DOI: https://doi.org/10.30650/jse.v1i2.2205

Using MDA to Improve Naïve Bayes Classification for Students Performance Prediction

Rofilde Hasudungan[1*], Wawan Joko Pranoto[2], Rudiman[3]
[1,2,3] Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia
* Corresponding Email: rofilde@umkt.ac.id

Abstract – The ability to predict student performance and find the influence factors are an important task. It can help students who are predicted to have low performance so that they have a better result in the future. Naïve bayes classifier is a classification algorithm based on naïve theorem. This algorithm has high accuracy and fast. However, Naïve Bayes has no ability to select the best features since all attributes are considered equal. Nevertheless, it is common that there are attributes that higher dependency degree than others and there are attributes that not important or superfluous or redundant that affect classification performance, hence this paper aim to improve Naïve Bayes model by employing Maximum Dependency Attribute (MDA) to select best attributes in predicting student performance. MDA is a feature selection technique based rough set that able to select and remove redundant attributes based on attribute dependency. The experiment is conducted to 40 students with 28 features show that the proposed model has an accuracy of 79%. The result has improved compared to Naïve bayes without MDA with an accuracy of 68%.

Keywords: Educational Data Mining; Feature Selection; Rough Sets; Maximum Dependency Attribute; Naïve Bayes

Submitted: 2 April 2020 - Revised: 25 April 2020 - Accepted: 27 June 2020

1. Introduction

Adoption of information technology in education sector become massive nowadays. Almost all activities conducting in online system such as registration and enrollment, assignment, and become more massive when facing the situation such as pandemic as covid-19. The data produced by the university is huge. However, the data become useless until we mine it and convert it into knowledge.

Data mining used to mine knowledge from huge amount of data. When we deal with data that came from educational sector it called educational data mining (EDM). EDM not only because the data, but the current technique cannot apply directly to handle this kind of situation since there are different objectives.

Predicting student performance is very important task. The ability to predict student performance can help the students that predicted have low performance so he/she can have better result in the future.

Naïve Bayes classifier is one of classification algorithm that has good accuracy, easy to implements, fast and able to handle large dataset as well as can handle numerical and categorical data[1]. This algorithm widely used such as text classification in [2], [3], [4], sentiment analysis in [5], [6], software defect prediction in [7], health in [8], and more. Naïve Bayes assume that all features are independent each other's. However, some features related to each other's and selecting the best attribute will affect the accuracy of classification.

MDA is rough set-based feature selection that able find the dependency attribute of attributes and eliminate redundant features. This algorithm proposed by Herawan[9] to select the best attribute for clustering. Furthermore, [10] implements that algorithm in classification to select the best feature in Malay musical instrument, and yield promising result. Hence, this research aims to combine MDA and Naïve bayes for finding important factors and predicting student performance.

2. Related Works

Data mining also called as Knowledge Discovery in Database (KDD) is a set tool used to reveal hidden knowledge in big data. It is used in many domains such as astronomy, medical, and education. In education its mainly call data mining in education (EDM) that used to reveal knowledge related to education[11]. Massive implementation of information technology in education institution yield massive data that contain information that can be mined[12].

In recent years many researchers applied EDM to predict student performance by using various data mining techniques and many parameters are used and proposed such Gowri et al. [13] has used several data mining technique (1) apriori algorithm used to extract pattern that are similar along with their associations in relation to various set of records (2) K-means cluster analysis used to generate group of students with similar characteristics. In this study, Gowri et al. [13] used for main indicator (1) Academic parameters, (2) family history, (3) learning methodology and (4) personal characteristic [13]. Rosadi et al.[14] using clustering technique (fuzzy C-mean) to group students based on GPA and graduation time. This clustering aim to divide students into four main group: (1) bad, (2) not good, (3) very good and (4) good. Khasanah and Harwati [15] used two data mining technique: Naïve Bayes and Decision Tree to predict and reveal the most influence indicators toward student performance. Khasanah and Harwati [15] reveal that attendance level and CGPA are an important indicators that most influence the student's performance. This research also found that Naïve bayes has better accuracy level that decision tree.

