

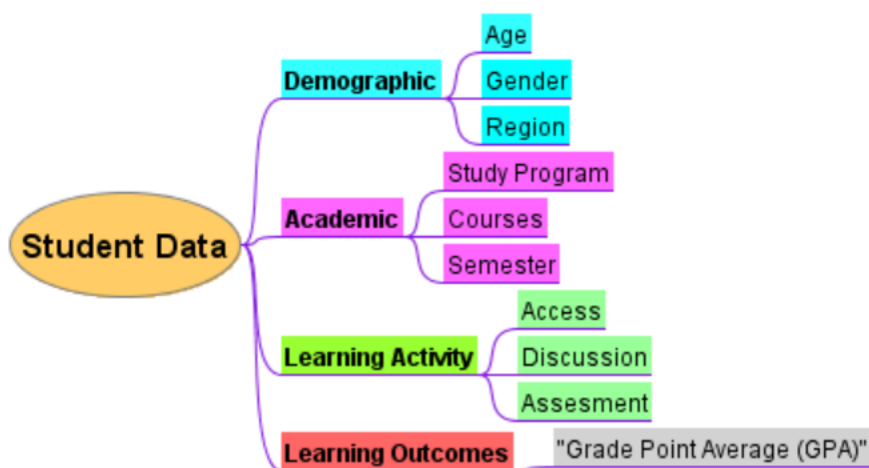


Figure 2. Student Data Structure in Research (Purwoningsih et al., 2020)

The research data for this study was obtained from a subset of students participating in e-learning courses at the Open University (UT). It consisted of two main sources. The first source was the Academic Information System (SIA), a software utilized for presenting academic-related information and managing administrative tasks related to academic activities. The SIA provided valuable demographic and academic profile data of the students. The second data source was the student activity log recorded on the Learning Management System at UT (LMS-UT). This log contained detailed records of students' tutorial activities and engagement within the e-learning platform.

FINDING & DISCUSSION

The study utilized both structured and semi-structured data for analysis. The student demographic and academic profile data were found to be structured data, ready for immediate processing in a fixed format. On the other hand, the LMS-UT logs containing student learning activities belonged to the semi-structured data type, requiring proper description analysis and extraction (Dietric et al., 2015). However, a challenge encountered during data collection was the absence of a centralized data system, resulting in the need for a query system to handle large volumes of scattered data.

The data used for this study focused on students who exclusively participated in Online Tutorials or fully online classes at UT during the periods 2018.2, 2019.1, and 2019.2. The dataset encompassed four main categories of data in CSV format. The analysis of the data revealed several significant findings related to student characteristics and behaviors in e-learning:

- 1) **Diversity in Student Demographics:** The demographic data analysis showed a diverse student population in e-learning, with varying ages, genders, and regional backgrounds. The study found that e-learning attracts learners from different walks of life, including individuals seeking flexible education options due to work commitments or geographical constraints.
- 2) **Learning Behavior Patterns:** Through exploratory data analysis of the LMS activity logs, patterns in student learning behaviors were identified. The study uncovered various engagement levels, study habits, and preferences for accessing e-learning materials. Some students exhibited consistent participation and interaction, while others showed intermittent or limited activity.
- 3) **Impact of Academic Background:** The academic profile data analysis provided insights into the impact of students' previous educational backgrounds on their performance in e-learning. Students with diverse academic experiences displayed different approaches to learning, with some leveraging their prior knowledge effectively and others requiring additional support to bridge knowledge gaps.

- 4) **Predictive Analytics for Learning Outcomes:** By applying predictive analytics, the study attempted to forecast students' learning outcomes based on their demographic profiles and learning behaviors. The analysis aimed to identify factors that contribute to better academic performance and to develop recommendations for improving learning effectiveness in e-learning settings.

To process the data effectively, the demographic and academic achievement data were merged to form comprehensive student profiles. Subsequently, the student learning activities and behaviors were extracted from the LMS-UT log. The extraction process involved calculating the frequency or number of hits for each student's activities recorded in the event column. This preprocessing technique utilized Exploratory Data Analysis (EDA) to gain initial insights and patterns from the dataset.

A. **Data Preprocessing dan Exploratory Data Analysis**

The study involved extensive data preprocessing and exploratory data analysis (EDA) to gain meaningful insights into student characteristics and behaviors in the context of e-learning. The data preprocessing and EDA stages were crucial in preparing the raw data for analysis. The process involved data cleaning, integration, and transformation to ensure data quality and consistency. Various mathematical and statistical techniques were applied to handle missing data, outliers, and inconsistencies, which required significant time and resources to complete successfully. By the end of the preprocessing stage, a refined and reliable dataset was obtained, ready for in-depth analysis.

a) **Data Extraction and Integration**

The study focused on UT e-Learning, specifically online tutorials or "tuton." These tuton activities spanned eight weeks per semester and were conducted fully online, including assignments in weeks 3, 5, and 7. The data extraction process involved collecting student activity data from the Learning Management System (LMS) using the Moodle platform. The extracted data from the activity log file contained essential attributes, providing information

on student interactions, engagement, and performance within the e-learning environment. Table 1 presents a comprehensive description of the attributes found in the Moodle activity log data.

