A Regression Model and a Combination of Academic and Non-Academic Features to Predict Student Academic Performance

Muhammad Arifin [1,2], Widowati Widowati [1], Farikhin Farikhin [1], Gudnanto Gudnanto [3]

[1]Doctoral of Information System Departement, Universitas Diponegoro Semarang, Indonesia
[2]Information System Departement, Universitas Muria Kudus, Indonesia
[3]Counseling Departement, Universitas Muria Kudus, Indonesia

Abstract **– Predicting academic performance provides an effective way for students and faculties to monitor their academic progress. The identification of the most significant features was a key outcome of this research, and the college/university databases from online learning platforms are the main academic data sets used to ascertain performance. However, previous research emphasized the addition of other significant features in the prediction of academic performance. Universities' organizational features include non-academic essential data used in determining student success, but no research has utilized this data to predict student academic performance. Generally, to evaluate binary classification, the relationship between the predicted classifications and the true classifications is analyzed, this approach can lead to the loss of important information from the data. Furthermore, to avoid such loss, this research proposes a regression model by comparing six regression algorithms, and combining academic and non-academic features for prediction student academic performance. Among the various models examined, the gradient-boosted trees regression model demonstrated the lowest error rate. The proposed features have been observed to have a significant impact on academic performance, making them suitable for use in predictions.**

DOI: 10.18421/TEM122-31
https://doi.org/10.18421/TEM122-31

Corresponding author: Muhammad Arifin,
Doctoral of Information System Departement,
Universitas Diponegoro Semarang
Email: arifin.m@umk.ac.id

Received: 29 January 2023.
Revised: 17 April 2023.
Accepted: 08 May 2023.
Published: 29 May 2023.

Keywords **– Regression algorithm, educational data mining, predicting student academic performance, lms, campus organization.**

1. Introduction

Education has a tremendous influence on economic expansion of a nation [1]. Economic growth is also influenced by labour, capital, and technological progress, as well as higher education on a local, state, and national level [2]. Furthermore, higher education also enhances the development of society [3]. Education is required to sustain national development in this era of rapid technological growth. Meanwhile the role of technology in education currently involves all processes including registration, selection, teaching and learning, assessment, payment, and graduation. These processes produce a large collection of data related to students in electronic forms. Furthermore, it is critical for stakeholders to effectively transform these enormous data sets into knowledge that enables teachers, administrators, and policymakers to enhance the quality of decision-making. The quality of the educational process is improved by technology in the provision of timely information to stakeholders [4]. In universities, there are enormous amounts of data that have not been transformed into useful information [5]. Changing these amounts of data renders the extraction of information or knowledge impossible using manual methods. Therefore, a specific method is needed to extract information quickly and precisely. This method is known as Data Mining (DM) and uses algorithms to view past data of the organization. It also locates confidential information that would be difficult to obtain using manual methods [6].

The term Educational Data Mining (EDM) is used to describe the implementation of DM techniques to educational data [7]. This is a growing discipline that expands classical DM methods and develops new techniques for finding data in educational systems [8].

DM evolved at a breakneck pace over the last two decades to enhance data processing for users in line with computer technology which is also developing rapidly [9], [10], [11]. The use of this technique is increasing daily and significantly improves the quality of education [4], [12], [13]. According to Bakhshinategh et al. [14] and Ibitoye et al. [15] it has also been used to evaluate the quality of the learning process. Moreover, there are at least 11 main areas in EDM. One of the popular and oldest areas is student performance prediction which estimates unknown values from variables describing students. The values usually predicted in education are performance, knowledge, scores, or grades [16]. The prediction of academic performance is highly crucial in the educational system as it enables both students and faculties to track progress [17]. In addition, the ability to predict academic performance has the potential to improve educational outcomes. A successful approach to predicting performance can enable educators to allocate resources and tailor instruction more accurately. Furthermore, early prediction enables decision-makers to take appropriate action and implement appropriate learning to improve student success rates [18], [19]. This research identifies the interrelated as well as the most influential features [20], [21]. Student academic performance is the most important component in higher education institutions as they are required to produce skilled graduates with high academic scores [22].

Data on academic performance mostly uses two data sets: college/university databases and online learning platform data [23]. Generally, the Learning management system (LMS) data is used to predict student academic performance, but some believe that other features should be added. Other essential factors include demographic and external assessments of college students, extracurricular activities, high school backgrounds, and social interaction networks [12].

The search for non-academic features that support academic performance has been carried out in previous research. Wolaver et al. [24] focused the analysis on the interaction between alcohol consumption and academic performance among students, while [25] examined the association between academic performance and substance use (alcohol, tobacco, and khat). Romer et al. [26] and Cohn et al. [27] investigated the relationship between attendance and academic performance. Many other features had a more significant influence, including demographic features such as gender and maternal occupation, and pre-enrollment features such as high school grades and university fee discounts [28]. Arifin et al. [29] stated that student involvement in organizations positively influences the acquisition of jobs after graduation.

Pinto et al. [30] observed that combining academic performance with extracurricular activities facilitates entrance into work. Furthermore, Fox et al. [31] discovered that students involved in co-curricular activities earn a higher average point and are more likely to hold leadership positions. Soria et al. [32] showed that extracurricular activities positively influence student leadership and competence development. Baker et al. [33] and Rahman et al. [34] stated that active participation in organizations positively affects academic performance.