Al-barrak and Al-Razgan [12] in another hand predict the CGPA based on the student's performance in particular courses. They believe some courses have more influences than other courses for determined CGPA. By using decision tree (J48) they analyze the courses for student in department information technology, King Saud University and found that some courses have more influence than other.

Ahmed and Elaraby[16] used classification (Decision tree) technique to predict student' final score. This research used previous score such as assignment, homework, mid test, seminar, participant, and attendance to predict final score. ID3 algorithm show that from all indicators, the mid semester is the main indicator that influence the final score.

3. Rough Set

Maximum dependency attribute (MDA) is a feature reduction technique based on rough set. It calculates the dependency of an attribute to other attributes and choose the subset of attribute based on the maximum degree of the attribute. In this section, we will discuss basic concept of rough set as main concept of MDA and MDA itself.

3.1. Basic theory of rough set

Rough set is a mathematical tool that proposed by Pawlak[17] to works with vague and uncertainty. There are several concepts in rough set theory such as information system and decision system, indiscernible relation, set approximation, and dependency attribute.

Information system and decision information system are tables that represent data in rough set theory. Information system is four tuples, $IS = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a non-empty set of attributes, $V = \bigcup_{a \in A} V_a$, V_a is the domain of attribute a, and $f: U \times A \rightarrow V$ is a function that map object to the with domain. Meanwhile, decision information system is defined as $DIS = (U, A \cup \{d\}, V, f)$ where d is decision attribute and $A \cap \{d\} = \emptyset$.

Indiscernible relation is relation between two objects. Two objects x and y is equivalence if $\forall b \in B \rightarrow b(x) = b(y)$, where $B \subseteq A$ and A is set attributes and $(x, y) \in U \times U$. This indiscernible relation induces partition of U. The partition of U induced by B is denoted as $IND(B)$. Meanwhile, $[x]_B$ denoted equivalence class inside $IND(B)$ that contain x.

Set approximation is approximation of a set by other sets. Let subset $X \subseteq U$ and $R \in IND(B)$, we associate two subsets $\underline{R}X = \{x \in U | [x]_R \subseteq X\}$ and $\overline{R}X = \{x \in U | [x]_R \cap X \neq \emptyset\}$ called the lower and upper approximation, respectively. From set approximation one can calculate dependency attributes by using following formula:

$$\gamma_C = \frac{|POS_C|}{|U|} \qquad (1)$$

Where γ_C represent the degree attribute D depend on attribute C (denoted as $C \implies D$), POS_C represent lower approximation, U represent all objects and $|.|$ represent cardinality.

3.2. Maximum Dependency Attribute (MDA)

Feature selection used to select a subset of features from all features that relevance and high dependency to the data. Maximum Dependency Attribute (MDA) is feature reduction based on Rough Set that initially proposed by [9] to select clustering attribute. This technique has advantages in term of finding attribute that has maximum dependency and eliminate redundancy. Furthermore, Senan et al [10] implements this technique to select attribute for classification in traditional Malay music instruments. The relation between properties of roughness of a subset $X \subseteq U$ with the dependency between two attributes presented in Proposition 1.

Proposition 1. Let $S = (U, A, V, f)$ be and information system and let D and C be any subsets of A. If D depends on totally on C, then $\alpha_B(X) \leq \alpha_C(X)$, for every $X \subseteq U$.

Proof. Let D and C by any subsets of A in information system $S = (U, A, V, f)$. From the hypothesis, we have $IND(C) \subseteq IND(D)$. Furthermore, the clustering U/C is finer that U/D, thus it is clear that any equivalence class induced by $IND(D)$ is a union of some equivalence class induced by $IND(C)$. Therefore, for every $x \in X \subseteq U$, we have $[x]_C \subseteq [X]_D$.