Table 1. Description of Attribute on Moodle Activity Log Data

Atribut	Description
Time	Date and timestamp when the activity was recorded
User full name	Full name of the students and ID
Affected user	Full names of Affected students
Event context	Context of the Activity to which the activity is subjected
Component	Components of an activity affected section
Event name	Activity name according to activity type and class
Description	Activity descriptions that describe ID activity and ID user on Moodle
Origin	Log Record Origin (Client / Web Server)
IP address	IP Address of the Device IP Address that the user uses to log in to the system

The activities recorded in the Moodle log were broadly categorized into four types: viewed, created, updated, and deleted. These activity types were utilized to differentiate between actions performed by teachers and those performed by students. The recorded log data was then compiled into various file formats, such as comma-separated values (csv), Microsoft Excel (.xlsx), HTML Table, JavaScript Object Notation (.json), or Open Document (.ods).

To create data features from the activity log, the "event" attribute was utilized, which provides atomic pieces of information that describe events occurring in Moodle (https://docs.moodle.org/38/en/Events_list_report). Events in Moodle could be the result of user actions or administrative processes/actions performed via the command line. The results were stored in a logging system that records all events and directs them to the plugin's logging storage in a controlled manner. For instance:

- 1) **Teaching** – Activities that usually pertain to what the teacher does and can impact students' learning experiences, such as updating course materials, grading assignments, or transitioning between class stages.
- 2) **Participate** – Activities that are typically related to students' learning experiences, such as participating in forum discussions or submitting assignments.
- 3) **Other** – This category includes actions that have no direct effect on teaching or learning, such as updating calendars, creating user accounts, or viewing messages.

Several studies (Balachandran, 2014; Bravo-Agapito et al., 2021; Figueira, 2017; Hussain et al., 2018; Purwoningsih et al., 2021; Ren, 2019) uses log data for learning improvement in higher education by predicting student success, predicting student performance, providing feedback for faculty, providing recommendations for students, and student modeling. Based on the results of the study of the College Personal Data Form and a review of the systematic literature we determined the features to be used as attributes in this study are as follows.

Table 2. Description of Attribute on Moodle Activity Log Data

Feature	Data	Description
Demographic Profile	Age	Student age
	Gender	Gender
	Profession	Types of student professions
	Region	Student residence areas categorized by archipelago in Indonesia
	Final Education	Students' educational background
	Year of the Last diploma	The year of the last diploma of student education
Academic Profile	Faculty	Faculty of Students
	Study Programs	Student Study Programs categorized into science and social

Feature	Data	Description
	Era	Period or Period of enrolled students
	Semester	Student semester when participating in <i>e-Learning</i>
	Number of online classes	Number of tuton classes attended in the current semester
	Courses	Courses followed in the current semester
	Value tuton	The tuton learning value obtained by students in the current semester
	uas_suo	Final semester exam scores obtained by students in the current semester
	grade	Grade of courses obtained by students in the current semester
Learning Behavior	N_access	Number of student access to LMS-UT
	N_days_access	Number of days students access LMS-UT
	average_hours_access	Average student access hours to LMS-UT
	N_weeks_access	Number of weeks students access LMS-UT
Learning Activities	N_viewed_resources	Frequency of students observing learning resources (<i>material learning</i>)
	N_posted_forum	Number of discussions sent by students on the forum
	N_created_forum	Number of discussions made by students on the forum
	N_viewed_forum	The frequency of students observing discussions on forums

Tuti Purwoningsih, Wahyu Inayanto, Muhammad Yunus
Exploring Learning Analytics In E-Learning:
A Comprehensive Analysis of Student Characteristics and Behavior

Feature	Data	Description
	<i>N_submitted_assignment</i>	Number of assignments sent or uploaded by students on <i>assignments</i>
	<i>N_created_assignment</i>	Number of assignments students create on <i>assignments</i>
	<i>N_viewed_assignment</i>	The frequency with which students observe the status page of assignment submissions on assignments
	<i>N_hits_T1</i>	Number of student hits in week 1
	<i>N_hits_T2</i>	Number of student hits in week 2
	<i>N_hits_T3</i>	Number of student hits in week 3
	<i>N_hits_T4</i>	Number of student hits in week 4
	<i>N_hits_T5</i>	Number of student hits in week 5
	<i>N_hits_T6</i>	Number of student hits in week 6
	<i>N_hits_T7</i>	Number of student hits in week 7
	<i>N_hits_T8</i>	Number of student hits in week 8
Academic Achievement	Semester Achievement Index	The predicate of Student Academic Achievement based on social studies obtained in the current semester (Failed, Graduated, Satisfactory, Very Satisfactory, and With Praise)

The dataset used in this study, as summarized in Table 2, consists of the extraction of the frequency and number of hits of student learning activities recorded at LMS-UT. This dataset is then enriched by integrating it with data encompassing various features of student demographic and academic profiles from the Student Information System, alongside assessment data and the Semester Academic Achievement Index from the Academic Information System, which serves as an indicator of student academic achievement. By combining these diverse sets of data, we aim to gain a comprehensive

understanding of the student's characteristics and behaviors within the e-learning environment.