Previous research determined approaches to predicting student academic performance. Amrieh et al. [35] introduced a new model, utilizing data mining techniques and incorporating novel data attributes referred to as "student behavior features" which pertain to student interaction with the LMS, was put forth for predicting student performance. Aluko et al. [36] proposed predicting academic performance by utilizing the information contained in previous academic achievements. Helal et al. [37] introduced various classification methods that predict academic performance based on data gathered from student enrollment and the university's LMS generates activity data, while enrollment data comprises student details such as socio-demographic characteristics, method of university entrance (i.e., via entrance exams or without exams), and attendance type (i.e., full-time or part-time). Additionally, Ramaswami et al. [38] obtained data on student interaction with the LMS (Xorro-Q) for one semester and one course with the data divided into two categories: participation in and outside the classroom.

The features that significantly influence student performance prediction vary between researches. Furthermore, the most influential include visited resources, daily attendance, active participation in class, viewing announcements, and parents answering surveys [35]. The most noteworthy feature is the value in mathematics, biology, and physics [36]. Meanwhile, Helal et al. [37] other influential features include gender, presence type, and attendance mode. In addition, Abu Saa et al. [12] included other influential features in the analysis, such as previous grades and class performance, social information, demographics, and e-learning activities. The most influential features on academic performance comprise critical reading, citizen competence, English scores, quantitative reasoning, and biology scores [39]. The features that significantly impact student performance prediction tend to differ between research studies.

The first research question aims to identify the most influential features on students' academic performance, asking: "What features have the greatest impact on students' academic performance?"

Currently, there is no available model for predicting academic performance that combines both academic features (LMS, academic data) and non-academic ones. The second research question aims to investigate whether combining academic and non-academic features can influence students' academic performance. Although various non-academic features can potentially affect predictions, this study is limited to exploring the impact of demography, economics, and campus organization, as these are the non-academic features proposed for analysis.

The researchers utilized a classification model to predict the target feature, which is the GPA, by categorizing it into different classifications such as good and bad [22], [36], pass and fail [40], excellent, very good, good, average, and poor [41], using a method called discretization or binning [42]. However, this approach can lead to the loss of important information from the data [43]. To avoid such loss, the researchers employed a regression model. Suleiman R. and Anane R. [44] compared various regression models, including linear regression (LR), supporting vector regression (SVR), decision trees (DT), and random forests (RF) to predict the CGPA at the end of the year. Arifin et al. [39] also compared six different regression algorithms, namely generalized linear model (GLM), deep learning (DL), DT, RF, gradient boosted trees (GBT), and support vector machine (SVM), to predict academic performance. Additionally, predicting academic performance through comparing LR with DL was conducted by [45]. The third research question of this study aims to identify the most appropriate regression model for predicting academic performance using academic and non-academic features.

In order to address the three research questions stated above, the researcher will calculate the weights of all variables used in this study and compare them with the findings of previous studies. In addition, to answer the second research question, the researcher will analyze the weights of both academic and non-academic features. Meanwhile, in order to address the third research question, the researcher will compare several regression models that have been used by previous researchers.

2. Related works

A comparison of data mining algorithms on EDM datasets has been done by [46]. They compared the C4.5 algorithm with Naive Bayes (NB) and used the 10-fold validation method.

A comparison of the methods shows that the NB method is better than C4.5. Amrieh et al. [35] compared multiple models, and selected the most appropriate.

The classification algorithms used include Decision Tree (DT), NB, and Artificial Neural Network (ANN) which applied ensemble methods (bagging, boosting, and RF). The DT ensemble method with boosting was the best with an accuracy rate of 82.2%. It was observed that the behavioural features of accessed sources were most significant in predicting academic performance. Meanwhile, Aluko et al. [36] proposed the models of Support Vector Machine (SVM) and Logistic Regression (LR). Accuracy-wise, the SVM model outperformed the LR model. The results also showed that previous academic performance is a good predictor of future.

Hellas et al. [20] proposed different classification method to predict academic performance using data gathering from student enrollment and activity data generated from the university's LMS from 2011 to 2013. The data obtained from enrollment includes details about students, such as their socio-demographic characteristics, mode of university entrance (through entrance exams or without exams), and attendance type (full-time or part-time). Additionally, to monitor student involvement with online learning activities, LMS data are collected. When creating prediction models, the research's study of student heterogeneity was an essential contribution, as students with distinct socio-demographic characteristics or learning styles may have diverse learning motives. Furthermore, experiments showed that enrollment features and lecture activities help to identify vulnerable students more precisely. The four algorithms of NB, J48, SMO, and JRip sufficiently predicted student academic performance. The experiments showed that no singular method for predicting student performance is superior in all aspects. The combination of J48 and JRip contributed significantly by producing intelligible outputs in the form of trees and rules, respectively.

Ramaswami et al. [38] collected data on student interaction with the LMS (Xorro-Q) for one semester and one course as well as divided the data into two categories: participation in and outside the classroom. Afterwards, the NB, k-Nearest Neighbor (KNN), LR, and RF algorithms were compared to determine the best accuracy for predicting academic performance. The results showed that the participation feature outside the classroom had a good but insignificant impact. Furthermore, the RF algorithm was the best among other algorithms.

The algorithms used to predict student academic performance differ between research.

Arifin et al. [39] compared several algorithms commonly used which include five Generalized Linear Models (GLM), Gradient-Boosted Tree (GBT), DT, Deep Learning (DL), RF, and SVM.

