G-Tech : Jurnal Teknologi Terapan
Vol. 9, No. 1, 2025, p. 511-519
E-ISSN: 2623-064X | P-ISSN: 2580-8737

Application of K-Means and Naïve Bayes Algorithms for Prediction Model of Student Interest Concentration (Case Study: Amikom University Yogyakarta)

Danang Eko Prayogo[1]✉, Kusrini[2]

[1,2] Magister Of Informatics Engineering, University AMIKOM Yogyakarta, Yogyakarta, Indonesia

✉ **Corresponding Author :** Danang Eko Prayogo (e-mail: danangekoprayogo@students.amikom.ac.id)

Article Information

Article History

Received : January 10, 2025
Revised : January 26, 2025
Published : January 30, 2025

***Keywords*:**

Student Concentration;
K-Means;
Naïve Bayes;
Amikom University Yogyakarta.

ABSTRACT

Amikom University Yogyakarta, has a Master of Informatics Engineering study program with three concentrations of specialization: Business Intelligence, Digital Information Intelligence, and Intelligence Animation. The choice of concentration by prospective students has been based on subjectivity, not on competence or work experience. To overcome this, this research proposes an algorithm-based concentration prediction and recommendation model to help prospective students choose the appropriate concentration. The dataset is obtained through questionnaires collected from active and inactive students. This research uses the K-Means algorithm for clustering raw data (unsupervised) in order to generate target classes, which are then classified using Naïve Bayes. The clustering process determines concentration labels such as Business Intelligence and others, while the SMOTE technique is used to balance the dataset to avoid data imbalance problems. This approach aims to produce more objective and accurate recommendations in determining student concentrations, reducing the tendency of subjectivity, and increasing the relevance of student competencies to the chosen field of specialization. From this research, the K-Means DBI score is 0.277 and the Naïve Bayes prediction accuracy score is 89%. This research aims to produce more objective and accurate recommendations in determining student concentrations, reducing subjectivity, and increasing the relevance of student competencies to the chosen field of specialization. The proposed model is expected to help universities in designing a more targeted admission strategy, as well as supporting students in making academic decisions that are in accordance with their abilities and interests, thereby increasing the effectiveness of the learning process and the suitability of graduates to the needs of the world of work.

INTRODUCTION

Amikom University Yogyakarta is a university with an information technology background, so the study programs in it are study programs that lead to information technology, one of which is the Master of Informatics Engineering study program. The study program has a concentration of student interests that must be chosen by prospective students when registering. There are 3 concentrations that can be chosen by prospective students, namely: Business Intelligence, Digital Information Intelligence and Intelligence Animation. The choice of

This is an open access article under the CC BY license

concentration made by prospective students will have implications for the composition of the course curriculum that they will receive during the lecture.

This fact and condition has a weakness gap, namely, almost all students choose a concentration based on subjectivity, not based on their competence or work experience. Therefore, the author proposes an algorithm model that can predict and provide recommendations to prospective students on what concentration to take based on a number of certain competencies and interests they have.

The dataset in this study was obtained through distributing questionnaires to active and inactive students in the Master of Information Engineering program. The attributes collected include educational background, work experience, technical skills, interests, and career goals. This data is used to build a more objective and relevant study concentration prediction model.

In this research, it is proposed to use the K-Means and Naïve Bayes algorithms as a prediction model for students' concentration of interest. The questionnaire data is unsupervised raw data, so it is very necessary to do clustering first before prediction. This aims to make the dataset have a target class and become supervised. The author can actually take data as well as the concentration chosen by students, but as described in the previous paragraph, the choice of concentration still has a subjective tendency, for example: prospective students want to take the Intelligence Animation concentration because they are interested in developing 3D animation films with certain theories as material for their final project, even though their abilities and competencies are in Digital Information Intelligence. This will certainly greatly affect the overall dataset pattern. So it is determined that in this research two clustering algorithms K-Means and Naïve Bayes classification are used. The results of the K-Means experiment will produce a dataset with the target class, the dataset will be interpreted by the cluster, which cluster is Business Intelligence and so on. Furthermore, the dataset will be balanced using the SMOTE technique. This is so that the class data is not imbalanced which will certainly affect the accuracy of the classification results and tend to result in overfit. Only then will the dataset be classified using the Naïve Bayes algorithm.

