Journal of Innovations in Teaching and Learning, 2024, Vol. 4, No. 1, 7-12
Available online at http://pubs.sciepub.com/jitl/4/1/2
Published by Science and Education Publishing
DOI:10.12691/jitl-4-1-2

Modeling Students' Procrastination Using Gaussian – Bernoulli Mixed Naïve Bayes Method

Shara M. Baylin*, Laila S. Lomibao

College of Science and Technology Education, University of Science and Technology of Southern Philippines, Cagayan de Oro City, Philippines
*Corresponding author: shara.baylin@ustp.edu.ph

Received December 16, 2023; Revised January 18, 2024; Accepted January 25, 2024

Abstract Academic procrastination appears to be widespread among university students. They delay in writing and turning in their assignments, presentations, and other academic requirements. Academic performance suffers as a result of incomplete and delayed assignments. This paper investigated the use of educational data mining techniques for the early detection of academic procrastination tendencies in online mathematics learning. Data are collected first-hand by the researchers using the research instruments. The features selected in this study were data that can be easily gathered even in the first week of school, and thus fits more for a model to predict as early as possible. The study was conducted among mathematics education students at the University of Science and Technology of Southern Philippines – CDO Campus. The K-means clustering method was used in the study to group students who procrastinate and Filter-based methods, particularly Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), Chi-square, and ReleifF were used to identify the most informative features that contribute most significantly to student performance, learning outcomes, or other relevant educational metrics. The results revealed that 5 of the identified features are relevant in predicting students' procrastination tendencies and the Gaussian-Bernoulli Mixed Naïve Bayes model can successfully predict students' later procrastination behaviors with a testing accuracy of 85% and Kappa score of 82%. Mathematics educators and administrators can use this model to predict and prevent the negative consequences of students' academic procrastination.

Keywords: *academic procrastination, educational data mining, online learning, procrastination*

Cite This Article: Shara M. Baylin, and Laila S. Lomibao, "Modeling Students' Procrastination Using Gaussian – Bernoulli Mixed Naïve Bayes Method." *Journal of Innovations in Teaching and Learning*, vol. 4, no. 1 (2024): 7-12. doi: 10.12691/jitl-4-1-2.

1. Introduction

Procrastination is a common phenomenon affecting individuals across various domains, including the academe. Despite the fact that definitional unanimity has been challenging to achieve regarding the idea of academic procrastination, there are numerous definitions in the literature. Academic procrastination refers to the tendency to delay or postpone academic tasks [1], leading to reduced productivity, increased stress levels [2], and compromised educational outcomes [3]. University students that procrastinate seem to postpone and put off their academic work, becoming self-excusive and ignoring their academic responsibilities for the duration of their studies. It appears to be a widespread tendency for university students to put off their academic work. They delay in writing and turning in assignments and presentations, finishing projects, and even preparing for major examinations [1]. It could be deliberate, accidental, or habitual, but it has an enormous impact on university students' learning and success. Nonetheless, procrastination comes in a variety of forms, as described by scholars [1] which includes realistic, unrealistic, and spiritual procrastination; chore, dream; behavioral, decisional and meta-cognitive procrastination.

A sudden shift in learning modality, due to the COVID-19 pandemic, incurred challenges to students that adversely impacted their completion of tasks, motivation to continue studying, and overall academic performance [4]. Several studies have already investigated the impact of online and distance learning on students [5,6] and found that lack of motivation and effort [7] among students is significantly higher in online learning. Procrastination then has become a prevailing problem since the onset of online learning. Based on observation, roughly 40% – 60% of students in the University of Science and Technology of Southern Philippines – CDO Campus had either missed submitting some or all of the tasks or do not submit at all. As a matter of fact, the degree of procrastination is amplified in an online setting [5] and academic procrastinators are less inclined to self-regulate, which will have a negative effect on performance [5,7,8]. Furthermore, while past and recent studies on online learning resulted to useful prescriptions for academic procrastinators, the research in academic procrastinators among university students during the conduct of online or

hybrid classes is still emerging. A study was conducted by Godinez and Lomibao [9] in St. Rita's College of Balingasag among junior high school students and identified 14 features out of the 35 identified features that are relevant in predicting students' procrastination tendencies.