The findings indicated that GBT was the most successful in predicting performance, with the least RMSE. Furthermore, to determine the best algorithm for specific research, a comparison is required. Previous research also highlighted that the best algorithms vary based on the research carried out. There is no optimal algorithm for, and this is influenced by the data obtained during preprocessing. Hence, researchers must compare algorithms before using them in predicting academic performance.

Conijn et al. [47] studied 17 hybrid courses containing 4,989 students with the Moodle LMS log in order to forecast the outcome using classification (pass/fail) and regression (GPA value) models. Student performance was successfully predicted during the first ten weeks. The accuracy improved slightly during the first week, with a significant improvement after week 5, when task grades became available. In the fifth week of the study, the regression model displayed an R2 adjustment of 0.43 and the binary classifier achieved an accuracy of 67% in the third week.

Gerritsen used Moodle log files for 17 subjects to predict the success or failure in a specific subject (binary classification) [48]. The perceptron multilayer model did the best out of the seven classifiers and selected the students who were at risk 66.1% of the time.

3. Methodology

This section holds great significance in a research paper as it outlines the research process and the methodologies employed to gather, visualize, model, and analyze data. Figure 1 indicates the main steps in the proposed methodology.

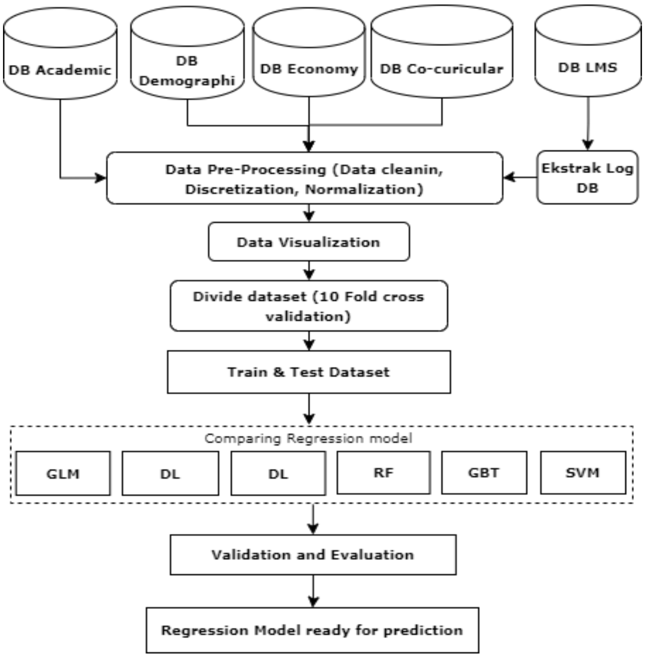

DB Academic
DB Demographi
DB Economy
DB Co-curicular
DB LMS
Data Pre-Processing (Data cleanin, Discretization, Normalization)
Ekstrak Log DB
Data Visualization
Divide dataset (10 Fold cross validation)
Train & Test Dataset
Comparing Regression model
GLM
DL
DL
RF
GBT
SVM
Validation and Evaluation
Regression Model ready for prediction

Figure 1. Research Steps

3.1. Data aggregation and preparation

This research combined academic data, namely activities obtained from the University LMS (Moodle) for one semester and GPA data. In combination, the non-academic data is of a demographic character consisting of gender and place of residence, economic information, and data on student organization activities. Table 1 represents the meanings of the various activity characteristics. Furthermore, Moodle records were extracted and combined using these data. To experiment with this data set, GPA was selected as the target column and a general reference for determining academic performance.

3.2. Data preprocessing

Student data was extracted from the Moodle LMS, and data from multiple sources were aggregated and filtered according to specific criteria. LMS records were taken for 19 weeks (one semester starting from February to July), with an average of each day producing a total of 199,700 records from 8,500 active students participating in lectures.

Information on academic data and demography were obtained from the Academic Information System (SIA), while economic data was obtained from the registration of new students. The co-curricular data obtained from the student affairs department was in the form of extractive results of organization decrees located at the university. Furthermore, data with inconsistent values were deleted, for example, students with academic data but no LMS record; students with a GPA below 1; students with very little activity level in the LMS, etc. A total of records from 4435 student data sets were consequently evaluated in this research.

Table 1. Features and Descriptions

Category	Feature	Description
Academic	Students number	Students ID
	CGPA	Semester GPA (Vulnerable grades from 1-4)
Demography	Gender	Gender, Male or female
	Domicile	Student residential address
Economy	Parents income	The amount of parental income per month
Co-curricular	Organization	Participation in campus organizations (Number 1, 2,3 etc.)
LMS	All student activities in accessing the LMS were recorded for analysis	Logins, access forum, access to teaching materials, assignments, questionnaires, and entries to the course.

3.3. Data visualization

One of the critical preprocessing tasks is data visualization, which employs graphical representation to simplify complex data. The use of visualisation techniques has been applied increasingly to depict various aspects of online learning. By leveraging graphical representations, instructors can gain a better understanding of their learners and gain insight into learning patterns. Figure 2 showed the data set based on gender features which included 1,814 males and 2,621 females, a total of 4435. In this research, both academic and non-academic data were visualized.

3.4. Model

This research introduces a predictive model of academic performance by applying the regression model, a learning approach that uses linear values as targets. This is a more appropriate model as students' final CGPA value at the end of the semester is used as the predictor value in the form of numbers instead of classes. After preprocessing, the five most frequently used regression models were compared. The best model with the lowest error rate is then selected and used to predict academic performance. Furthermore, a comparison of algorithms on EDM data has also been made by [49] to ascertain the best algorithm.