Relevant studies, such as (Berry, 2017), have reported similar findings, demonstrating that the amalgamation of factors like IP address, age, and years since graduating from a previous education can serve as an accurate differentiator between college students who are likely to successfully complete online courses and those who may encounter difficulties. This integration of data sources enables a more robust analysis, facilitating the identification of significant patterns and trends in student performance and engagement in e-learning settings.

b) Handling Missing Values

To ensure optimal results when processing data with Machine Learning Algorithms, it is essential to address missing values in the dataset. Missing data values can significantly impact the accuracy of conclusions drawn from data analytics if not handled appropriately. In this study, we focused on two methods for handling missing data: data deletion (drop) and data imputation (fill). By comparing datasets that underwent either a missing value drop process or a missing value fill process, we assessed the impact of these approaches on the overall analysis. Each method has its advantages and limitations, and the chosen approach should be based on the specific characteristics of the data and the research objectives.

c) Handling Outlier Data

Detecting and managing outlier data is a crucial step in data preprocessing, especially in learning analytics. In this study, we employed the Interquartile Range (IQR) score, a widely used statistical method for identifying outliers (Yang et al., 2019). The IQR is calculated as the difference between the 75th and 25th percentiles of the data, defining the range within the box-and-whisker plot. Any data point falling below the 25th percentile minus 1.5 times the IQR or above the 75th percentile plus 1.5 times the IQR is considered an outlier. These extreme values can significantly impact the analysis, and proper handling is crucial to avoid distortion of the results. The utilization of Python's

NumPy library facilitated the computation of the IQR and the identification of outliers in the dataset.

Effectively managing missing values and outliers plays a pivotal role in ensuring the integrity and accuracy of the subsequent analysis. The chosen methods for handling these data issues can significantly influence the overall findings and conclusions of the study. Through appropriate data preprocessing techniques, we aim to achieve a comprehensive analysis of student characteristics and behaviors in the e-learning environment, shedding light on key insights that can inform educational practices and enhance learning outcomes.

B. Data Analysis

The data analysis in this study focused on the pre-processed dataset, which has undergone thorough cleaning to ensure data quality. The preprocessing steps involved handling missing values and outliers, as well as cleaning up invalid data based on data categories. Moreover, to accommodate instances where students enrolled in multiple Tuton courses with different periods, we formed super keys by combining student IDs, course IDs, and periods.

a) Univariate Analysis

The univariate analysis served as the initial exploration of the dataset, examining individual variables to gain insights into student characteristics and behavior within the e-learning environment. This analysis represents the simplest form of statistical examination, aiming to describe the data and identify underlying patterns.

For each specific learning period, the dataset includes one row per student, with "student_periode" representing the instance of the concept dataset. To provide an overview of the entire dataset, data aggregation was performed using median statistical techniques. The results of the univariate analysis are summarized in Table 3, presenting essential descriptive statistics.

Table 3. Descriptive Statistics Research Datasets

a) Numeric Data

Feature	Mean	Std	Min	25%	50%	75%	Max
n_module_viewed	50,13	27,22	0	30	47,5	67	213
n_discussion_viewed	1,79	5,16	0	0	0	1,5	136
n_discussion_created	5,54	3,08	0	3	6,5	8	53
n_assignment_viewed	16,07	5,44	0	14	16	18,5	59,5
n_assignment_upload	9,03	4,7	0	5	10	12,5	56
n_assignment_created	2,08	1,16	0	1	3	3	6
n_hits_t1	18,97	17,97	0	0	17	31	187
n_hits_t2	18,07	14,56	0	7	17,5	26	161
n_hits_t3	20,11	14,65	0	10	20	27,5	164
n_hits_t4	15,79	12,12	0	8	16	22	216
n_hits_t5	13,4	12,27	0	0	12	21,5	195
n_hits_t6	14,09	11,74	0	2	15	20	266
n_hits_t7	15,23	12,92	0	0	15,5	23	262
n_hits_t8	10,08	12,67	0	0	7	18	255,5
n_access	162,62	69,61	1	114	160	205	355,5
n_day_access	13,47	6,85	1	8	13	18	50
mode_jam_access	13,87	2,7	0	12,11	13,87	15,66	23
mode_hari_access	3,5	2,06	0	2	4	6	6
access_time_mode	1,86	0,96	0	1	2	3	3
n_week_access	6,65	1,81	1	6	7,5	8	8

Tuti Purwoningsih, Wahyu Inayanto, Muhammad Yunus
Exploring Learning Analytics In E-Learning:
A Comprehensive Analysis of Student Characteristics and Behavior