This research aims to explore academic procrastination among students, shedding light on the underlying causes, consequences, and potential interventions. By addressing this knowledge gap, the study seeks to provide valuable insights for educational institutions, policymakers, and stakeholders to develop effective strategies for mitigating academic procrastination. To conduct this, the researchers will utilize Educational Data Mining (EDM) used in the study of Godinez and Lomibao [9]. Educational Data Mining is an emerging discipline that connects mining and its application to education. Many predictive and classification models have been made with EDM techniques. However, most of these predictive models focus on predicting features with students' overall performance and grades. Instead of these identified predicting features, the researcher would like to specifically predict the teachers' and students' procrastination tendencies.

The researchers hope to use the built model to support school administrators and teachers in improving the quality of their instruction in the conduct of online or hybrid learning in mathematics and by taking into account the suggested intervention activities for students identified by the model according to the degree of their procrastination.

2. Methods

This research employed educational data mining techniques to extract potentially important hidden patterns from diverse data sets. The K-means clustering method was used in the study to group students who procrastinate, and Filter-based methods, particularly Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), Chi-square, and ReleifF, offer powerful tools for identifying the most informative features that contribute most significantly to student performance, learning outcomes, or other relevant educational metrics.

This study was restricted to the University of Science and Technology of Southern Philippines – CDO Campus' faculty and students with Mathematics courses. The features selected in this study are data that can be easily gathered even in the first week of school. A survey questionnaire was sent by the researcher via email. The researcher used closed questionnaires by Likert Scale. The researcher used the Pure Procrastination Scale, Procrastination Scale for Students, Multidimensional Perfectionism Scale, Satisfaction with Life Scale, The Overall Attitudes Towards Online Learning Questionnaire, and the Time Management Scale. Figure 1 below provides a visual presentation that depicts the flow of the procedures that went into creating the model.

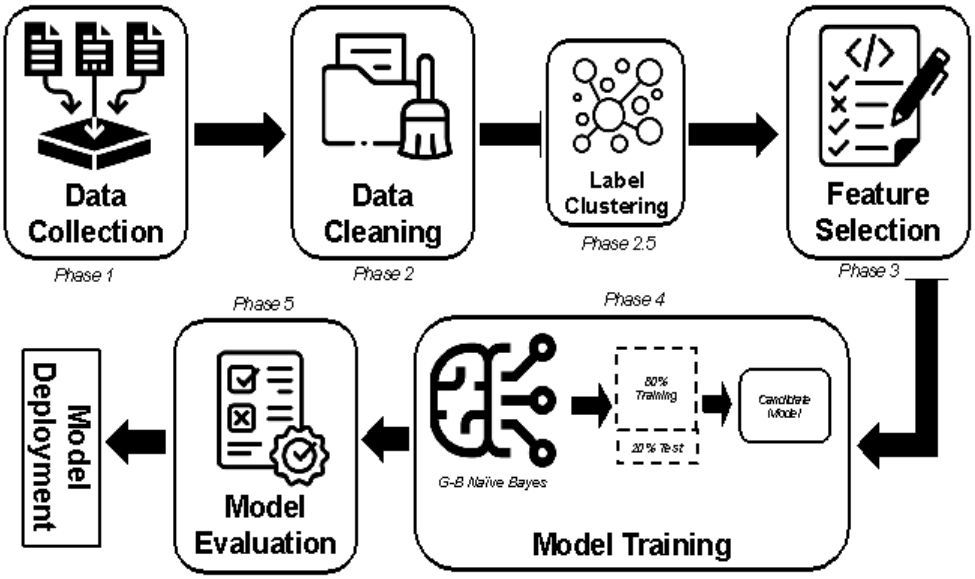

Figure 1. Methodological Framework for Model Building

2.1. Research Procedure

Phase 1: Sampling and Data Collection

Since there was no available, pre-existing data on the university's repositories related to students' procrastination and to the features set in this study, data are collected first-hand by the researchers using the research instruments. A total of 308 students were the respondents in this survey through purposive sampling. Students that are enrolled during the first semester of SY 2023-2024 were sent the questionnaires via email. The features selected in this study were data that can be easily gathered even in the first week of school, and thus fits more for a model to predict as early as possible. Data on students' demographics, such as year level, gender, age, number of course units, and Mathematics grades in the previous semester year are considered 'institutional data' that can be easily retrieved from the school records, while features including Math Anxiety, Overall Attitudes Towards Online Learning, Time Management, Multi-dimensional Perfectionism, and Procrastination were