In previous studies, there is relevant research, where prediction of student concentration interests is carried out, the features used are more likely to be the grades of disciplinary courses or fields. In this study, the Naïve Bayes algorithm was used, in the target class, the labeling was done manually. From this research, an accuracy of 84.27% was obtained (Farid et al, 2022). There have been so many problems that have been solved using the algorithm, the algorithm is claimed to be a popular algorithm not only in its ease of understanding (explainabillity), but also effective and resilient (Wickramasinghe et al, 2021) (Anagora et al, 2022).

As for the use of K-Means itself, previously there were also similar studies, the K-Means algorithm was used for COVID-19 cases in Lampung province, using this algorithm, researchers succeeded in dividing the Lampung province area into 4 clusters of red, orange, yellow and green zones representing areas prone to COVID-19 (Nabila et al, 2021), clustering the nutritional status of toddlers in the Karangsongo area, successfully identifying 5 cluster groups, ranging from malnutrition clusters to obesity, which will indirectly produce insight into other diseases that correlate with nutritional problems in toddlers (Mega, 2015).

The use of both algorithms: K-Means and Naïve Bayes is a good integration, where the two algorithms are algorithms that have different scopes, the integration between the two has been done by previous research, predicting academic performance learning, where the dataset is unsupervised and requires a target class for Naïve Bayes classification, therefore the labeling process is carried out with the K-Means algorithm (Ali et al, 2020) (Fadhil, 2021).

From the description that the author describes, the author proposes to make a predictive model as a recommendation for choosing a concentration interest that suits prospective students using the K-Means and Naïve Bayes algorithms. K-Means and Naïve Bayes algorithms are data mining algorithms with different approaches. If K-Means algorithm is a clustering algorithm, then

Naïve Bayes algorithm is a classification algorithm. Both are used to create a predictive model of student interest concentration.

RESEARCH METHODS

Data Cleaning

The questionnaire dataset will be cleaned first. The author did some cleaning such as: casefolding (this is done only to tidy up the dataset by making all data font characters small), feature engineering (selecting which features will be used as attributes for clustering), data normalization (changing and equalizing data that has the same essence of meaning but different values), data discretization (categorizing data from a unique numerical attribute) and finally performing one hot encoding (converting categorical data into numerical data that represents the data).

Dimensionality Reduction with AutoEncoder

The dataset used has 17 attributes, making it high dimensional. This high dimensionality can affect the performance of the K-Means algorithm in clustering and result in a less than optimal evaluation. To overcome this, dimensionality reduction is performed using AutoEncoder, a Deep Learning-based algorithm that is able to reduce the dimensionality of the dataset without losing important information, thus helping to improve the quality of clustering and visualization of data patterns.

Clustering K-Means

The cleaned data is then clustered using the K-Means algorithm, the author uses the Scikit Learn library and the Google Collaboratory platform as the experimental medium for this research. The K-Means algorithm itself is a clustering algorithm with a partitioning-based clustering approach that is most popularly used by scientists and researchers when clustering data, which will divide data into several classes (Hutagalung et al, 2021) (Gustientiedina et al, 2019). The K-Means algorithm clustering process itself uses the Euclidean formula which is shown in formula no 1.

$$\boldsymbol{Min} \sum_{\boldsymbol{k}}^{\boldsymbol{k}} = \boldsymbol{d_{ik}} = \sqrt{\sum_{\boldsymbol{j}}^{\boldsymbol{m}} (\boldsymbol{C_{ij}} - \boldsymbol{C_{kj}})^2} \quad (1)$$

Description:

dik = distance of each data object point to the data center point

ckj = data object point

By using the Eucliden formula, the distance of each data object point to its centroid that has been determined randomly is calculated. After clustering, the next thing to do is to evaluate the clustering model created using an evaluation matrix, such as: DBI (Harahap et al, 2021) and Sillhoute (Febrianto et al, 2021). In the DBI evaluation, if the accuracy value is close to 0, the better the model evaluation, on the other hand, in Sillhoute, if the accuracy value is greater, the better the model evaluation. To determine whether the clustering model is good or not, not only from the accuracy of the model, but also analyze the visualization of the scatterplot generated by the model, whether it overlaps or not, if it overlaps, the clustering model is not good at clustering and needs to be changed to another clustering algorithm model (Singh, 2019).