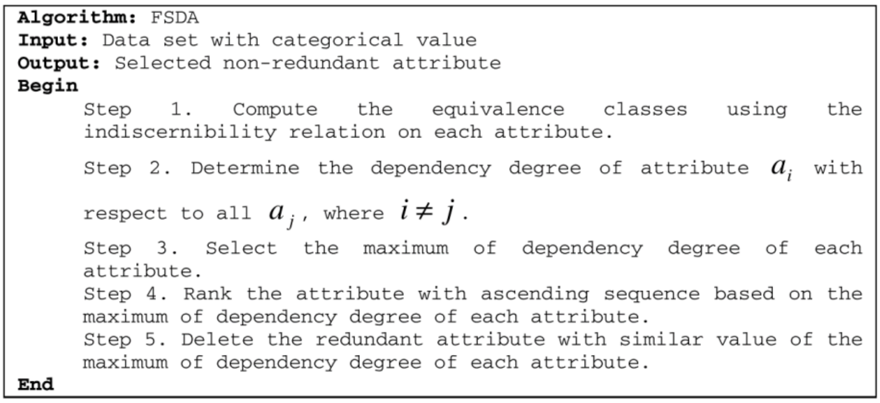
Hence, for every $X \subseteq U$, we have the following relation:

$$\underline{D}(X) \subseteq \underline{C}(X) \subseteq X \subset \bar{C}(X) \subseteq \bar{D}X$$

Consequently,

$$\alpha_D(X) = \frac{|\underline{D}(X)|}{|\bar{D}(X)|} \leq \frac{|\underline{C}(X)|}{|\bar{C}(X)|} = \alpha_C(X)$$

The generalization of Proposition 1 expressed in proposition 2.



```
Algorithm: FSDA
Input: Data set with categorical value
Output: Selected non-redundant attribute
Begin
     Step 1. Compute the equivalence classes using the
     indiscernibility relation on each attribute.
     Step 2. Determine the dependency degree of attribute $a_i$ with
     respect to all $a_j$, where $i \neq j$.
     Step 3. Select the maximum of dependency degree of each
     attribute.
     Step 4. Rank the attribute with ascending sequence based on the
     maximum of dependency degree of each attribute.
     Step 5. Delete the redundant attribute with similar value of the
     maximum of dependency degree of each attribute.
End
```

Figure 1 MDA algorithm for attribute selection

Proposition 2. Let $S = (U, A, V, f)$ be information system and let $C_1, C_2, \dots, C_n$ and D be any subsets of A. If $C_1 \Rightarrow_{k_1} D,\ C_2 \Rightarrow_{k_2} D,\ \dots C_n \Rightarrow_{k_n} D,$ where $k_n \leq k_{n-1} \leq \cdots \leq k_2 \leq k_1$, then

$$\alpha_D(X) \leq \alpha_{C_n}(X) \leq \alpha_{C_{n-1}}(X) \leq \cdots \leq \alpha_{C_2}(X) \leq \alpha_{C_1}(X)$$

For every $X \subseteq U$.

Proof. Let $C_1, C_2, \dots, C_n$ and D be any subsets of A in information system S. From the hypothesis and Proposition 1, the accuracies of roughness are given as

$$\alpha_D(X) \leq \alpha_{C_1}(X)$$
$$\alpha_D(X) \leq \alpha_{C_2}(X)$$
$$\vdots$$
$$\alpha_D(X) \leq \alpha_{C_n}(X)$$

Since $k_n \leq k_{n-1} \leq \cdots \leq k_2 \leq k_1$, then

$$[x]_{C_n} \subseteq [x]_{C_{n-1}}$$
$$[x]_{C_{n-1}} \subseteq [x]_{C_{n-2}}$$
$$\vdots$$
$$[x]_{C_2} \subseteq [x]_{C_1}.$$

Obviously,

$$\alpha_D(X) \leq \alpha_{C_n}(X) \leq \alpha_{C_{n-1}}(X) \leq \cdots \leq \alpha_{C_2}(X) \leq \alpha_{C_1}(X)$$

Figure 1 shows the pseudo-code of selecting feature based on this technique. The algorithm computes the dependency attribute and finds the dependency maximum for each attribute and eliminate attributes that have similar values.