collected from the students using the research instruments. The study was conducted in accordance with the principles of informed consent and data privacy. All students' data were collected with the explicit consent of the students, and all data were stored securely and in accordance with the institutional policies.

Phase 2: Data Cleaning and Transformation

The study prioritized data quality by employing a thorough pre-processing phase. This phase involved identifying and removing missing values to mitigate potential bias. Additionally, categorical data was transformed into numerical representations through appropriate encoding techniques.

Phase 2.5: Output Data Clustering

After collecting students' procrastination data during the data collection phase, students' answers were then employed with a Machine Learning method called K-Means Clustering to identify which among our students can be branded as Non-procrastinators, Low Procrastinators, Moderate Procrastinators, and High Procrastinators. K-means clustering has emerged as a valuable tool for analyzing educational data, offering insights into patterns and hidden structures within student performance, engagement, and learning behaviors. This allows researchers to identify distinct student clusters with potentially shared characteristics, paving the way for personalized learning strategies and targeted interventions [10]. For instance, K-means clustering has been successfully applied to segment students based on their academic performance, revealing groups struggling with specific concepts or excelling in particular areas [11]. By analyzing the features associated with each cluster, educators can pinpoint potential factors influencing performance and tailor instructional approaches accordingly. Additionally, K-means has been used to group students based on engagement patterns, identifying those at risk of disengagement and allowing for the development of targeted interventions to enhance motivation and participation [12]. The procrastination data were separated into 4 clusters to identify students that were considered non-procrastinators, low procrastinators, moderate procrastinators and high procrastinators. After the input features were structured and the output data were identified, the two data were merged in one dataset ready for the feature selection phase.

Phase 3: Feature Selection

Feature selection plays a crucial role in extracting meaningful insights from educational data, often plagued by dimensionality and irrelevant information. Filter-based methods, particularly Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), Chi-square, and ReleifF, offer powerful tools for identifying the most informative features that contribute most significantly to student performance, learning outcomes, or other relevant educational metrics.

Information Gain, measuring the reduction in entropy after splitting data based on a feature, and Gain Ratio, normalizing IG for feature variations, are popular choices for prioritizing features that effectively discriminate between different student groups [13]. Gini Decrease, focusing on the purity increase within each data partition, excels in identifying features that best separate students based on their performance or engagement levels [14]. Chi-square, assessing the statistical dependence between features and target variables, is valuable for uncovering features significantly associated with specific educational outcomes [15]. Finally, ReleifF, a multivariate technique, identifies features that differentiate instances belonging to the same class while simultaneously distinguishing instances belonging to different classes, proving effective in discovering relevant features for personalized learning or student clustering tasks [16].

By employing these filter-based methods, researchers can significantly reduce the dimensionality of educational data, improve the efficiency of subsequent analysis, and gain deeper understanding of the factors influencing student success. However, it is crucial to consider the limitations of each method, such as IG's susceptibility to bias towards features with more values, and carefully evaluate the selected features through further analysis to ensure their suitability for the specific research question [17].

Thus, in this study, filter-based feature selection methods were utilized by the researcher to extract the most valuable information from existing data, to improve the predictive power of the model.