3.5. Validation and evaluation

In this research, k-fold cross-validation technique was utilized to assess the effectiveness of machine learning algorithms. One of the principal benefits of the k-fold cross-validation is that all data points are used for training, testing, and validating the algorithm [49]. The data set was randomly divided into 80% samples for training and 20% for algorithm testing and evaluation. A variety of evaluation metrics like MSE, MAE, and RMSE were employed for assessment. Moreover, the algorithm was trained and evaluated on the entire dataset by randomly separating and evaluating it five times.

4. Results and discussion

This section aims to provide a comprehensive understanding of the study's outcomes and offer an in-depth discussion of the implications and significance of the results.

4.1. Results of dataset collection

Academic data consists of CGPA for the current semester, there are a total of 10,044 active students.

The demographic and economic data were obtained from new student registration for two batches, namely 2019, with 2998 students, and 2020, with 2622 students. Furthermore, a total of 1400 students are active in campus organizations, and some students participate in more than one organization. This data was obtained from the student affairs department through the rector's decree from each organization. The LMS data was obtained from the Moodle LMS with nine columns, where each column contains an overview of the learning activity data in the LMS. The column description is listed in Table 2.

4.2. Data preprocessing

CGPA data were selected and classified: CGPAs with a value of 0 or below were deleted because it signifies that the student has either left school or is a final student. The data has been filtered based on student batches, specifically the 2019 and 2020 batches. The final dataset, consisting of 4,435 entries, will be utilized for predicting student academic performance.

The university categorized the economy into seven categories: >10.000.000, 7.500.000-10.000.000, 5.000.000-7.500.000, 2.500.000-5.000.000, 1.000.000-2.500.000, 500.000-1.000.000, and 100.000-600.000. Meanwhile, domicile data consisting of 57 variants, are further classified into five: the first group is student cities, the second includes the cities adjacent to student cities, the third contains a group of cities that are one city apart and in one area, the fourth represents cities outside the territory, and the fifth group is cities outside the island. According to organizational data, 4 participated in four organizations, 19 in three organizations, 155 in two organizations, and 1012 in one organization.

The LMS log data was extracted into 16 variables in line with [50]. According to the results of the LMS log extraction, three variables are worth 0. Therefore, the LMS variables used in this research are 13.

Table 2. Field log of LMS

Field	Description
Time	Date and time of the event. Eg: 6/06/20, 14:03
User full name	Name of the user
Affected user	Which user gets affected by this task
Event context	Under which item was the event performed, Eg: Course Name
Component	Name of the component, Eg: Quiz, Test
Event name	Action/event performed by the user
Description	More details about the action/event performed
Origin	Origin of the event Eg: Web
IP address	From which IP address is the event generated

4.3. Data visualization

Academic CGPA data ranged from 1 to 4, with an average of 3.5 and a standard deviation of 0.5. Figure 2. showed the proportions of male and female students, where the number of women was more than that of men.

Figure 3. showed that parents of students had a major income between 1,000,000–2,500,000 and were ranked third. This majority comprised a total of 1832 parents out of 4435. Therefore, the income factor was not balanced across the economic groups. The domicile feature in Figure 4 showed that students living in rings 1 and 2 were the highest compared to other rings.

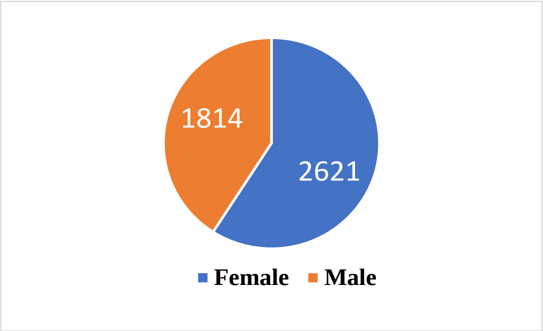


Figure 2. Proportions of male and female students

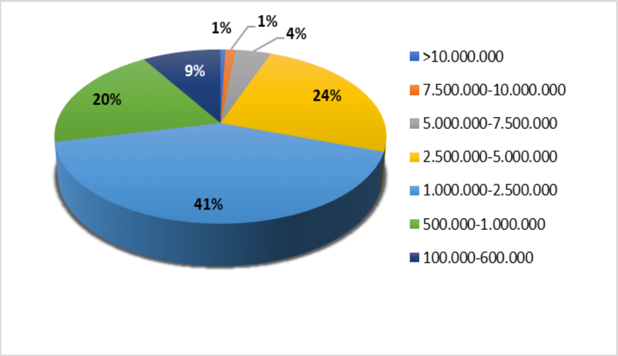


Figure 3. Visualization of economic features

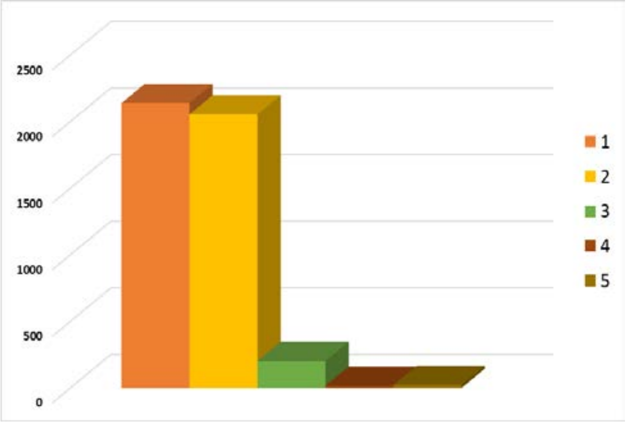


Figure 4. Domicile Visualization

Figure 5 showed the co-curricular features of students who were not members of campus organizations. More than 77% were not members, 20% followed only one organization, and 3% followed two organizations. However, fewer students participated in organizations than those who do not.