Feature	Mean	Std	Min	25%	50%	75%	Max
age	25,45	6,24	17	21	23	29	43
distance_year_last_diploma_diploma	6,39	5,69	0	2	5	9	26
semester	2,8	3,16	1	1	1	4	47
n_subject	4,69	2,1	1	3	5	6	10
last_education_name	2,58	1,02	0	3	3	3	3
Gender	0,6	0,49	0	0	1	1	1
Work	3,09	1,91	0	2	3	4	7
cat_region	2,42	2,15	0	1	1	5	7
cat_prodi	0,95	0,23	0	1	1	1	1
faculty	0,95	0,87	0	0	1	2	3
tuton	65	23,61	20	44	74	85	100
assesment	22,6	5,21	0	19	22	25,5	99
Semester_GPA	2,12	0,79	0,08	1,54	2,15	2,69	4

b) Categorical Data

Feature	Unique	Top	Frequency
mode_day_access	7	Monday	29.291
mode_time_access	4	afternoon	35.852
name_education_end	4	Sma	94.168
Gender	2	Woman	66.149
work	8	private	32.750
cat_region	8	javanese	60.604

Feature	Unique	Top	Frequency
cat_prodi	2	social	104.929
faculty	4	FHISIP	41.650
grade	8	B-	35.506
cat_gpa	5	1_gagal	47.143

Table 3 presents a comprehensive descriptive analysis of the pre-processed research dataset, providing key statistical measures for both numerical and categorical data. The dataset consists of 110,902 rows, and the descriptive statistics offer valuable insights into the distribution and central tendencies of the data. For numerical data, the table showcases essential measures such as minimum, maximum, mean, median, and quartiles. Notably, in most numerical features, the mean value exceeds the median, indicating a right-skewed distribution or positive slope. This skewness suggests that certain student behaviors or characteristics may be more prevalent among a subgroup of learners, leading to deviations from a perfectly symmetrical distribution.

In particular, the feature semester GPA which serves as the basis for classifying academic achievement predicates for e-learning students fully online, exhibits a unique comb-shaped distribution or comb distribution. This type of distribution indicates distinct groupings or clusters within the dataset, suggesting that student academic performance may vary significantly across different groups or categories.

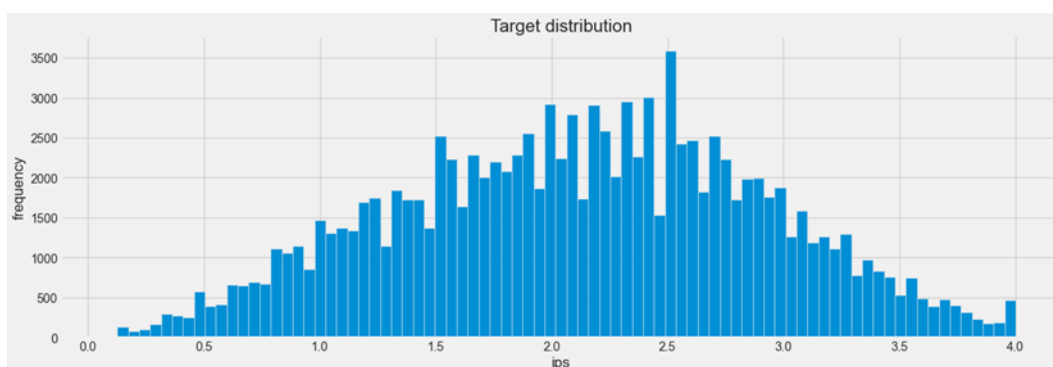
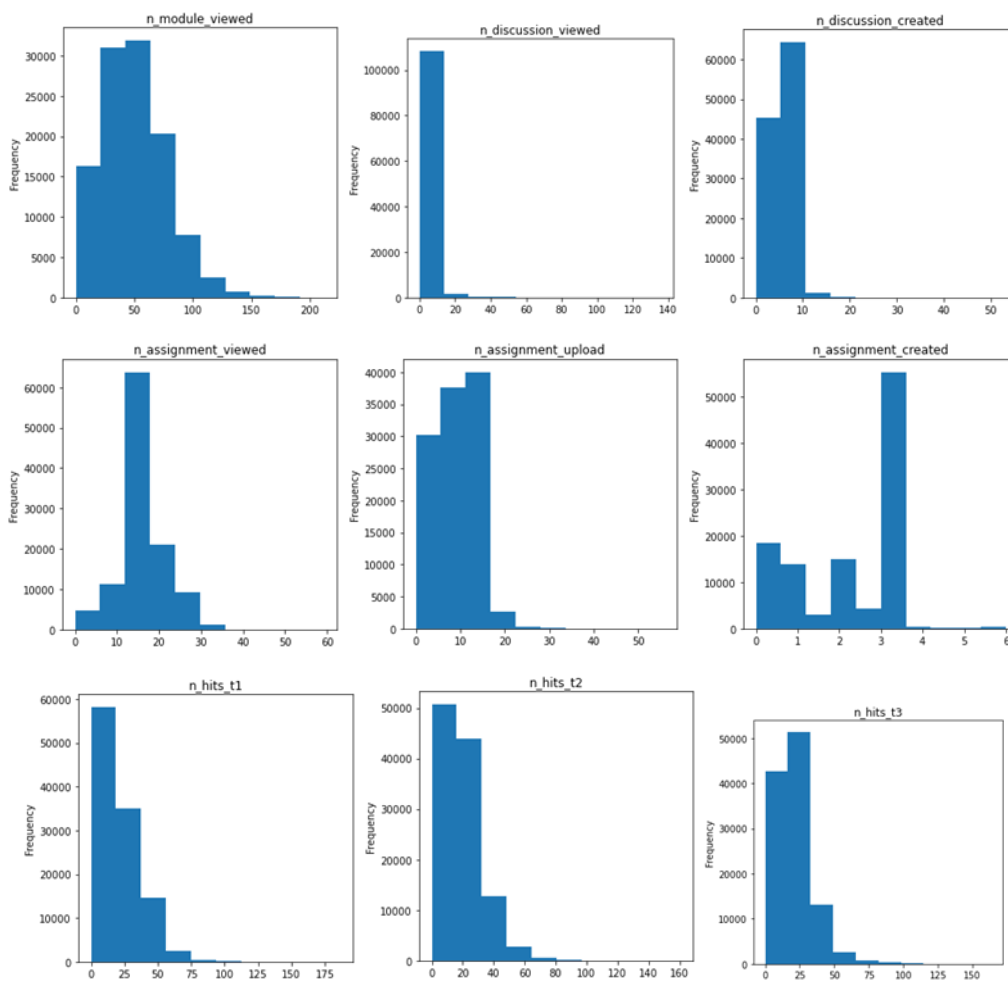