Phase 4: Model Training using Mixed Gaussian – Bernoulli Naïve Bayes Algorithm

The two Naïve Bayes algorithm variations will be employed in this study to handle mixed data because our input features include both continuous and binary data. In a MapReduce setting, two approaches for managing mixed data for the Naïve Bayes model were examined in a study by [18]. Data discretization is the first approach used in this study. The first approach uses the discretized values to build the NB model once the continuous values have been discretized. Discretizing the continuous data is an additional pre-processing step needed for this option. This additional step could require a lot of time and resources. For continuous values, the second approach makes use of the probability density function of the Gaussian distribution. This model, also known as the Mixed NB model, was able to handle discrete and continuous variables. The chance of discrete values was estimated using the Multinomial distribution, and the probability of continuous values was estimated using the Gaussian distribution. There is no need for a pre-processing step with this procedure according to the study of [18]

Phase 5: Model Evaluation

Over two thirds (80%) of the data were divided at random and used as the training dataset. Twenty percent (20%) of the remaining set was designated as the test set, which was then utilized to assess the model. The data must be trained before using a data mining method to develop a prediction model. In addition, a 10-fold cross validation was utilized to attain optimal training accuracy. The dataset is divided into ten equal subsamples by the test in a 10-fold cross validation. The remaining 10–1 subsamples are used for training, and one subsample was retained for data validation. Until all subsamples had been used for validation, this process was repeated. Then, using confusion matrix analysis, the model with the highest training accuracy in the 10-fold cross validation was verified.

A Confusion Matrix Analysis (CMA) displays instances that are real and those that classifiers anticipate. A CMA table is typically used in supervised learning to display an algorithm's performance. According to [19] the number of occurrences of an exact class is shown by its rows, whereas the number of occurrences of a projected class is indicated by its columns. Since there are four class labels in this study—Non-procrastinator, Low-procrastinator, Moderate-procrastinator, and High-procrastinator—Table 1 displays the 4-class confusion matrix that was employed.

Table 1. The Confusion Matrix Analysis

Confusion Matrix Analysis		PREDICTED			
		Non Procrastinator	Low Procrastinator	Moderate Procrastinator	High Procrastinator
ACTUAL	Non Procrastinator	A	B	C	D
	Low Procrastinator	E	F	G	H
	Moderate Procrastinator	I	J	K	L
	High Procrastinator	M	N	O	P

True Positives | True Negatives | Misclassified Cases

The accuracy, precision, and recall of the system can then be evaluated based on the entries in the CMA matrix. Accuracy per class is calculated by dividing the number of instances correctly classified as belonging to the actual class by the total number of instances in that class. Sensitivity is calculated as the number of positive predicted instances divided by the total number of predicted instances. Conversely, sensitivity can be calculated by dividing the total number of projected cases by the true negatives. The following formulae express specificity, sensitivity, and accuracy:

$$Accuracy_{class} = \frac{TP_{class} + TN_{class}}{Positives + Negatives}$$

$$Sensitivity_{class} = \frac{TP_{class}}{TP_{class} + FN_{class}}$$

$$Specificity_{class} = \frac{TN_{class}}{TN_{class} + FP_{class}}$$

3. Results and Discussion

Clustering Students' Procrastination

To properly label students' procrastination, 308 college students answered the Procrastination Assessment Scale – Students [20], a 44-item instrument that quantified students' the prevalence of academic procrastination, the reasons for academic procrastination, and compared scores on the PASS with behavioral indices of procrastination and other related constructs. A K-means Clustering was employed to identify and categorize the procrastination tendencies of these students. Table 3 below shows the distribution of students in terms of their procrastination tendencies.

The K-modes Clustering, with re-runs set to 10 and a maximum iteration set to 300, categorized 46 Non Procrastinators (14.97%), 97 Low Procrastinators (31.49%), 80 Moderate Procrastinators (25.97), and 85 High Procrastinators (27.6) from the 308 total students. The four clusters of procrastination tendencies were also suggested to be the best number of clusters according to the K-means clustering.

Table 2. Clustering Results of Students' Procrastination Tendencies

Procrastination Tendency	Frequency	%
Non Procrastinators	46	14.94
Low Procrastinators	97	31.49
Moderate Procrastinators	80	25.97
High Procrastinators	85	27.6
Total	308	100

Feature Selection

Nine variables or input features made up the dataset of this study after the data cleaning and transformation phase were conducted. These are – Year Level, Age, Sex, Number of Units/Course Load, Average Grade in Math, Math Anxiety, Overall Attitude Towards Online Learning, Time Management, Multi-dimensional Perfectionism scores. To reduce possible redundant data or unnecessary predictor variables, many filter-based feature selection methods were employed – Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), Chi-square, and ReleifF – to identify only the more important input or predictor variables and thus increase the overall accuracy of our model. From the nine predictor variables, only the highest five predictor variables were pushed for the model building process.