The LMS log data visualization in Figure 6 showed that the N_entries_course variable had the most significant number, followed by the Total_assignments variable. Simultaneously, the n_questionnaires_submitted variable had the smallest amount.

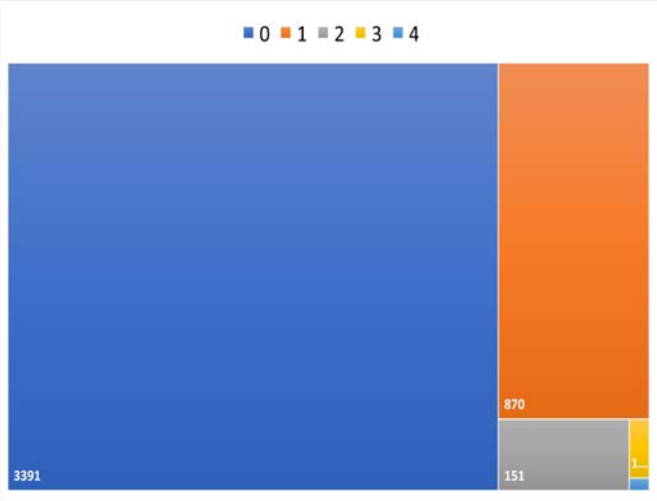
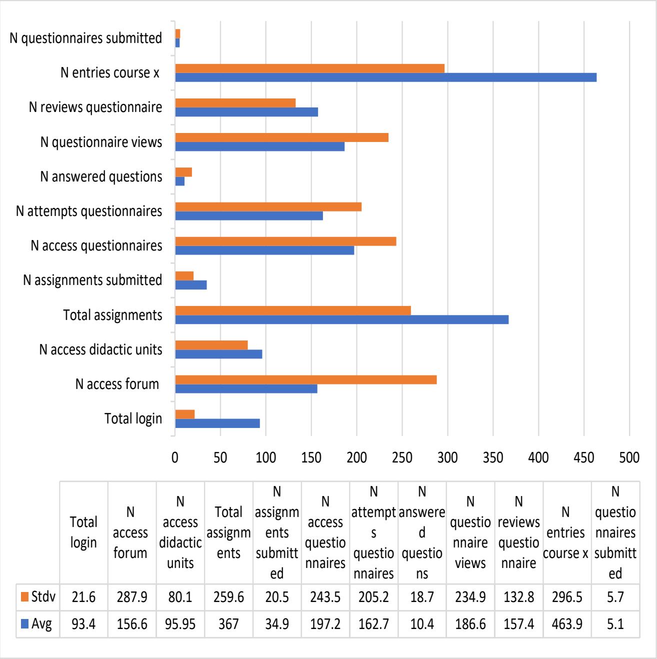


Figure 5. Features of student organizations



	Total login	N access forum	N access didactic units	Total assignments	N assignments submitted	N access questionnaires	N attempts questionnaires	N answered questions	N questionnaire views	N reviews questionnaire	N entries course x	N questionnaires submitted
Stdv	21.6	287.9	80.1	259.6	20.5	243.5	205.2	18.7	234.9	132.8	296.5	5.7
Avg	93.4	156.6	95.95	367	34.9	197.2	162.7	10.4	186.6	157.4	463.9	5.1

Figure 6. LMS Log feature visualization

4.4. Evaluation of the model

After comparing the experimental results of several regression models with the processed data, it was found that the GBT regression algorithm had the lowest error rate and the highest correlation when compared to the other models.

Table 3. Regression model comparison

Model	**RMSE**	**AE**	**RE**	**SE**	**Corr**
Generalized Linear Model	0.446	0.32	0.09	0.20	0.50
Deep Learning	0.413	0.29	0.09	0.17	0.61
Decision Tree	0.462	0.32	0.09	0.21	0.51
Random Forest	0.408	0.30	0.09	0.17	0.63
Gradient Boosted Trees	**0.379**	**0.26**	**0.08**	**0.14**	**0.68**
Support Vector Machine	0.434	0.29	0.08	0.19	0.56

In Table 3, the prediction of academic performance using regression models can be applied effectively in predicting student academic performance with a value of RMSE 0.38. This regression model can also be used as a reference using the target variable regression.

Figure 7 showed the weight of each variable, where the most influential variable in the prediction of student academic performance was N_assignments_submitted, followed by total login. Furthermore, the proposed variables, namely domiciliation, campus organization, and economy, ranked 4th, 6th, and 8th out of the 17 existing variables.

