AN EDUCATIONAL DATA MINING FOR STUDENT ACADEMIC PREDICTION USING K-MEANS CLUSTERING AND NAÏVE BAYES CLASSIFIER

Dewi Ayu Nur Wulandari[1]; Riski Annisa[2]; Lestari Yusuf[3],Titin Prihatin[4]

Sistem Informasi Kampus Kota Bogor[1*]; Sistem Informasi Kampus Kota Pontianak[2]
Universitas Bina Sarana Informatika[1,2]
www.bsi.ac.id[1,2]
dewi.dan@bsi.ac.id[1], riski.rnc@bsi.ac.id[2]

Sistem Informasi[3]; Teknik Informatika[4]
Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri [3,4]
www.nusamandiri.ac.id [3,4]
lestari.lyf@nusamandiri.ac.id[3], titin.tpn@nusamandiri.ac.id[4]
(*) Corresponding Author

Abstract— Data mining in education is something that can be used to analyze data obtained from processes in the world of education to obtain new information from existing data so that it is useful for improving the quality of learning. One way that can be done to improve the quality of learning is to predict student academic achievement. Student academic classification based on academic potential can be a strategy to increase student graduation, improve learning achievement, and also for better management of student academic data. This study proposes combining the K-Means data mining clustering method and the Naïve Bayes classifier (K-Means Bayes) for better results in processing student academic performance data. The data were taken from the Student Academic Performance dataset which was used as a test case. The amount of data used in this study were 131 data and 21 attributes. The accuracy of the results obtained from the combination of the proposed methods is 97.44%. Although the initial centroid determination in the K-Means method is done randomly, the impact can be reduced by adding the Naïve Bayes Classifier method which results in a better accuracy value, thereby increasing the accuracy of the method used. Compared with the K-Means and Naïve Bayes methods, the proposed method increases the accuracy of about 27% of the Naïve Bayes algorithm and about 23% of the K-Means algorithm. With the results obtained, it can be concluded that the proposed method can improve the prediction of student achievement data.

Keywords: *K-Means, Naive Bayes, Data Mining*

Abstrak— Data mining dalam dunia pendidikan merupakan teknik yang dapat digunakan untuk menganalisis data yang diperoleh dari proses dalam dunia pendidikan untuk memperoleh informasi baru dari data yang dimiliki sehingga berguna untuk meningkatkan kualitas pembelajaran. Salah satu cara yang dapat dilakukan untuk meningkatkan kualitas pembelajaran adalah dengan memprediksi prestasi akademik siswa. Klasifikasi akademik siswa berdasarkan potensi akademik dapat menjadi strategi untuk meningkatkan kelulusan mahasiswa, meningkatkan prestasi belajar, dan juga untuk pengelolaan data akademik siswa yang lebih baik. Penelitian ini mengusulkan penggabungan metode K-Means clustering data mining dan Naïve Bayes classifier (K-Means Bayes) untuk hasil yang lebih baik dalam pengolahan data kinerja akademik siswa. Data diambil dari dataset Kinerja Akademik Mahasiswa yang digunakan sebagai test case. Jumlah data yang digunakan dalam penelitian ini adalah 131 data dan 21 atribut. Akurasi hasil yang diperoleh dari kombinasi metode yang diusulkan adalah 97,44%. Meskipun penentuan centroid awal pada metode K-Means dilakukan secara acak, namun dampaknya dapat dikurangi dengan menambahkan metode Naïve Bayes Classifier yang menghasilkan nilai akurasi yang lebih baik, sehingga meningkatkan akurasi metode yang diusulkan. Dibandingkan dengan metode K-Means dan Naïve Bayes, metode yang diusulkan meningkatkan akurasi sekitar 27% dari algoritma Naïve Bayes dan sekitar 23% dari algoritma K-Means. Dengan hasil yang diperoleh, dapat disimpulkan bahwa metode yang diusulkan dapat meningkatkan prediksi data prestasi akademik siswa.

Kata Kunci*: K-Means, Naive Bayes, Data Mining*

INTRODUCTION

Data mining in education is something that can be used to analyze data obtained from processes in the world of education to obtain new information from existing data so that it is useful for improving the quality of learning. One way that can be done to improve the quality of learning is to predict student academic achievement. Student academic classification based on academic potential can be a strategy to increase student graduation, improve learning achievement, and also for better management of student academic data.