Thus, the predictor variables that the Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), Chi-square, and ReleifF considered to be more important and the only input features that was considered in the model training were: MDP or The Students' Multi-Dimensional Perfectionism (1), Time Management (2), Math Anxiety (3), Number of Units/Course Load in the 1st Sem SY 2023-2024 (4), and Year Level (5). Table 3 shows the list of all the features and the decisions to whether confirm or reject.

Table 3. Result of the Feature Selection Process

	IG	GR	GINI	χ^2	ReliefF	Decision
MPS	0.179	0.089	0.061	52.775	0.031	Confirm
TMQ	0.093	0.047	0.031	25.872	0.009	Confirm
MASA	0.061	0.031	0.022	14.436	0.013	Confirm
Number of Units/Course Load 1st Sem SY 2023-2024	0.039	0.020	0.013	4.824	0.002	Confirm
Year Level	0.037	0.019	0 013	7.041	0.004	Confirm
Average Grade in Math (Previous Sem)	0.035	0.018	0.012	5.582	0.002	Reject
OATOL	0.029	0.014	0.011	5.872	0.014	Reject
Age	0.020	0.010	0.007	3.376	- 0.002	Reject
Sex	0.006	0.008	0.007	0.576	- 0.006	Reject

Our analysis using five filter-based methods (Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), Chi-square (χ^2), and ReliefF) revealed that MPS emerged as the most relevant feature for predicting the target variable. Its consistently high scores across all measures (IG: 0.179, GR: 0.089, GINI: 0.061, χ^2: 52.775, ReliefF: 0.031) indicate its strong potential for discriminating between different classes [15]. TMQ followed MPS in importance, exhibiting moderate scores in most measures (IG: 0.093, GR: 0.047, GINI: 0.031, χ^2: 25.872, ReliefF: 0.009). MASA also showed notable relevance (IG: 0.061, GR: 0.031, GINI: 0.022, χ^2: 14.436, ReliefF: 0.013). These findings suggest that these three features likely hold valuable information for the prediction task. The remaining features displayed progressively lower scores, implying a weaker association with the target variable. *Number of Units and Year Level* exhibited some potential (IG: 0.039, GR: 0.020, GINI: 0.013, χ^2: 4.824, ReliefF: 0.002 for Number of Units; IG: 0.037, GR: 0.019, GINI: 0.013, χ^2: 7.041, ReliefF: 0.004 for Year Level), while *Grade and OATOL* showed minimal but measurable relevance (IG: 0.035, GR: 0.018, GINI: 0.012, χ^2: 5.582, ReliefF: 0.002 for Grade; IG: 0.029, GR: 0.014, GINI: 0.011, χ^2: 5.872, ReliefF: 0.014 for OATOL). *Age and Sex* displayed negligible or even negative scores (IG: 0.02, GR: 0.01, GINI: 0.007, χ^2: 3.376, ReliefF: -0.002 for Age; IG: 0.006, GR: 0.008, GINI: 0.007, χ^2: 0.576, ReliefF: -0.006 for Sex), suggesting they may not contribute significantly to the prediction task. It is crucial to acknowledge that each filter-based method evaluates feature importance from a distinct [21]. Therefore, considering all five measures provides a more comprehensive understanding of feature relevance. IG prioritizes uncertainty reduction, GR corrects for potential bias, GINI measures decision tree impurity reduction, χ^2 tests independence between variables, and ReliefF emphasizes distinguishing instances from nearest neighbors of different classes. This multifaceted analysis contributes to a robust assessment of feature importance.

Performance of the Gaussian-Bernoulli Mixed Naïve Bayes Model

Table 4 shows the different performance results of the different models generated in terms of Accuracy and Kappa Scores. The four models under comparison are as follows: Model A employs all input features; Model B uses only the five input features that were verified through filter-based feature selection techniques during the study's feature selection phase; Model C has been implemented with a 10-fold cross validation in addition to using only the five most pertinent input features, but with zero Laplace Smoothing; and Model D has also been implemented with a 10-fold cross validation in addition to using only the five most relevant input features, but with Laplace Smoothing set to one. These models are all still implemented in the RStudio environment.