SMOTE Technique

The dataset generated from K-Means clustering is then subjected to data balancing using SMOTE. SMOTE will work by creating synthetic data by studying existing data. The data size of each class will be equalized to the size of the largest class data, so the frequency distribution of each class will be equal and normal. If ignored, this can affect the classification evaluation results and tend to overfit. This technique works by generating new data based on interpolation between existing minority data. Here is how SMOTE overcomes data imbalance:

1. Minority Class Identification : SMOTE focuses on classes with less data (minority classes) than majority classes.
2. Nearest Neighbor Selection : For each sample in the minority class, SMOTE selects several nearest neighbors using algorithms such as K-Nearest Neighbors (KNN).
3. Interpolation of New Data : New samples are generated by taking values that lie between the original sample and its nearest neighbors, based on linear interpolation. This creates data variation without simply duplicating the original sample.
4. Synthetic Data Addition : The resulting synthetic data is then added to the dataset, thus increasing the amount of data in the minority class until it reaches a more balanced proportion.

With this technique, SMOTE helps machine learning algorithms avoid bias towards the majority class, improves the model's performance in predicting the minority class, and prevents overfitting that often occurs when only adding data by direct duplication.

Classification Naïve Bayes

The results of the dataset that has been balanced using SMOTE will have a normal frequency distribution, then the next classification experiment will be carried out using the Naïve Bayes algorithm. At this stage, the dataset is again carried out feature engineering, feature selection of attributes that will be used for the classification process.

Similar to the clustering stage, the classification process uses the Scikit Learn library and the Google Collaboratory media platform.

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)} \quad (2)$$

Description:

P(C | X) = Probability of class C based on data X

P(X | C) = Probability of data X if class C is known

P(C) = Probability of class C

P(X) = Probability of data X

The Naïve Bayes algorithm will calculate the probability of each attribute based on the target class, so that if a test data with attribute X is known, then the probability for the target class is Y (Widodo et al, 2021) (Byna et al, 2020). The Naïve Bayes algorithm itself is a classification algorithm with a probabilistic based learning approach, which is an algorithm that solves data classification problems based on the assumption of calculating probability scores. Within the scope of the classification field itself, there are many approaches such as: logic based learning, instance based learning, rule based learning, SVM and others (Kotsiantis, 2007) (Amalina et al., 2020).

RESULTS AND DISCUSSION

Data Cleaning

Data cleaning at this stage includes the process:

1. Case folding
2. Feature selection
3. Data normalization
4. Data discretization
5. One Hot Encodin

These stages will make the dataset new and ready for experimentation.

Dimensionality Reduction with AutoEncoder

The author performs dimension reduction to simplify the dimensions of the dataset. This is so that the K-Means algorithm is able to do Clustering well. The author uses the AutoEncoder algorithm to reduce the dimensions of the dataset as shown in Figure 1.