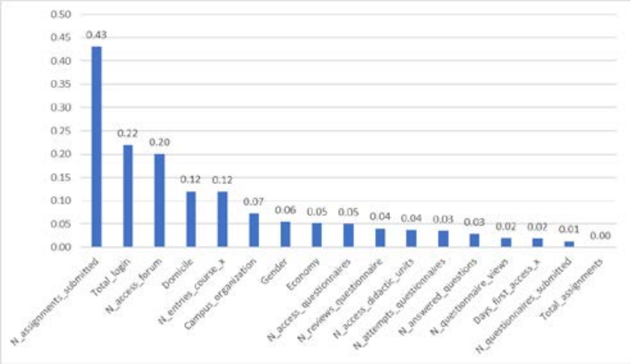


Figure 7. Variable weight values

4.5. Validation

Validation evaluates the acceptability of numerical results in the measurement of predicted relationships between variables. The k-fold cross-validation was the popular technique used, and operates in the following manner:

1. Randomly divide a dataset into k approximately equal folds.
2. Select one-fold to act as the holdout set. Fit the model on the remaining k-1 folds and evaluate its performance on the held-out fold using metrics such as test MSE and MAE.
3. Repeat the process k times, with each fold serving as the holdout set exactly once.
4. Compute the average of the k-test MSEs, which represents the overall test MSE.

Upon comparing the regression algorithms, it was found that the gradient boosting tree algorithm outperformed the other five algorithm. Validation was carried out on the selected algorithm using cross-validation with a value of k=5, and the Table 4 outlines the outcomes.

Table 4. Model validation using cross-validation

K	MSE	MAE
1	0.148763	0.268474
2	0.141609	0.262551
3	0.159986	0.273623
4	**0.135967**	**0.260845**
5	0.169836	0.279075

Each validation's MSE and MAE values have a small vulnerable value. However, the fourth validation has the smallest value, which is MSE: 0.136 and MAE: 0.260.

5. Discussion

Based on the research questions, the experimental results were examined

RQ1. What features have the greatest impact on students' academic performance?

Based on the results, the gender feature is commonly used, as it has a fairly good influence on the academic performance of students [35], [37], [51], [52]. The LMS is the most commonly used feature [37], [50], [48], [40], [47], [52], [53], although other researchers apply economic features [39], [49]. According to Abu Saa et al. [12], e-learning activity features were ranked third, while in this research, this feature was first. Furthermore, the average co-curricular feature is used to determine the extent of the effect of this feature on the waiting period for graduates to acquire a job [29], facilitate the world of work [30], obtain a high GPA [31], positively affect competence and leadership [32], and have a good influence on academic performance [33], [34].

RQ2. Does the combination of academic and non-academic features influence academic performance?

The feature analysis showed that non-academic features like campus organization influence the prediction of student academic performance. In light of this, it is possible that this feature could be leveraged in the future to accurately predict academic performance and could even be integrated with other academic features to further enhance accuracy. Different research reports show varying influence values for the same features. Helal et al. [37] concludes that Gender characteristics were the most influential, although in this research, the feature ranked seventh, [12] demographic features were ranked fourth in the same position as this research, and [13] economic features were at the thirteenth level, although in this research, this feature was ranked eighth.

Economic, domicile, and campus organizational features have unbalanced values between each other. In the future, further exploration is required to determine the influence of these features on the prediction of student academic performance.

RQ3. What is the regression model most appropriate for predicting the academic performance?

Several regression models have been studied to determine the best model for predicting student academic performance. Researchers in this study have compared six regression models, commonly used by previous researchers to predict academic performance, to address RQ3. The experiment's findings show that the GBT regression model has the smallest RMSE value of 0.374%, AE of 0.258%, and SE of 0.140%. Based on these results, the GBT model can be deemed as the most appropriate regression model for predicting academic performance using academic and non-academic features, effectively answering RQ3.

6. Conclusion

According to the results and subsequent discussion, it is evident that both academic and non-academic features proposed in this research significantly influence the prediction of academic performance. A conclusion can be drawn from these findings. Active involvement in organizations influences performance by 7%, gender by 6%, geographical location by 12%, and economic factors by 5%. Meanwhile, the most influential factor is the LMS log of N_assignments_submitted variables of 42%, and total_login of 22%. These features can be combined with LMS features because this complements each other, and nothing dominates between one feature and another.

Although GBT sequentially creates an ensemble of shallow trees, with each tree learning from the previous one and improving the overall performance. As a result, GBT achieves the lowest error rate. Although shallow trees are weak predictive models, they can be improved through proper tuning to form a powerful "committee", is difficult to beat with other algorithms. This should be explored in further research and tuned with hyperparameters to obtain a lower level of accuracy.

Acknowledgements

We express our deepest gratitude to Diponegoro University for supporting this research with valuable laboratory and human resources. Without the support and assistance from this institution, we will not be able to achieve the expected results in a short time. We also thank all those who have played a role in helping to complete this research. Thank you to Muria Kudus University for providing the funding for this research.

References

[1]. Deda, E., Pacukaj, S., & Vardari, L. (2021). Education and its role in the economic development of the country and government policies to be undertaken to increase the quality of education, the case of albania. *Journal of Educational and Social Research*, *11*(1), 188–199.
Doi: 10.36941/jesr-2021-0018

[2]. Labini, M. S. (2015). Higher education and economic welfare. In: *The University and the Economy: Pathways to Growth and Economic Development, February*, 28–46.
Doi: 10.4337/9781782549499.00010.

[3]. Pinheiro, R., Wangenge-Ouma, G., Balbachevsky, E. & Cai, Y. (2015). The Role of Higher Education in Society and the Changing Institutionalized Features in Higher Education. In: Huisman, J., de Boer, H., Dill, D.D., Souto-Otero, M. (eds.) , 225-242, The Palgrave International Handbook of Higher Education Policy and Governance. Palgrave Macmillan, London.

[4]. Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers and Education*, *113*, 177–194.
Doi: 10.1016/j.compedu.2017.05.007.