The Training Accuracies, Testing Accuracies, and Kappa scores of the various models that were generated are showed in Table 4. The researcher provided Model A, the model that incorporates all of the recognized input features and includes redundant and unneeded input features, for comparison's sake. As stated in the claim, Model A has the lowest Kappa score and accuracies of all the models. In contrast, Model B, which employs only the five input features that were verified through filter-based feature selection techniques during the study's feature selection phase, shows a notable increase in accuracies. In a Naïve Bayes model, Laplace smoothing is sometimes applied to help address the issue of zero probability, which guarantees that the posterior probabilities are never zero. Our best model is Model C, which was developed using a 10-fold cross-validation process and only employs the five key features. We may conclude that zero probability did not exist in our data because Model D, which was implemented using a Laplace Smoothing, did not alter the model's conclusions. Model C performs exceptionally well in predicting students' future procrastinating tendencies in online mathematics learning, with an accuracy rate of 85% and an 82% Kappa Statistic.

Table 4. Performance Results of the Different Models

Model	Training Accuracy	Testing Accuracy	Kappa	Laplace Smoother
A (Using all input features)	0.55	0.44	0.23	0
B (Using only the relevant features)	0.74	0.69	0.53	0
C (Using the only the relevant features and using a 10-fold cross validation)	0.89	0.85	0.82	0
D (Using the only the relevant features and using a 10-fold cross validation)	0.89	0.85	0.82	1

Table 5. The Confusion Matrix Analysis of the Best Model Generated

Confusion Matrix Analysis		ACTUAL				Total
		High Procrastinator	Moderate Procrastinator	Low Procrastinator	Non Procrastinator	
PREDICTED	High Procrastinator	15	2	0	0	17
	Moderate Procrastinator	1	13	2	0	16
	Low Procrastinator	0	2	15	1	18
	Non Procrastinator	0	0	1	10	11
Total		16	17	18	11	

Table 5 displays the Confusion Matrix Analysis for the various classes of procrastination inclinations, with Model C being our top choice. Of the 53 predictions the model produced, 47 turned out to be accurate. The confusion matrix illustrates how few cases are misclassified by the model among the various classes or procrastinating inclinations. The model can produce very good results if it is especially used to identify the High Procrastinators and Moderate Procrastinators within a group of students, as less misclassifications are seen in those groups. Table 6 provides a summary of each class's performance based on the Confusion Matrix Analysis, which may be used to better understand the model's performance in each class or procrastinating tendencies.

Table 6. The Sensitivity and Specificity of Each Class in the Best Model Generated

Class	Sensitivity	Specificity
High Procrastinator	0.94	0.96
Moderate Procrastinator	0.76	0.93
Low Procrastinator	0.83	0.93
Non Procrastinator	0.91	0.98

The figures presented in Table 6 are obtained from the Confusion Matrix Analysis in Table 6. Simply put, Sensitivity demonstrates how well our approach can recognize pupils who exhibit that particular procrastinating inclination. Conversely, specificity demonstrates our model's ability to accurately identify kids who do not exhibit that particular procrastinating inclination. The bulk of the High Procrastinators and Non Procrastinators are accurately predicted by the model, according to the Confusion Matrix Analysis.

All things considered, our model performs better at identifying the students who are predicted to put off doing anything or not at all, or at differentiating between high and low procrastinators. This model can certainly fulfill its intended goal if it is used by a teacher to identify the pupils who procrastinate the most and those who may not have any problems at all.

3. Conclusions and Recommendations

The researchers draw the conclusion that the generated Gaussian – Bernoulli Mixed Naïve Bayes Model is a useful tool that mathematics teachers can use to anticipate students' procrastinating behaviors in their online math learning. Mathematics teachers can utilize the tool's predictions to identify students who are suitable for a particular task, take prompt action, or take early safeguards. Based on students' experiences in online classes, certain treatments and activities can be suggested. Since Naïve Bayes is only one of many Machine Learning models available, the researchers advise looking into more advanced and recent algorithms, such as Neural Networks and others, with more data or input features that may also have a great deal of predictive value for students' academic procrastination.