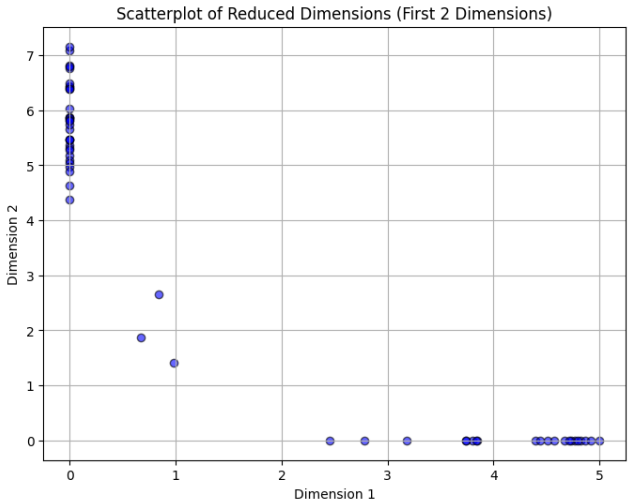

Figure 1. Visualization of dataset scatterplot with AutoEncoder

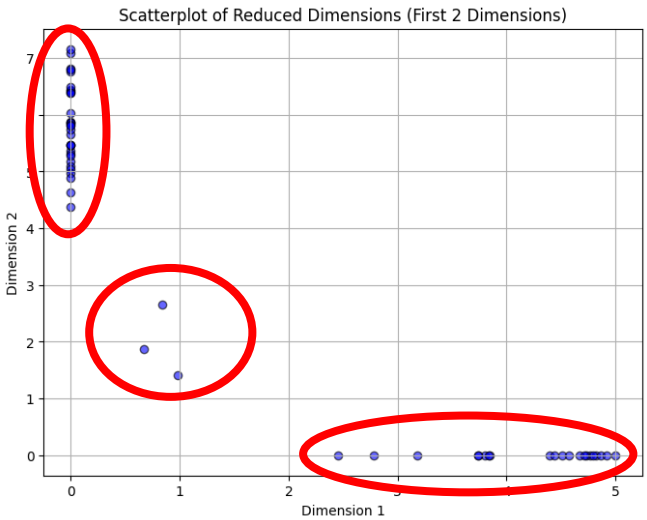

Figure 2. Visualization of dataset scatterplot with AutoEncoder

Figure 1 shows the results of visualizing the dataset scatterplot with AutoEncoder, as previously described, the author divides the dataset into 3 clusters (K = 3) which corresponds to the concentration of student interests. In the figure, it can be analyzed if it will be divided into 3 clusters, then 3 groups should be formed according to Figure 2.

Clustering K-Means

The author performs clustering based on the dataset that has been reduced based on the 1st figure. By using K=3, the clustering results are obtained as shown in Figure 3. From the K-Means clustering results shown in Figure 3, 3 color groups are obtained: green, blue and red. In the visualization, the centroid position is shown in the star symbol in the middle of the data group.

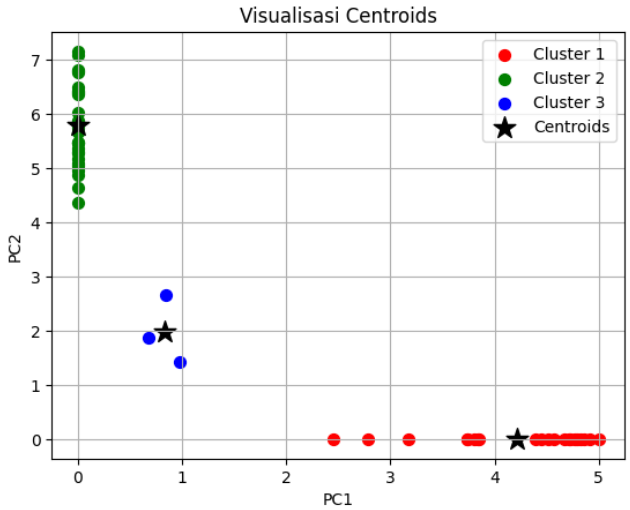

Figure 3. K-Means algorithm clustering results K=3

From the clustering results, the clustering image is in accordance with Figure 2, so it is considered that the clustering results are successful and good. A DBI (Davies Bouldin Index) score of 0.277 was obtained. The results of the clustering are then downloaded as Excel.

SMOTE Technique

Based on the clustering results, from a total of 55 data, it is divided into 3 clusters, namely cluster 0 of 21 data records, cluster 1 of 31 data records and cluster 2 of 3 data records. The cluster results are then interpreted, where the BI concentration group is a group with characteristics of having high coding and mathematical analytical skills, while the DTI group is a group with characteristics of good coding skills, followed by multimedia skills, art in multimedia, technopreneurship and SI / IT governance. While IA is a group with characteristics of having high multimedia skills followed by high art in multimedia. Here are the statistics of the number of each class.