[5]. Morris, L. V. (2016). Mining Data for Student Success. *Innovative Higher Education*, *41*(3), 183–185. Doi: 10.1007/s10755-016-9367-6.

[6]. Kumar, M., & Salal, Y. K. (2019). Systematic review of predicting student's performance in academics. *International Journal of Engineering and Advanced Technology*, *8*(3), 54–61.
Doi: 10.13140/RG.2.2.26667.69923.

[7]. Baker, R. S. J. D., & Yacef, K. (2009). The State of Educational Data Mining in 2009 : A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3–16.

[8]. Alsuwaiket, M. (2018). *Measuring Academic Performance of Students in Higher Education Using Data Mining Techniques*. [Doctoral dissertation, Loughborough University].

[9]. Romero, Cristobal, & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), 1–21. Doi: 10.1002/widm.1355.

[10]. Maphosa, M., & Maphosa, V. (2020). Educational data mining in higher education in sub-saharan africa. *Proceedings of the 2nd International Conference on Intelligent and Innovative Computing Applications*, 1–7. Doi: 10.1145/3415088.3415096.

[11]. Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A Systematic Review of Deep Learning Approaches to Educational Data Mining. *Complexity*, *2019*. Doi: 10.1155/2019/1306039.

[12]. Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques. In *Technology, Knowledge and Learning 24*. Springer Netherlands. Doi: 10.1007/s10758-019-09408-7.

[13]. Al-Emran, M., Mezhuyev, V., Kamaludin, A., & Shaalan, K. (2018). The impact of knowledge management processes on information systems: A systematic review. *International Journal of Information Management*, *43*(2), 173–187.

[14]. Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, *23*(1), 537–553. Doi: 10.1007/s10639-017-9616-z.

[15]. Ibitoye, A. O. J., Borokini, B., & Alabi, J. O. (2019). Knowledge Based Performance Evaluation and Predictive Model for Undergraduate Students. *Asian Journal of Research in Computer Science*, *2*(3), 1–7.

[16]. Romero, Cristbal, & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *40*(6), 601–618. Doi: 10.1109/TSMCC.2010.2053532.

[17]. Ragab, M., Abdel Aal, A. M. K., Jifri, A. O., & Omran, N. F. (2021). Enhancement of Predicting Students Performance Model Using Ensemble Approaches and Educational Data Mining Techniques. *Wireless Communications and Mobile Computing*, *2021*. Doi: 10.1155/2021/6241676.

[18]. Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1). Doi: 10.1186/s41239-020-0177-7.

[19]. Al-Ashoor, A., & Abdullah, S. (2022). Examining Techniques to Solving Imbalanced Datasets in Educational Data Mining Systems. *International Journal of Computing*, *21*(2), 205–213. Doi: 10.47839/ijc.21.2.2589.

[20]. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: A systematic literature review. *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE*, 175–199. Doi: 10.1145/3293881.3295783.

[21]. Romero, Cristóbal, Ventura, Sebastian Pechenizkiy, M., & Baker, R. S. J. (2010). *Handbook of Educational Data Mining*. CRC Press.

[22]. Jenitha, T., Santhi, S., & Monisha Privthy Jeba, J. (2021). Prediction of Students' Performance based on Academic, Behaviour, Extra and Co-Curricular Activities. *Webology*, *18*, 262–279. Doi: 10.14704/WEB/V18SI01/WEB18058.

[23]. Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences*, *11*(9). Doi: 10.3390/educsci11090552.

[24]. Wolaver, A. M. (2002). Effects of heavy drinking in college on study effort, grade point average, and major choice. *Contemporary Economic Policy*, *20*(4), 415–428. Doi: 10.1093/cep/20.4.415.

[25]. Mekonen, T., Fekadu, W., Mekonnen, T. C., & Workie, S. B. (2017). Substance Use as a Strong Predictor of Poor Academic Achievement among University Students. *Psychiatry Journal*, *2017*, 1–9. Doi: 10.1155/2017/7517450.

[26]. Romer, D. (1993). Do Students Go to Class? Should They? *Journal of Economic Perspectives*, *7*(3), 167–174. Doi: 10.1257/jep.7.3.167.

[27]. Cohn, E., Cohn, S., Hult, R. E., Balch, D. C., & Bradley, J. (1998). The Effects of Mathematics Background on Student Learning in Principles of Economics. *Journal of Education for Business*, *74*(1), 18–22. Doi: 10.1080/08832329809601655.

[28]. Saa, A. A. (2016). Educational Data Mining & Students' Performance Prediction. *International Journal of Advanced Computer Science and Applications*, *7*(5), 212–220. Doi: 10.14569/ijacsa.2016.070531.

[29]. Arifin, M., Widowati, W., & Farikhin, F. (2022). Using Education Data Mining (EDM) and Tracer Study (TS) Data as Materials for Evaluating Higher Education Curriculum and Policies. *KnE Social Sciences*. Doi: 10.18502/kss.v7i14.11948.

[30]. Pinto, L. H., & Ramalheira, D. C. (2017). Perceived employability of business graduates: The effect of academic performance and extracurricular activities. *Journal of Vocational Behavior*, *99*, 165–178. Doi: 10.1016/j.jvb.2017.01.005.

[31]. Fox, L. M., & Sease, J. M. (2019). Impact of co-curricular involvement on academic success of pharmacy students. *Currents in Pharmacy Teaching and Learning*, *11*(5), 461–468. Doi: 10.1016/j.cptl.2019.02.004.