References

[1] Hussain, I., & Sultan, S. (2010b). Analysis of procrastination among university students. Procedia - Social and Behavioral Sciences, 5, 1897–1904.

[2] Flett, G.L.; Hewitt, P.L.; Martin, T.R. Dimensions of perfectionism and procrastination. In Procrastination and Task Avoidance: Theory, Research, and Treatment; Ferrari, J.R., Johnson, J.L., McCown, W.G., Eds.; Plenum Press: New York, NY, USA, 1995; pp. 113–136.

[3] Cerino, E.S. Relationships between academic motivation, self-efficacy, and academic procrastination. Psi Chi J. Psychol. Res. 2014, 19, 156–163.

[4] Barrot, J. S., Llenares, I. I., & Del Rosario, L. S. (2021). Students' online learning challenges during the pandemic and how they cope with them: The case of the Philippines. Education and Information Technologies, 26(6), 7321–7338.

[5] Elvers, G. C., Polzella, D. J., & Graetz, K. (2003). "Procrastination in online courses: Performance and attitudinal differences"." Teaching of Psychology". 30: 159-162.

[6] Klingsieck, K. B., Fries, S., Horz, C., & Hofer, M. (2012). "Procrastination in a distance university setting"." Distance Education". 33: 295-310.

[7] Rakes, G. C., & Dunn, K. E. (2010). "The Impact of Online Graduate Students' Motivation and Self-Regulation on Academic Procrastination"." Journal of interactive online learning". 9.

[8] Tuckman, B. W. (2011). The effect of motivational scaffolding on procrastinators' distance learning outcomes: A multiple case study. Journal of Marketing Education, 33(1), 5-17.

[9] Godinez, C. D. O., & Lomibao, L. S. (2022). A Gaussian-Bernoulli Mixed Naïve Bayes Approach to Predict Students' Academic Procrastination Tendencies in Online Mathematics Learning. American Journal of Educational Research, 10(4), 223-232.

[10] Kaur, A., Singh, J. P., & Sharma, V. (2020). K-means clustering for student grouping in online learning environments. International Journal of Engineering & Technology, 7(4), 31-36.

[11] Chavez-Luque, A., Maldonado-Basurto, L. Á., & Mora-Olvera, A. T. (2023). Educational data analysis using K-means clustering and decision trees to predict academic performance. Applied Sciences, 13(12), 6053.

[12] Cano-Plata, M. I., Duque-Méndez, M. Á., & Martínez-Gómez, P. (2022). Student engagement in virtual courses: A study using k-means clustering and decision trees. Sustainability, 14(12), 7866.

[13] Dash, M., & Dash, P. K. (2012). Feature selection for classification: A data mining perspective. New York: Springer.

[14] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: with applications in R. Springer.

[15] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3, 1157-1182.

[16] Robnik-Šikonja, M., & Kononenko, I. (2010). An empirical analysis of ReliefF and ReliefF for feature selection in regression. In Proceedings of the International Conference on Machine Learning (pp. 1227-1234).

[17] Yu, L., & Liu, H. (2004. Feature selection for data mining with evolutionary algorithms. Springer.

[18] S. Bagui, K. Devulapalli, and S. John, "MapReduce Implementation of a Multinomial and Mixed Naive Bayes Classifier," Int. J. Intell. Inf. Technol., vol. 16, no. 2, pp. 123, 2020.

[19] S. Visa, B. Ramsay, A. Ralescu, and E. Van Der Knaap, "Confusion matrix-based feature selection," CEUR Workshop Proc., vol. 710, pp. 120-127, 2011.

[20] Solomon, L. J., & Rothblum, E. D. (1984). Procrastination Assessment Scale--Students [Dataset]. In PsycTESTS Dataset.

[21] Brown, G., Humphreys, G., & ListNode, M. (2014). Feature selection for machine learning and data mining. John Wiley & Sons.