Figure 3. Visualization of semester GPA Distribution in Training Data

Source: Field data

The visualization in Figure 3 provides a clear depiction of the semester GPA, distribution in the training data, further highlighting the presence of comb-shaped clusters. Understanding such patterns can aid educators and institutions in identifying factors that contribute to academic success and designing targeted interventions to support students at various performance levels. Learning analytics plays a crucial role in leveraging these findings to enhance the e-learning experience. By analyzing student characteristics and behaviors, educators can personalize learning pathways, provide timely interventions, and offer targeted support to improve overall learning outcomes.

In addition to the descriptive statistics presented in Table 3, the distribution of other data features is further illustrated in Figure 4. The graph displayed is a histogram, which is a commonly used graphical representation to showcase the frequency distribution of data.



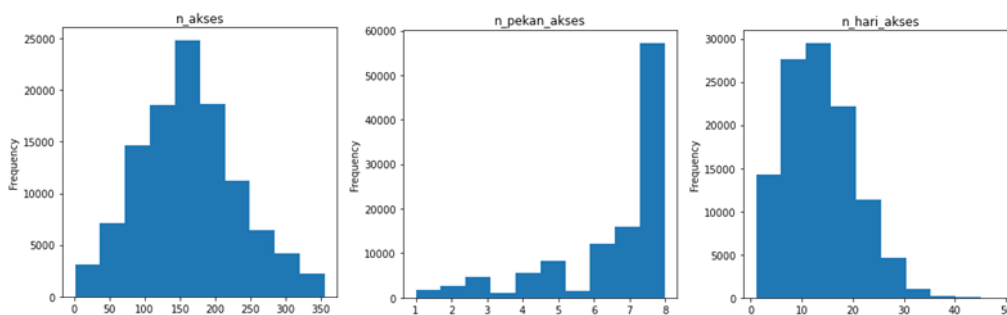


Figure 4. Histogram of Research Data Features

Histograms are valuable tools for understanding the distribution of data within each attribute. By analyzing the histogram, we can identify whether the data follows a normal distribution or exhibits other specific patterns. Our analysis of the histograms in Figure 4 revealed a diverse range of dataset distributions. Some attributes demonstrated normal distribution patterns, while others displayed distinct non-normal distributions. This diversity in data distribution highlights the heterogeneity of student characteristics and behaviors in the e-learning environment.

The insights gained from analyzing the dataset distributions have significant implications for learning analytics. Understanding the distribution of data features enables us to identify potential trends, clusters, or outliers that may influence student performance and engagement. With this knowledge, educational institutions can develop targeted interventions and personalized learning approaches to optimize student outcomes. Moreover, the identification of normally distributed data can help in modeling and prediction tasks, allowing us to apply appropriate statistical methods to gain deeper insights into students' learning patterns and preferences.

The findings obtained through univariate analysis highlight student behavior, learning patterns, and engagement in e-learning platforms. By investigating individual variables, we can uncover valuable insights about student preferences, learning styles, and performance. The implications of univariate analysis extend beyond this research, as the patterns and trends identified can be used to enhance e-learning experiences, personalize learning content, and optimize educational strategies. Educators and institutions can leverage this insight to design more effective online courses and support students in achieving their academic goals.

b) Bivariate Analysis

In this study, we conducted bivariate analysis to explore the relationships between two variables. The bivariate analysis method utilized the correlation coefficient to assess the potential associations between the variables. The main objective of this analysis was to identify any overlapping patterns and connections between the category data and the Semester GPA, which is a numeric type representing students' academic performance. The relationship between the categorical data and semester GPA is visualized using a boxplot, as depicted in Figure 5.