[32]. Soria, K. M., Werner, L., Chandiramani, N., Day, M., & Asmundson, A. (2019). Cocurricular Engagement as Catalysts Toward Students' Leadership Development and Multicultural Competence. *Journal of Student Affairs Research and Practice*, *56*(2), 207–220. Doi: 10.1080/19496591.2018.1519439.

[33]. Baker, C. N. (2008). Under-represented college students and extracurricular involvement: the effects of various student organizations on academic performance. *Social Psychology of Education*, *11*(3), 273–298. Doi: 10.1007/s11218-007-9050-y.

[34]. Rahman, S. R., Islam, M. A., Akash, P. P., Parvin, M., Moon, N. N., & Nur, F. N. (2021). Effects of co-curricular activities on student's academic performance by machine learning. *Current Research in Behavioral Sciences*, *2*, 100057. Doi: 10.1016/j.crbeha.2021.100057.

[35]. Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. *International Journal of Database Theory and Application*, *9*(8), 119–136. Doi: 10.14257/ijdta.2016.9.8.13.

[36]. Aluko, R. O., Daniel, E. I., Shamsideen Oshodi, O., Aigbavboa, C. O., & Abisuga, A. O. (2018). Towards reliable prediction of academic performance of architecture students using data mining techniques. *Journal of Engineering, Design and Technology*, *16*(3), 385–397. Doi: 10.1108/JEDT-08-2017-0081.

[37]. Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D. J., & Long, Q. (2018). Predicting academic performance by considering student heterogeneity. *Knowledge-Based Systems*, *161*, 134–146. Doi: 10.1016/j.knosys.2018.07.042.

[38]. Ramaswami, G., Susnjak, T., Mathrani, A., Lim, J., & Garcia, P. (2019). Using educational data mining techniques to increase the prediction accuracy of student academic performance. *Information and Learning Science*, *120*, 451–467. Doi: 10.1108/ILS-03-2019-0017.

[39]. Arifin, M., Widowati, Farikhin, Wibowo, A., & Warsito, B. (2021). Comparative Analysis on Educational Data Mining Algorithm to Predict Academic Performance. *Proceedings - 2021 International Seminar on Application for Technology of Information and Communication: IT Opportunities and Creativities for Digital Innovation and Communication within Global Pandemic, iSemantic 2021*, 173–178. Doi: 10.1109/iSemantic52711.2021.9573185.

[40]. Karalar, H., Kapucu, C., & Gürüler, H. (2021). Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. *International Journal of Educational Technology in Higher Education, 18*(1). Doi: 10.1186/s41239-021-00300-y.

[41]. Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, *8*, 55462–55470. Doi: 10.1109/ACCESS.2020.2981905.

[42]. Oreski, D., Visnjic, D., & Kadoic, N. (2022). Discretization of numerical meta-features into categorical: analysis of educational and business data sets. *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology, MIPRO 2022 - Proceedings*, 1179–1184. Doi: 10.23919/MIPRO55190.2022.9803574

[43]. García, S., Luengo, J., Sáez, J. A., López, V., & Herrera, F. (2013). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, *25*(4), 734–750. Doi: 10.1109/TKDE.2012.35.

[44]. Suleiman, R., & Anane, R. (2022). Institutional Data Analysis and Machine Learning Prediction of Student Performance. *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 1480–1485. Doi: 10.1109/CSCWD54268.2022.9776102

[45]. Hussain, S., Gaftandzhieva, S., Maniruzzaman, M., Doneva, R., & Muhsin, Z. F. (2021). Regression analysis of student academic performance using deep learning. *Education and Information Technologies*, *26*(1), 783–798. Doi: 10.1007/s10639-020-10241-0.

[46]. Santoso, J. T., Ginantra, N. L. W. S. R., Arifin, M., Riinawati, R., Sudrajat, D., & Rahim, R. (2021). Comparison of Classification Data Mining C4.5 and Naïve Bayes Algorithms of EDM Dataset. *TEM Journal*, *10*(4), 1738–1744. Doi: 10.18421/TEM104-34.

[47]. Conijn, Rianne, Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS. *IEEE Transactions on Learning Technologies*, *10*(1), 17–29. Doi: 10.1109/TLT.2016.2616312.

[48]. Gerritsen, L. (2017). *Predicting Student Performance with Neural Networks*. [Master thesis, Tilburg University, Netherlands]. Retrieved from: http://arno.uvt.nl/show.cgi?fid=143628 [accessed: 15 February 2023].

[49]. Abdullah, S. A., & Al-Ashoor, A. (2020). An Artificial Deep Neural Network for the Binary Classification of Network Traffic. *International Journal of Advanced Computer Science and Applications*, *11*(1). Doi: 10.14569/IJACSA.2020.0110150.

[50]. 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*, 106595. Doi: 10.1016/j.chb.2020.106595.

[51]. Al-Sudani, S., & Palaniappan, R. (2019). Predicting students' final degree classification using an extended profile. *Education and Information Technologies*, *24*(4), 2357–2369. Doi: 10.1007/s10639-019-09873-8.

[52]. Santoso, L. W., & Yulia. (2020). Predicting student performance in higher education using multi-regression models. *Telkomnika (Telecommunication Computing Electronics and Control)*, *18*(3), 1354–1360. Doi: 10.12928/TELKOMNIKA.v18i3.14802.

[53]. Conijn, R. (2018). Predicting student performance in a blended MOOC. *Journal of Computer Assisted Learning*, *34*(5), 615–628. Doi: 10.1111/jcal.12270