4. Naïve Bayes Classifier

Naïve bayes classifier is a classification method that can be used to predict the probability the membership of the class. This method based on Bayes theorem that provided a way to calculate the probability of a prior event by using another subsequent event has occurred. The main formula of the Bayes theorem is given as bellow:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad (1)$$

Where X is data with unknown class, H is the hypothesis of X data is a specific class, $P(H|X)$ the probability of hypothesis H is based on X condition, $P(H)$ is H hypothesis probability, $P(X|H)$ is probability X under these conditions, $P(X)$ is the probability of X.

Naïve Bayes classifier is one of the most simple but sophisticated technique based on Bayes theorem. This technique assumes that all features all independence to each other that why it called Naïve Bayes.

Naïve Bayes classifier has several stages as follows ([18]):

1) Let D be training set of tuples and their associated class labels.
2) Suppose that there are *m* classes, $C_1, C_2, \dots, C_m$. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, condition on X. Naïve bayes classifier predict that object X belongs to class C_i if only if $P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m, n \neq i$. $P(C_i|X)$ is calculated by using following equation:
$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (2)$$
3) As $P(X)$ is constant for all classes, only $P(X|C_i)P(C_i)$ need to be maximized.
4) Calculate probability of $P(X|C_i)$ by using following equation:
$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \qquad (3)$$
5) To predict the class label of X, $P(X|C_i)P(C_i)$ is evaluated for each class C_i. The classifier predicts that the class label of tuple X is the class C_i if and only if $P(X|C_i)P(C_i) > P\left(X\middle|C_j\right)P(C_j)$ for $i \leq j \leq m, j \neq i$.

5. Research Method

To solve the main issue, there are several stages used in this research as described in Figure 2. The research started with collecting data based on current standard lecturer evaluation form which are 28 parameters as describe in Table 1. Each parameter has several options that student should choose one of them. The possible values are (1) Very poor, (2) Poor, (3) Fair, (4) Good, (5) Excellent. Through questioner, students evaluate their lecturer performance in teaching. Instead of collecting data in the end of semester like current configuration, this research collecting data in the beginning of semester, where the data are collected before quiz 1 occurred. By this configuration, we can investigate the problem in teaching and learning earlier and still have time to reconfigure or improve the performance before end of semester. There are 47 students involve in this stage. The students are from fundamental programming class, semester 1, academic years 2018/2019, Universitas Muhammadiyah Kalimantan Timur.