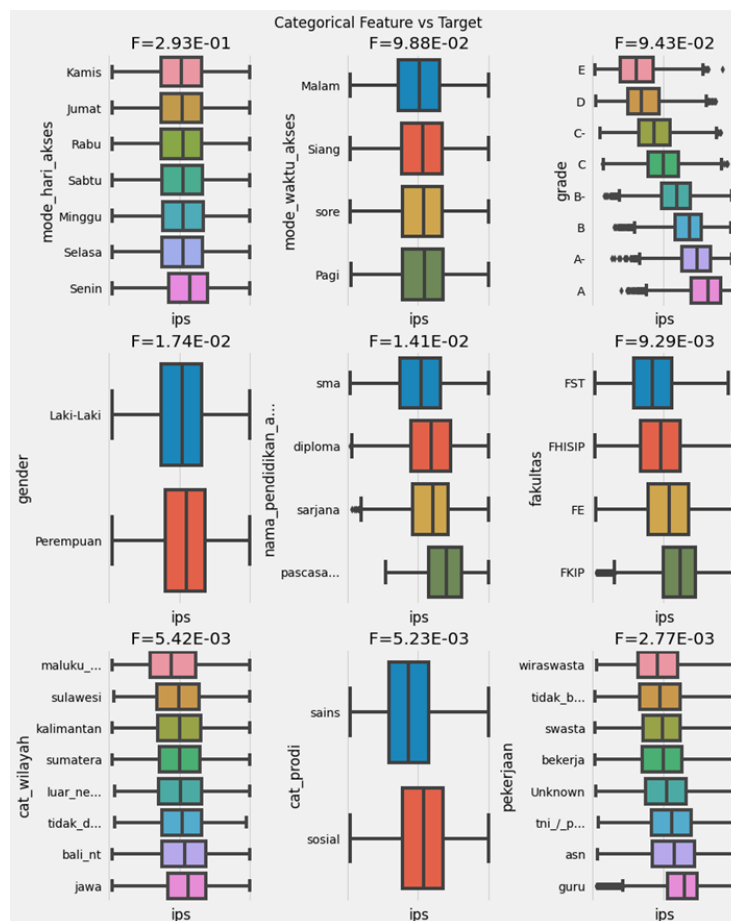


Figure 5. Graph of Relationship between Categorical Data and Numerical Semester GPA

Upon analyzing Figure 5, we observed significant correlations between ten features and academic performance in social studies. These features include:

1. Days of e-Learning Access: Students accessing e-Learning predominantly on Mondays showed a notable correlation with higher academic achievement in social studies.
2. Time of Access: Students accessing the e-Learning platform mostly in the morning demonstrated a positive association with improved performance in social studies.
3. Student Profession: Among various professions, students with a teaching background exhibited the highest average social studies scores.
4. Diploma Background: Students with a diploma educational background showed a strong correlation with better academic performance in social studies.
5. Student-Faculty Category: Students from the Faculty of Teacher Training and Education displayed the highest social studies scores.

These findings shed light on the key factors and behaviours that impact student success in e-Learning environments, specifically concerning academic performance in social studies. Understanding these correlations can inform educational institutions in tailoring instructional approaches and support systems to address students' unique needs and enhance their learning experiences. Moreover, the insights gained from the bivariate analysis can be utilized to develop predictive models for identifying students at risk of academic underachievement. By intervening early and providing targeted support, institutions can foster a conducive learning environment and facilitate the academic growth of their students.

c) Multivariate Analysis

Multivariate analysis is a powerful technique used to explore complex trends and relationships that arise from combinations of attributes in the dataset. In the context of this study, multivariate analysis was applied to understand the dynamics of student attendance and access to e-Learning at UT over the course of 8 weeks, with assignments scheduled in weeks 3, 5, and 7. The student attendance or access to e-Learning is influenced by their level of independence and self-regulated learning (SRL) capabilities. To visualize the patterns of student attendance per week, a box plot was constructed, as shown in Figure 6.