The large number of students who do not graduate on time creates its problems. New students should be given input about what factors support students to graduate on time. So far, tertiary institutions have not explored the data they have to obtain new information related to data generation and graduate data that can be used to analyze the factors that affect student graduation rates (Risnawati, 2018).

The education sector is a field that gets a good impact from an advance in technology and information. Data mining in education is something that can be used to analyze data to obtain new information from the data that is owned so that it is useful for the university to improve the quality of learning. Technological development makes educational institutions carry out the process of digitizing academic data, which is important for educational institutions to process and analyze academic data to obtain new information that helps in the decision-making process (Asif et al., 2017).

Currently, the information on the number of student passes can be known after the announcement of the exam results is announced. Schools cannot take early anticipatory steps to increase the number of passes if it does not predict the number of student graduations (Satrianansyah & Wulandari, 2019).

To predict the student graduation rate according to predetermined standards, a data mining method with a classification function is needed which is useful for taking early anticipatory steps to overcome the occurrence of problems in the academic fields (Widaningsih, 2019)

To improve the quality of decision making to improve the quality of learning, many educational institutions increasingly recognize the importance of analyzing academics data obtained (Miguéis et al., 2018). To improve the education of quality, this is necessary to predict the academic students, so that improvements and actions can be done early in improving the quality and quality of learning (Hamsa et al., 2016). New information that can be obtained from the many academic data obtained provides an opportunity to optimize user technology to enhance the learning experiences (Waheed et al., 2020).

Many techniques can be used in evaluating student performance. The information provides clues about previously unknown trends related to student performance and learning behavior, teaching efficiency, and quality, prediction of student potential, student tendencies that lead to dropouts, etc. This provides an opportunity to identify quality gaps and to propose improvements and policies needed to improve the quality also the delivery of the education systems (Adekitan & Salau, 2019).

Data mining which also known as Knowledge Discovery in Database (KDD), refers to extracting information from a large amount of data to obtain new information. Data mining techniques are used on large amounts of data to discover new patterns that can be used to assist in decision making. Data mining is one of the most popular techniques for analyzing student performance.

Useful information and patterns can be used to predict student performances (Shahiri et al., 2015). Educational Data Mining (EDM) is a data mining that aims to develop methods in exploring data that used these methods in understanding students' and students' learning experiences (Bansode, 2016). Educational Data Mining (EDM) can be defined as data mining techniques to help in analyzing educational data. EDM uses a database of the education system to understand students and their learning styles more comprehensively to design educational policies that will improve their academic achievement, and reduce the failure rate at the end of each school year.

EDM can find new patterns and new knowledge about the students learning process (Fernandes et al., 2019). Data mining techniques that can be used to apply Educational Data Mining (EDM) include the K-Means clustering algorithm and the Naïve Bayes algorithm. We used hybrid learning through a combination of two methods of classification and grouping, namely Naive Bayes and the K-Means to solve this problem. Groupon of all data will be showed proposed in the relevant groups before applying them to the classification. And the result of the combination shows greater behavior with the data(Muda et al., 2011).

This study will discuss the modification of the K-Means clustering algorithm and the Naive Bayes classification to improve accuracy.

MATERIALS AND METHODS

An educational data mining is a new thing related to the discipline that has recently emerged relating to the development of methods in exploring educational data taken from an interactive learning environment sourced from

schools or universities which is because a large amount of data often has different meanings.

In this research the method used is the K-Means clustering algorithm to group data, then used the Naïve Bayes to classification the data.

A. Datasets

In this study, we used the Academic Performance dataset taken at the UCI Repository, because it aims to find end-of-semester predictions based on social, economic, and academic attributes (Hussain et al., 2018). In this study, data were obtained and collected from 3 different universities, namely Duliajan College, Doomdooma College, and Digboi College of Assam, India. This data is easily retrieved because it is freely available in the UCI Machine Learning Repository.