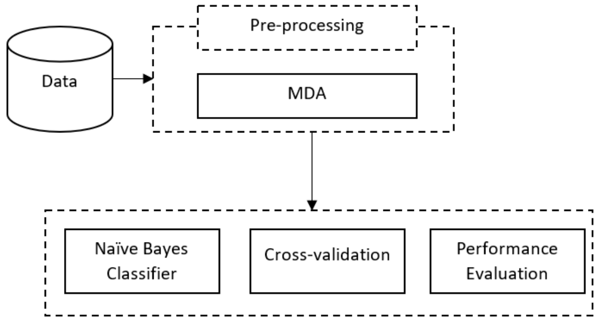


Figure 2 Propose Model

Table 1
The parameter

Variable	Description
P1	Readiness to teach
P2	Regularity and order organization of lectures
P3	The ability to revive the class atmosphere
P4	Clarity conveys material and answers the questions.
P5	Usage of media and learning technology.
P6	Diversity of ways to measure learning outcomes
P7	Providing feedback on assignments
P8	Suitability between exam and/or assignments to learning objective
P9	Suitability the grade provided with learning outcomes
P10	The ability to explain the subject/topic correctly
P11	The ability to give relevant examples from given concepts
P12	The ability to explain the correlation between a subject/topic taught with other subjects/topics
P13	The ability to explain the correlation of the subject/topic taught in the context of live
P14	Mastery of current issues in the field being taught
P15	Use of research results to improve the quality of lectures
P16	The involvement of students in research / study and / or development / engineering / design is done by lecturers
P17	Ability to use various communication technologies
P18	Authority as a lecturer
P19	Wisdom in making decisions
P20	Became example in attitude and behavior
P21	One words and actions
P22	The ability to control herself/himself in various situations and conditions
P23	Fair in treating students
P24	The ability to express opinions
P25	The ability to accept criticism, suggestions, and opinions of others
P26	Get to know students who attend their lectures
P27	Get along easily with colleagues, employees, and students
P28	Tolerance to the diversity of students

Furthermore, the pre-processing stage is carried out by eliminating attribute that redundant. In this stage, Maximum Dependency Attribute algorithm is employed to select the best attributes by calculating dependency value and eliminate the attributes that have equal value where the processes are described in Figure 1.

From previous result, the best attributes are obtained. By using that attributes, Naïve Bayes classifier is run to predict to predict students' performances. In Naïve Bayes process, the probability of students fails, or pass will be calculated as well as calculate the probability for each attribute, so based on that probability, the algorithm can predict student performance. To evaluate the model, cross-validation with $k = 10$ is employed. This evaluation enables us to calculate the accuracy, precession and recall of MDA + Naïve Bayes and compared it to Naïve Bayes. Accuracy, precession and recalled are calculated by using equation (4), (5), and (6), respectively.

$$accuracy = \frac{TP + FN}{TP + TN + FN + FP} \tag{4}$$

$$precession = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

6. Result and Discussion

6.1. Feature Selecting using MDA

In this stage, we employed MDA to select the best features. By using MDA, we reduce features from 28 to 7 as depicted in Table 2. In this process, we calculate the dependency for each feature, and eliminate redundant features. The redundant features are features that have same maximum dependency. The removing process consider the next maximum dependency. For example, there two attributes A, and B with dependency attribute A: {0.6, 0.5} and B: {0.6, 0.4}. Since those attributes have same maximum dependency, so we have eliminated one of them. Since the next dependency attribute of A(0.5) greater that B(0.4), so we eliminate B.

Table 2
Importance Features based on MDA

Variable	Maximum Dependency
P_{18}	0.34042553191489
P_7	0.17021276595745
P_9	0.12765957446809
P_{16}	0.1063829787234
P_{15}	0.085106382978723
P_{10}	0.063829787234043
P_2	0.042553191489362

Based on this process we found that P_{18} has highest dependency value among other attributes.

6.2. Classification using Naïve Bayes Classifier

In this process, we are using Rapid Miner to run Naïve Bayes. There are two experiments: (1) running naïve