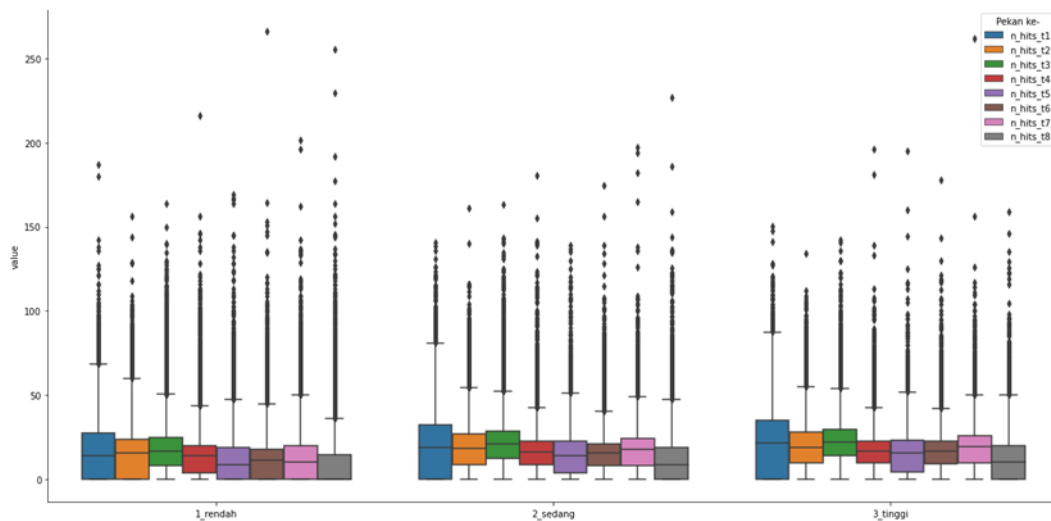


Figure 6. Box Plot of Student Attendance per Week per Class Predicate Academic Achievement of e-Learning Students

Figure 6 represents the distribution of student attendance throughout the e-Learning program via LMS-UT. The box plot provides valuable insights, with the boxes indicating the interquartile range (IQR), the top line representing Q3 (the maximum value), the middle line representing the median, and the bottom line representing Q1. Dots outside the box represent outliers. The analysis of Figure 6 revealed noteworthy observations. Specifically, a significant number of outliers were found in the data of students belonging to the academic achievement group labeled as "low." In this group, some students exhibited no access during weeks 1, 2, 5, 6, 7, and 8. Conversely, in the groups with "medium" and "high" academic achievement predicates, students without access were only observed in weeks 1 and 8. Notably, the number of student attendance showed a substantial increase in their academic achievement class, with students in the "high" predicate group displaying higher attendance rates compared to those in the "low" predicate group.

These findings suggest a strong correlation between academic achievement and e-Learning attendance. Students with higher academic performance demonstrated more consistent and active participation in the e-Learning activities. On the other hand, students with lower academic achievement showed a tendency to have irregular or minimal engagement with the e-Learning platform. Understanding these patterns can assist educators and

institutions in devising strategies to support and motivate students with lower academic performance. By identifying students who may be at risk of disengagement early on, interventions and personalized support can be provided to enhance their learning experience and improve academic outcomes.

C. Featured Engineering

To achieve a robust and accurate prediction model, the researchers leveraged various features extracted from the dataset under investigation (Mubarak et al., 2019). In this study, a feature refers to a numerical representation of raw data. Numerous methods exist to transform raw data into numerical measurements, leading to multiple perspectives in the form of features. Selecting the appropriate features is crucial, as they must be relevant to the task at hand and easily interpretable by the model. Therefore, a systematic process of formulating the most suitable features based on the data, model, and model's objective is employed, known as Feature Engineering. In this study, Feature Engineering was applied to the categorical data features. For instance, the study program category was formulated based on the philosophical division of study programs into those focusing on exact sciences (science) and those concentrating on social sciences (social). The regional category was determined based on the division of internet penetration areas established by the Indonesian Internet Service Providers Association (APJII).

Furthermore, the target data in this study was the predicate of student academic achievement, which was categorized based on the student's Semester GPA. Previous research (Bolliger et al., 2014; Morris et al., 2005) highlights that students with higher IPS tend to excel in navigating online learning environments and demonstrate more successful academic behaviors. It is essential to acknowledge that the Semester GPA can vary every semester or period. Consequently, in this study, the academic achievements of social studies students were grouped based on the social studies scores obtained in the current semester. The process of Feature Engineering plays a pivotal role in preparing the dataset for modeling, ensuring that the selected features offer meaningful insights and contribute significantly to the accurate prediction of student academic achievements. By carefully engineering the features, the

model can effectively capture the underlying patterns and relationships within the data, thereby enhancing the overall performance and reliability of the predictive model.

CONCLUSION

The present study focused on exploring learning analytics in the context of e-Learning, analyzing various aspects of student characteristics and behaviors. The process of exploratory data analysis (EDA) yielded structured datasets that were utilized as features in prediction models. These features were categorized into two main groups: 1) Data related to e-Learning students' learning activities and behaviors; and 2) Data pertaining to students' demographic, academic, and study habits. The academic achievement predicate, as measured by the Semester GPA, was found to be significantly associated with students' demographic and academic profiles, as well as their learning activities and behaviors. The descriptive analysis revealed several dominant features that played a key role in influencing the predicate of student academic achievement. Notably, factors such as the student's profession, faculty, educational background, and level of activeness in submitting assignments emerged as crucial determinants of academic success. Furthermore, students' study habits were found to be characterized by increased access to the LMS-UT on Mondays, particularly during the morning hours.

The findings from this comprehensive analysis of student characteristics and behaviors in e-learning carry substantial implications for educational institutions and e-learning platforms. By gaining insights into the diverse student population and their learning behaviors, educators can adopt a targeted approach to developing instructional strategies, interventions, and support systems that optimize students' learning experiences and academic outcomes. By harnessing the potential of learning analytics, educational institutions can proactively identify students at risk, provide personalized feedback, and implement interventions to enhance student retention and success. Moreover, this comprehensive analysis contributes to the

advancement of learning analytics in the e-Learning domain, empowering educators to make data-driven decisions and continually improve the effectiveness of their e-learning initiatives.