Tabel 1. Statistics on the number of each class

No	Concentration	Quantity per Record
1	BI	31
2	DTI	21
3	IA	3

From the amount shown in table 1, if the dataset is classified, the accuracy performance will decrease and tend to overfit, so the author then balances the data using SMOTE. By using SMOTE, a number of random records in the DTI and IA classes are generated and equalized with the highest number of records, namely BI, so that they become the same, as shown in the following figure:

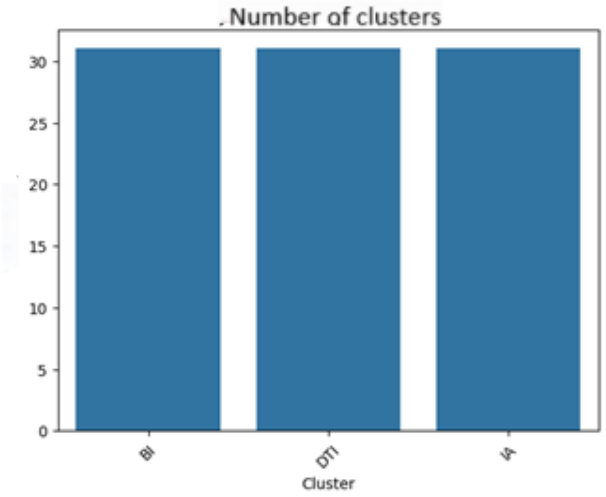

Figure 4. Dataset balancing results with SMOTE

Figure 4 shows the number of records that are the same between BI, DTI and IA classes, 31 records for each class.

Classification Naïve Bayes

Furthermore, the author performs classification based on data that has been balanced, from Naïve Bayes classification, an accuracy of 0.89 or 89% is obtained, with an error value of 11%. So that the prediction model can be said to be good. Then the value can be considered good because:

1. Low Error Rate: With an error rate of only 11%, the model can predict students' concentration of interest quite accurately in most cases.
2. Standard Model Performance: In the context of educational applications, 89% accuracy is usually considered quite high as the data collection process and student behavior tend to be complex and difficult to predict.
3. Use of the Smoothed Dataset: Dimensionality reduction with AutoEncoder and dataset balancing with SMOTE have helped to create a more representative dataset and support the performance of Naïve Bayes.

The author also predicts testing 10 random data records and gets the following results:

Tabel 2. Prediction testing 10 random data

No	Actual Result	Predicted Result
1	1	1
2	1	1
3	2	2
4	2	2
5	0	0
6	0	0
7	0	0
8	0	0
9	0	0
10	1	1

Table 2 shows that from the prediction test of 10 random data, the algorithm model successfully predicted all the actual result data. The numerical data is still in the One Hot Encoding dimension.

CONCLUSION

Based on the experiments conducted, the author succeeded in building a prediction model for the concentration of students' fields of interest with a Naïve Bayes algorithm classification accuracy score of 89% with an error score of 11%. This explains that the prediction model made is good. In the dimension reduction that occurs before K-Means clustering, the AutoEncoder algorithm successfully reduces the dimensions of the dataset and maps it into 3 groups. The 3 groups have been successfully detected by the K-Means Clustering algorithm with a DBI score of 0.277, for the DBI score it is said to be very good, because if the DBI is closer to 0 then the evaluation of the resulting cluster model is getting better, and the K-Means clustering results do not overlap. Recommendations for Future Research:

1. Model Performance Improvement: Future research can explore other algorithms, such as Random Forest or Gradient Boosting, to compare prediction accuracy with existing models.
2. Larger Dataset Evaluation: Expand the dataset by adding data from the next academic year or from other universities to improve the generalization of the model.
3. Integration of Other Factors: Adding additional attributes, such as satisfaction level or career success of graduates, to enrich the prediction model.
4. Recommendation System Implementation: Develop a web-based system or app to recommend concentrations directly to prospective students, which can be integrated into the admission process.
5. Exploration of Other High-Dimensional Algorithms: In addition to AutoEncoder, algorithms such as t-SNE or PCA can be tested to understand the comparison of dimensionality reduction results.

REFERENCES

Ali, Z. M., Hassoon, N. H., Ahmed, W. S., & Abed, H. N. (2020). The application of data mining for predicting academic performance using K-means clustering and Naïve Bayes classification. *International Journal of Psychosocial Rehabilitation, 24*(3), 2143–2151. https://doi.org/10.37200/ijpr/v24i3/pr200962

Amalina, N., Suhaimi, D., & Abas, H. (2020). A systematic literature review on supervised machine learning algorithms. *Perintis eJournal, 10*(1).