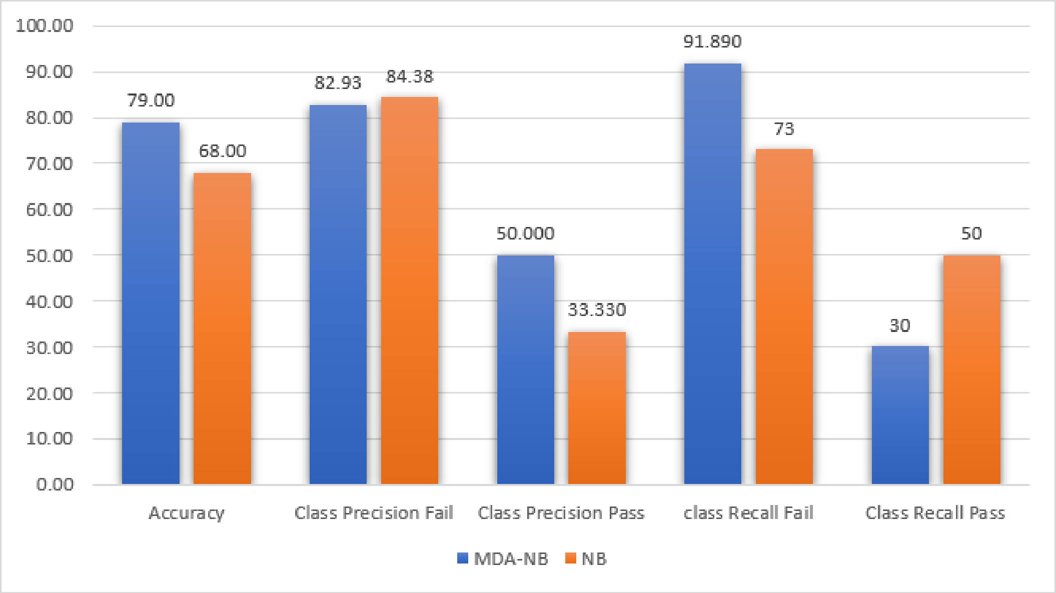


Figure 3 Comparison between MDA-Naïve Bayes and Naïve Bayes

bayes with all features (without MDA), and (2) running naïve bayes with features taken from MDA feature selection. To validate our approaches, cross-validation is employed, with k=10. Based on these experiments, two confusion matrices are built as shown in Table 3 and Table 4 for Naïve bayes with MDA and Naïve bayes without MDA, respectively.

Table 3
Confusion Matrix for Naïve Bayes Classifier with MDA

Variable	True Fail	True Pass
Pred. Fail	34	7
Pred. Pass	3	3

Table 4
Confusion Matrix for Naïve Bayes Classifier

Variable	True Fail	True Pass
Pred. Fail	27	5
Pred. Pass	10	5

As shown in Table 3 and Table 4, there is improvement for students who are predicted fail in quiz from 27 to 34 students. However, for students who are predicted pass decreased from 7 to 5 students. Detail comparison for the value of confusion matrix shown in Figure 2.

Based on value in confusion matrix for each model, we calculated accuracy, class precision, and recall. The comparison between Naïve Bayes-MDA and Naïve Bayes shown in Figure 4. Figure 4 shown that by adding MDA as feature selection increase accuracy significantly from 68% to 79%, and class precision fail from 73% to 91.89%. The improvement is due to the number of students that predicted fail increase from 27 to 34 closer to real data. However, the students who predicted pass decrease from 5 to 3 far away from real data, causing the precision for class fail, class recall pass decrease from 84.38% to 82.93% and 50% to 30%. respectively.

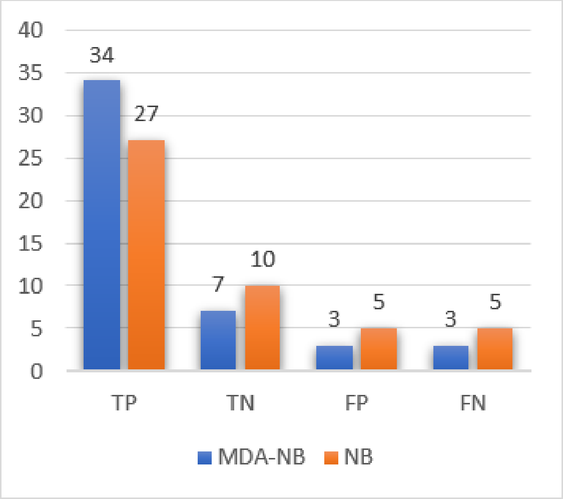


Figure 4 Confusion values comparison

7. Conclusion

This paper presents predicting student performance based on Naïve Bayes and MDA as feature selecting to reduce the attributes. In this paper we use data from questioner taken from students in subject fundamental programming. The data has 47 rows with 28 attributes. By using MDA, we succeeded to reduce attributes become 7 by eliminating redundant attributes. Based on selected attributes, classification processing is conducted we found that it has significant improvement in accuracy from 68% to 79%, so this combination is very promising to improve naive bayes classification.