ACKNOWLEDGMENT

The author would like to thank Prof. Zainal A. Hasibuan, Ph.D. (Universitas Dian Nuswantoro) and Harry B. Santoso, Ph.D (Universitas Indonesia). for their guidance and review of the research process. The author also expresses his deepest gratitude to the Learning Assistance Center, the Institute for Research and Community Development, and the Vice-Chancellor for Academic Affairs of the Open University for their support in data collection and insights.

REFERENCES

- Balachandran, A. (2014). *Large scale data analytics of user behavior for improving content delivery (Doctoral thesis, Carnegie Mellon University, 2014)*. Carnegie Mellon University.
- Berry, L. J. (2017). Using learning analytics to predict academic success in online and face-to-face learning environments (Dissertation, Doctor of Education in Educational Technology, Boise State University, 2017) [Boise State University]. In *Boise State University* (Issue May). <https://doi.org/10.1038/132817a0>
- Bolliger, D. U., Inan, F. A., & Wasilik, O. (2014). Development and Validation of the Online Instructor Satisfaction Measure (OISM). *Educational Technology & Society*, 17(2), 183–195. <https://doi.org/10.1080/01587910902845949>; Bourne, J., Moore, J.C., (2005) Elements of quality online education: Engaging communities., http://sloanconsortium.org/publications/books/vol6_summary.pdf, Needham, MA: Sloan Consortium. Retrieved from (Eds.); Bower, B.L., Distance education: Facing the faculty challenge (2001) *Online Journal of Distance Learning Administration*, 4 (2), <http://www.westga.edu/~distance/ojdla/summer42/bower42.html>, Retrieved from; Burden, K., Atkinson, S., Evaluating pedagogical affo
- Bravo-Agapito, J., Romero, S. J., & Pamplona, S. (2021). Early prediction of undergraduate Student's academic performance in completely online learning: A five-year study. *Computers in Human Behavior*, 115(02), 106595. <https://doi.org/10.1016/j.chb.2020.106595>

Tuti Purwoningsih, Wahyu Inayanto, Muhammad Yunus
Exploring Learning Analytics In E-Learning:
A Comprehensive Analysis of Student Characteristics and Behavior

- Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*. <https://doi.org/10.1504/IJTEL.2012.051815>
- Dick, W., Carey, L., & Carey, J. O. (2015). *The systematic design of instruction* (J. Johnston (ed.); Eighth). Pearson Education, Inc., www.ablongman.com.
- Dietric, D., Heller, B., & Yang, B. (2015). *Data science and big data analytics: Discovering, analyzing, visualizing, and presenting data* (C. Long (ed.); 1st ed.). John Wiley & Sons, Inc. www.wiley.com
- Figueira, Á. (2017). Mining Moodle Logs for Grade Prediction. *TEEM 2017, October 18–20, 2017, Cádiz, Spain*, 1–8. <https://doi.org/10.1145/3144826.3145394>
- Hussain, S., Çifçi, M. A., Tamayo, J. D., & Safdar, A. (2018). Big Data and Learning Analytics Model. *International Journal of Computer Sciences and Engineering*, *6*(7), 654–3. <https://doi.org/https://doi.org/10.26438/ijcse/v6i7.654663>
- Leedy, P. D., & Ormrod, J. E. (2015). *Practical research: planning and design* (J. W. Johnston (ed.); Eleventh). Pearson Education Limited.
- Moodle Statistics. (2023). Moodle Statistics. Retrieved from <https://stats.moodle.org/>
- Morris, L. V., Finnegan, C., & Wu, S. S. (2005). Tracking student behavior, persistence, and achievement in online courses. *Internet and Higher Education*, *8*(3), 221–231. <https://doi.org/10.1016/j.iheduc.2005.06.009>
- Mubarak, A., Cao, H., & Zhanga, W. (2022). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*, *30*(8), 1414–1433. <https://doi.org/10.1080/10494820.2020.1727529>
- Purwoningsih, T., Santoso, H. B., & Hasibuan, Z. A. (2020). Data analytics of students' profiles and activities in a full online learning context. *5th International Conference on Informatics and Computing, ICIC 2020*, 1–8. <https://doi.org/10.1109/ICIC50835.2020.9288540>
- Purwoningsih, T., Santoso, H. B., Puspitasari, K. A., & Hasibuan, Z. A. (2021). Early prediction of students' academic achievement: categorical data from fully online learning on machine-learning classification algorithms. *Journal of Hunan University Natural Sciences*, *48*(9), 131–141. <https://doi.org/http://jonuns.com/index.php/journal/article/view/713/710>
- Ren, Z. (2019). *Academic Performance Prediction with Machine Learning Techniques*.
- Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*, *46*(5), 30.

<http://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education>

Yang, J., Rahardja, S., & Fränti, P. (2019). Outlier detection: How to threshold outlier scores? *ACM International Conference Proceeding Series, AIIICC2019, December 2019, Sanya, China*, 1–6. <https://doi.org/10.1145/3371425.3371427>



© 2023 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>).