There is a lot of research discussing students' academic performance and predictions of students dropping out of school and their job prospects to measure the quality of education held by universities. Most studies consider grade point averages (GPA) because the response variables and explanatory variables vary. But this dataset uses final semester percentages as variables. There are 131 data which is all data are nominal data, 22 attributes, and 3 clusters.

```
@RELATION Sapfile1

@ATTRIBUTE ge  {M,F}
@ATTRIBUTE cst {G,ST,SC,OBC,MOBC}
@ATTRIBUTE tnp {Best,Vg,Good,Pass,Fail}
@ATTRIBUTE twp {Best,Vg,Good,Pass,Fail}
@ATTRIBUTE iap {Best,Vg,Good,Pass,Fail}
@ATTRIBUTE esp {Best,Vg,Good,Pass,Fail}
@ATTRIBUTE arr {Y,N}
@ATTRIBUTE ms  {Married,Unmarried}
@ATTRIBUTE ls  {T,V}
@ATTRIBUTE as  {Free,Paid}
@ATTRIBUTE fmi {Vh,High,Am,Medium,Low}
@ATTRIBUTE fs  {Large,Average,Small}
@ATTRIBUTE fq  {Il,Um,10,12,Degree,Pg}
@ATTRIBUTE mq  {Il,Um,10,12,Degree,Pg}
@ATTRIBUTE fo  {Service,Business,Retired,Farmer,Others}
@ATTRIBUTE mo  {Service,Business,Retired,Housewife,Others}
@ATTRIBUTE nf  {Large,Average,Small}
@ATTRIBUTE sh  {Good,Average,Poor}
@ATTRIBUTE ss  {Govt,Private}
@ATTRIBUTE me  {Eng,Asm,Hin,Ben}
@ATTRIBUTE tt  {Large,Average,Small}
@ATTRIBUTE atd {Good,Average,Poor}

@DATA
F,G,Good,Good,Vg,Good,Y,Unmarried,V,Paid,Medium,Average,Um,10,Farmer,Housewife,Large,Poor,Govt,.
```

Source : (Wulandari et al., 2020)

Figure 1. Attributes Used In Research

Figure 3 shows the attributes used in this study, totaling 22 attributes.

B. Preparation

Data will be processed through several stages of initial data processing (data preparation). The data processing stage is the grouping of data by dividing data into three groups using the K-Means algorithm. K-Means algorithm has a weakness because it cannot process nominal data. The data will be compared with numeric values so that finally there are three clusters obtained. Furthermore, the data will be grouped and divided using ten cross-validation techniques to be divided into two parts data, namely training data and testing data. The test data will be implemented using the Naïve Bayes classification algorithm. Furthermore, the evaluation results will be written in a confusion matrix.

Table 1 Confusion Matrix

	Act. True	Act. False
Pred. True	TP	FP
Pred. False	FN	TN

Source : (Wulandari et al., 2020)

Evaluation of measurement data classification mining is done by measuring and calculating accuracy based on the confusion matrix formed. The data in Table 1 is an example of the results of the confusion matrix calculation. TP (true positive) is the amount of data predicted true and reality is true. TN (true negative) is the amount of data that is predicted false and false facts. FP (false positive) is the amount of data that is predicted true, but the reality is false. And FN (false negative) is the amount of data predicted to be wrong, but the reality is true. The formula for calculating accuracy uses equation 1

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{..... (1)}$$

C. Modeling

Data mining modeling in this study was carried out using K-Means Clustering and Naive Bayes Classification

1) The K-means clustering

The K-means clustering algorithm is a cluster that is affected by the initial centroid cluster selection. The data element K is chosen as the initial center and then the distance of all data elements is calculated by the Euclidean distance formula. Data elements that have less distance to the center of mass are moved to the appropriate cluster. The process continues until no more changes occur in the cluster. The following are the basic steps of the clustering algorithm in K-mean clustering.

2) Naive Bayes Classifier

Naive Bayes classifier is one of the statistical classifiers, which can predict the probability of class membership of tuple data that will enter a certain class, according to probability calculations. This method is often used to solve problems in the field of machine learning because it is known to have a high degree of accuracy with simple calculations (Handayani & Pribadi, 2015)

Bayes's theorem comes from

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{.....(2)}$$

The Naïve Bayes classification estimates the following probability equation:

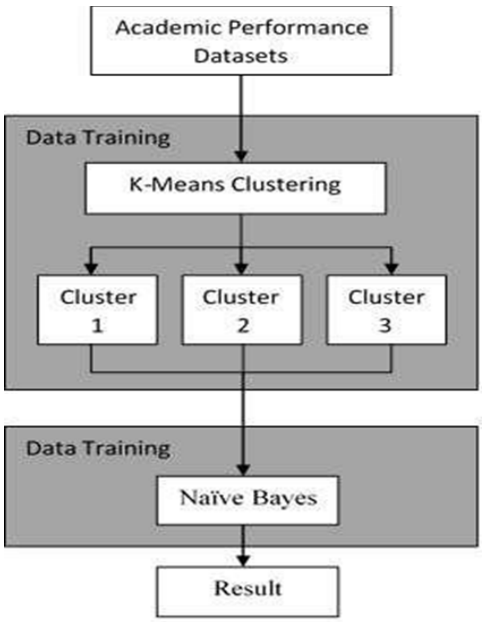
$$P(y) = \frac{n_y}{n} \quad \text{.....(3)}$$

$$P(x_i|y) = \frac{n_{y\&xi}}{n_y} \quad \text{.....(4)}$$

Where n, the total number of data points in the training data set, n_y number of class y target data points, $n_{y\&xi}$, several data points with target class y, and i the attribute variable that takes the value from x.

D. **Proposed Method**

This study will discuss the modification of the K-Means clustering algorithm and the Naive Bayes classification to improve accuracy. The flow stages of the algorithm to be modified are shown in Figure 2.



Source : (Wulandari et al., 2020)

Figure 2. Design Study

In the first phase, the clustering process using the K-Means algorithm is performed first to determine training data. Academic Performance data collection has 3 classes, which are good, average, and bad. In the second phase, data testing will be carried out using the Naive Bayes algorithm.

RESULT AND DISCUSSION

In conducting the calculation process, this study used a computer with Intel (R) Core i5 CPU 2.70GHz CPU, 8GB RAM, Windows 10 Operating System, 64 bit, and application programs using RapidMiner software. In this study applying the k-means clustering algorithm to group data then the data is classified using the Naive Bayes algorithm.

Data were taken from the Student Academic Performance Dataset (Hussain et al., 2018), used in this study as a test case. The amount of data used in this study were 131 data and 21 attributes.

The dataset is tested by ignoring the original label then by applying the k-means algorithm to label the new data by grouping the data using the k-means algorithm. By applying clustering techniques to the original dataset, divide the data into 4, 3, or 2 clusters as a test tool. Next to predict the results of educational data using the Naïve Bayes classification algorithm. Then compare the results of classification, grouping, and classification integration. This research identifies classification rules for classification rules through experimental studies to classify academic performance results.

Data is presented in a database in the ARFF format then tested using the RapidMiner data tools. The result of dataset student academic performance which calculated used data tool rapid miner can be seen in table 2:

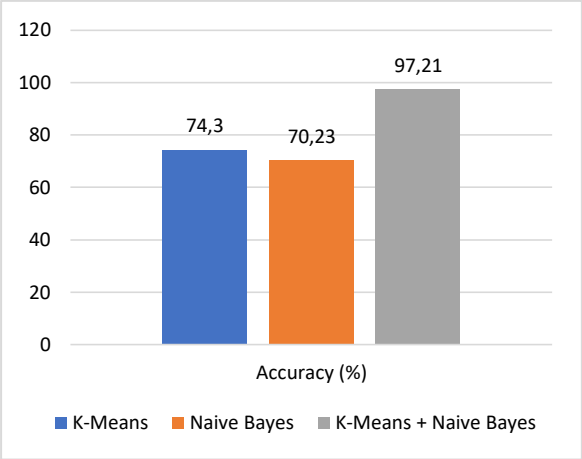
Table 2. Experimental Result

	Accuracy	Precision	Recall	F-Measure
K-Means	74.3%	84%	60.66%	70.44%
Naïve Bayes	70.23%	75.4%	84.14%	79.5%
K-Means +Naïve Bayes	97.21%	97.44%	97.44%	97.44%

Source : (Wulandari et al., 2020)

Based on the table and chart above shows the accurate result of K-means and Naïve Bayes collaboration is better than the other algorithm without collaboration. The collaborative method between K-Means and Naïve Bayes which proposed giving 97.44% accurate from the data test. Compared to K-Means and Naïve Bayes, the proposed method increases the accuracy of about 27% of the Naïve Bayes classification.

From Figure 3, we can see that the proposed method used can reduce the error rate of the process by 27%



Source : (Wulandari et al., 2020)

Figure 3. Accuracy Result

Clusters produced by k-Means classification still use random centroid selection. That is because the first step of the K-Means method is by selecting a centroid value. The result of the initial random centroid determination method according to the K-Means original technique is faster and easier. Then by utilizing the Naive Bayes classify the next test data.