References

[1] K. Yadav and R. Thareja, "Comparing the Performance of Naive Bayes And Decision Tree Classification Using R," *Int. J. Intell. Syst. Appl.*, vol. 11, no. 12, pp. 11–19, 2019, doi: 10.5815/ijisa.2019.12.02.

[2] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," *J. Inf. Sci.*, vol. 44, no. 1, pp. 48–59, 2018, doi: 10.1177/0165551516677946.

[3] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," *2017 IEEE 1st Ukr. Conf. Electr. Comput. Eng. UKRCON 2017 - Proc.*, pp. 900–903, 2017, doi: 10.1109/UKRCON.2017.8100379.

[4] T. Pranckevičius and V. Marcinkevičius, "Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification," *Balt. J. Mod. Comput.*, vol. 5, no. 2, pp. 221–232, 2017, doi: 10.22364/bjmc.2017.5.2.05.

[5] M. S. Mubarok, A. Adiwijaya, and M. D. Aldhi, "Aspect-based sentiment analysis to review products using Naïve Bayes," *AIP Conf. Proc.*, vol. 1867, no. August, 2017, doi: 10.1063/1.4994463.

[6] F. Xu, Z. Pan, and R. Xia, "E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework," *Inf. Process. Manag.*, vol. 57, no. 5, p. 102221, 2020, doi: 10.1016/j.ipm.2020.102221.

[7] Ö. F. Arar and K. Ayan, "A feature dependent Naive Bayes approach and its application to the software defect prediction problem," *Appl. Soft Comput. J.*, vol. 59, pp. 197–209, 2017, doi: 10.1016/j.asoc.2017.05.043.

[8] L. Ali *et al.*, "A Feature-Driven Decision Support System for Heart Failure Prediction Based on $\chi 2$ Statistical Model and Gaussian Naive Bayes," *Comput. Math. Methods Med.*, vol. 2019, 2019, doi: 10.1155/2019/6314328.

[9] T. Herawan, M. M. Deris, and J. H. Abawajy, "A rough set approach for selecting clustering attribute," *Knowledge-Based Syst.*, vol. 23, no. 3, pp. 220–231, 2010, doi: 10.1016/j.knosys.2009.12.003.

[10] N. Senan, R. Ibrahim, N. Mohd Nawi, I. T. R. Yanto, and T. Herawan, "Rough set approach for attributes selection of traditional Malay musical instruments sounds classification," *Commun. Comput. Inf. Sci.*, vol. 151 CCIS, no. PART 2, pp. 509–525, 2011, doi: 10.1007/978-3-642-20998-7_59.

[11] M. Anand, "Advances in EDM: A State of the Art," in *Software Engineering*, 2019, pp. 193–201.

[12] M. A. Al-Barrak and M. Al-Razgan, "Predicting Students Final GPA Using Decision Trees: A Case Study," *Int. J. Inf. Educ. Technol.*, 2016, doi: 10.7763/IJIET.2016.V6.745.

[13] G. S. Gowri, R. Thulasiram, and M. A. Baburao, "Educational Data Mining Application for Estimating Students Performance in Weka Environment," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 263, no. 3, p. 032002, Nov. 2017, doi: 10.1088/1757-899X/263/3/032002.

[14] R. Rosadi, R. Sudrajat, B. Kharismawan, and Y. A. Hambali, "Student academic performance analysis using fuzzy C-means clustering," in *IOP Conference Series: Materials Science and Engineering*, 2017, vol. 166, no. 1, p. 12036.

[15] A. U. Khasanah and Harwati, "A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques," *{IOP} Conf. Ser. Mater. Sci. Eng.*, vol. 215, p. 12036, Jun. 2017, doi: 10.1088/1757-899x/215/1/012036.

[16] A. B. E. D. Ahmed and I. S. Elaraby, "Data Mining: A prediction for Student's Performance Using Classification Method," *World J. Comput. Appl. Technol.*, 2014, doi: 10.13189/wjcat.2014.020203.

[17] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, 1982, doi: 10.1007/BF01001956.

[18] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.