Anagora, R., Taufiq, R., Dedi Jubaedi, A., Wirawan, R., & Syah Putra, A. (2022). The classification of phishing websites using Naive Bayes classifier algorithm. *International Journal of Science.* Retrieved from http://ijstm.inarah.co.id

Byna, A., & Basit, M. (2020). Application of Adaboost method to optimize stroke disease prediction with Naïve Bayes algorithm. *Sisfokom Journal (Information and Computer Systems), 9*(3), 407–411. https://doi.org/10.32736/sisfokom.v9i3.1023

Fadhil, Z. M. (2021). Hybrid of K-means clustering and Naive Bayes classifier for predicting performance of an employee. *Journal of Information Technology, 9*(2), 799–807.

Farid, M., Wibowo, S., Puspitasari, N. F., & Satya, B. (2022). Application of data mining and Naïve Bayes algorithm for student concentration selection using classification method. *Journal of Information Systems Management (JOISM), 3*(2).

Febrianto, A., Achmadi, S., & Sasmito, A. P. (2021). Application of K-Means method for clustering visitors to ITN Malang library. *Jurnal Mahasiswa Teknik Informatika, 5*(1).

Gustientiedina, G., Adiya, M. H., & Desnelita, Y. (2019). Application of K-Means algorithm for drug data clustering. *National Journal of Information Technology and Systems, 5*(1), 17–24. https://doi.org/10.25077/teknosi.v5i1.2019.17-24

Harahap, F. (2021). Comparison of K-Means and K-Medoids algorithms for clustering classes of Tunagrahita students. *Applied Informatics Nusantara, 2*(4).

Hediyati, D., & Suartana, I. M. (2021). Application of principal component analysis (PCA) for dimension reduction in the clustering process of agricultural production data in Bojonegoro Regency.

Hubbs, C. D., Li, C., Sahinidis, N. V., Grossmann, I. E., & Wassick, J. M. (2020). A deep reinforcement learning approach for chemical production scheduling. *Computers and Chemical Engineering, 141.* https://doi.org/10.1016/j.compchemeng.2020.106982

Hutagalung, J., Ginantra, N. L. W. S. R., Bhawika, G. W., Parwita, W. G. S., Wanto, A., & Panjaitan, P. D. (2021). COVID-19 cases and deaths in Southeast Asia clustering using K-Means algorithm. *Journal of Physics: Conference Series, 1783*(1). https://doi.org/10.1088/1742-6596/1783/1/012027

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31.*

Liaw, L. C. M., Tan, S. C., Goh, P. Y., & Lim, C. P. (2025). A histogram SMOTE-based sampling algorithm with incremental learning for imbalanced data classification. *Information Sciences, 686,* 121193.

Lubis, A., Irawan, Y., Junadhi, J., & Defit, S. (2024). Leveraging K-Nearest Neighbors with SMOTE and boosting techniques for data imbalance and accuracy improvement. *Journal of Applied Data Sciences, 5*(4), 1625–1638.

Mega, W. (2015). Clustering using K-Means methods to determine literary nutrition status. *National Journal of Information Technology, 15*(2).

Nabila, Z., Rahman Isnain, A., & Abidin, Z. (2021). Data mining analysis for clustering COVID-19 cases in Lampung Province with K-Means algorithm. *Journal of Information Technology and Systems (JTSI), 2*(2), 100. Retrieved from http://jim.teknokrat.ac.id/index.php/JTSI

Singh, A. (2019, October 31). Build better and accurate clusters with Gaussian mixture models. *Analytics Vidhya.* Retrieved from https://www.analyticsvidhya.com/blog/2019/10/gaussian-mixture-models-clustering/

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: Applications, variations and vulnerabilities: A review of literature with code snippets for implementation. *Soft Computing, 25*(3), 2277–2293. https://doi.org/10.1007/s00500-020-05297-6

Widodo, Y. B., Anggraeini, S. A., & Sutabri, T. (2021). Design of a web-based diabetes disease diagnosis expert system using the Naive Bayes algorithm. *Journal of Informatics and Computer Technology, 7*(1), 112–123. https://doi.org/10.37012/jtik.v7i1.507