In the stage of using the K-Means algorithm produces majority data that has been grouped based on the original class. Data prediction in new test clusters using the Naive Bayes classification is related to insignificant data clusters, the accuracy results are increasing although not significant. This is the impact of Naive Bayes classification which requires sufficient training data to carry out an optimal classification process. The original K-Means randomly produce an initial centroid which makes the quality of grouping accuracy dependent on the initial centroid. When centroids are incorrect, the accuracy results will be relatively lower.

By using a combination of the K-Means clustering algorithm and the Naive Bayes classification, the process of determining the initial centroid K-Means also influences the accuracy results. However, the impact can be reduced by the addition of the Naive Bayes classifier which results in better accuracy, although not better than the proposed method.

CONCLUSION

This study proposed a combination of K-Means clustering and Naive Bayes classification (K-Means Bayes) data mining techniques in students' academic performance data to produce higher data accuracy.

The accuracy of the proposed method combination is 97.44%. Calculations with the K-Means algorithm and calculations with the Naive Bayes algorithm, the proposed method gives better results. Although the initial centroid determination in the K-Means method is carried out randomly, the impact can be reduced by the addition of the Naive Bayes Classifier method resulting in better accuracy and increasing the accuracy of the existing methods.

Compared to K-Means and Naïve Bayes, the proposed method increases the accuracy of about 27% of the Naïve Bayes algorithm and about 23% of the K-Means algorithm. With the results obtained, it can be concluded that the proposed method can improve predictions of student academic performance data. The initial centroid determination in the K-Means method affects that the quality of grouping accuracy depends on the initial centroid.

For further research, the techniques needed to determine the initial centroid that will be used to improve the curation of the data to be predicted.

REFERENCES

Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, *5*(2), e01250. https://doi.org/10.1016/j.heliyon.2019.e01250

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. https://doi.org/10.1016/j.compedu.2017.05.007

Bansode, J. (2016). *Mining Educational Data to Predict Student's Academic Performance*. *January*, 1–5.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. Van. (2019). Educational data mining: Predictive analysis of the academic performance of public school students in the capital of Brazil. *Journal of Business Research*, *94*(August 2017), 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Hamsa, H., Indiradevi, S., & Kizhakkethottam, J. J. (2016). Student Academic Performance Prediction Model Using Decision Tree and

Fuzzy Genetic Algorithm. *Procedia Technology*, *25*, 326–332. https://doi.org/10.1016/j.protcy.2016.08.114

Handayani, F., & Pribadi, S. (2015). Implementasi Algoritma Naive Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110. *Jurnal Teknik Elektro*, *7*(1), 19–24.

Hussain, S., Dahan, N. A., Ba-alwi, F. M., & Ribata, N. (2018). *Educational Data Mining and Analysis of Students ' Academic Performance Using WEKA*. *9*(2), 447–459. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

Miguéis, V. L., Freitas, A., Garcia, P. J. V., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modeling approach. *Decision Support Systems*, *115*(July 2017), 36–51. https://doi.org/10.1016/j.dss.2018.09.001

Muda, Z., Yassin, W., Sulaiman, M. N., & Udzir, N. I. (2011). Intrusion detection based on K-Means clustering and Naïve Bayes classification. *7th International Conference on Information Technology in Asia*.

Risnawati. (2018). Analisis Kelulusan Mahasiswa Menggunakan Algoritma C.45. *Jurnal Mantik Penusa*, *2*(1), 71–76.

Satrianansyah, S., & Wulandari, C. (2019). Penerapan Algoritma Naive Bayes Untuk Memprediksi Kelulusan Ujian Siswa Berbasis Web Pada Smk. *Seminar Nasional AVoER XI 2019*, 23–24.

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting the academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*(October 2019), 106189. https://doi.org/10.1016/j.chb.2019.106189

Widaningsih, S. (2019). Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naïve Bayes, Knn Dan Svm. *Jurnal Tekno Insentif*, *13*(1), 16–25. https://doi.org/10.36787/jti.v13i1.78

Wulandari, D. A. N., Annisa, R., & Yusuf, L. (2020). *An Educational Data Mining For Student Academic Prediction Using K-Means Clustering And Naïve Bayes Classifier*.