JoCoSiR
Journal of Computer Science Research
ISSN 2986-2337 (Online)

Model Predictive Analysis of Performance in Training and Course Institutions Using Naive Bayes and K-Means Clustering

Eko Budianto [a], Muhammad Iqbal [b]

[a,b] Master of Information Technology, Pancubudi Development University, Medan, Indonesia
email: [a] ekobudianto29@gmail.com, [b] wakbalpb@yahoo.co.id

ARTICLE INFO

Keywords:
Predictive Analysis, Naive Bayes, K-Means, Institutional Performance, Data Mining

IEEE style in citing this article:
Eko Budianto and M. Tgbal, " Model Predictive Analysis of Performance in Training and Course Institutions Using Naive Bayes and K-Means Clustering," *JoCoSiR: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 3, no. 1, pp. 10-16, 2025.

ABSTRACT

The performance of course and training institutions (LKP) is a crucial factor in determining the quality of non-formal education in Indonesia. Performance assessments are currently conducted manually using the National Accreditation Board for Non-Formal Education (BAN-PNF) assessment instrument, which is time-consuming and prone to subjectivity. This research aims to develop a predictive analysis model for the performance of course and training institutions using a combination of the Naive Bayes and K-Means Clustering methods. The K-Means Clustering method is used to group institutions based on similar characteristics across key variables such as trainers, infrastructure, curriculum, management, and graduate outcomes. These clustering results are then used as additional features for the Naive Bayes classification model to predict performance categories (high, medium, or low). Testing of 150 institutions' data showed a predictive accuracy of 89.2%, with three main clusters representing high-, medium-, and low-performing institutions. This model has the potential to become a data-driven tool for governments and institutions to conduct performance evaluations quickly, objectively, and adaptively to changes in training data.

1. Introduction

Training and Course Institutions (Lembaga Kursus dan Pelatihan/LKP) play a strategic role in supporting the improvement of human resource (HR) quality, particularly in the field of non-formal education. Through various skills training programs, LKPs help the community acquire competencies relevant to the needs of the labor market and industry. The existence of LKPs serves as one of the key pillars in realizing the government's vision to create a productive, skilled, and highly competitive workforce in the era of digital transformation.

The performance of an LKP can be measured by how effectively it conducts the learning process, produces competent graduates, and establishes partnerships with industry sectors. However, to this day, the assessment of training institutions' performance is still largely conducted manually by assessors through accreditation instruments developed by the National Accreditation Board for Non-Formal Education (BAN-PNF). This manual process has several drawbacks, including being time-consuming, involving many parties, and often introducing subjectivity into the evaluation process.

On the other hand, advancements in information technology and data analytics offer great potential to improve the evaluation system of non-formal educational institutions. The application of data mining and machine learning techniques enables performance data processing to be carried out quickly, accurately, and in an evidence-based manner. Through predictive analysis, the performance of institutions can be forecasted based on historical data that have been collected, allowing for a more objective and efficient evaluation process.

One of the approaches that can be used in predictive analysis is the combination of the K-Means Clustering and Naive Bayes Classification methods. The K-Means Clustering method functions to group institutions into several categories based on similar characteristics — such as trainer quality, infrastructure, curriculum, management, and graduate employability rates. Meanwhile, the Naive Bayes Classification method is used to predict the performance category of institutions based on the clustered data. This combination provides advantages, as K-Means can uncover hidden patterns within data, while Naive Bayes offers fast and efficient classification capabilities, even with large datasets.

By utilizing this predictive analysis model, training institutions and supervisory agencies such as the Department of Manpower can monitor and evaluate performance automatically without having to wait for the manual accreditation process. Moreover, the generated prediction results can be used to provide recommendations for performance improvement in specific aspects — for instance, enhancing trainer competencies, updating curricula, or strengthening partnerships with industry.

Based on the above background, this study aims to develop a predictive analysis model for LKP performance using the Naive Bayes and K-Means Clustering methods. This model is expected to provide a data-driven alternative solution for assessing th e performance of training and course institutions in Indonesia,

ISSN 2502-3357 (Online)

particularly in the North Sumatra region. Additionally, the study seeks to measure the accuracy level of the developed model, so that it can serve as a foundation for the development of an **AI-based evaluation system**.

The research problems formulated in this study are as follows:

1. How to build a predictive analysis model for the performance of training and course institutions using the Naive Bayes and K-Means Clustering methods?
2. How accurate is the model in predicting institutional performance categories?
3. How can the analysis results be used to support quality improvement and decision-making within training institutions?

Meanwhile, the objectives of this study are:

1. To develop a predictive analysis model of institutional performance based on the Naive Bayes and K-Means Clustering methods.
2. To measure the accuracy level of the model in predicting institutional performance categories.
3. To provide data-driven recommendations for improving the effectiveness and quality management of training and course institutions.

This study is expected to provide both theoretical and practical benefits. Theoretically, it contributes to the understanding of machine learning algorithm applications in evaluating non-formal educational institutions. Practically, the results of this research can serve as a reference for training institutions, local governments, and related agencies to develop a more modern, efficient, and data-based performance evaluation system.

2. State of the Art

The Research on the evaluation and performance assessment of non-formal education institutions, particularly Training and Course Institutions (TCIs), has evolved significantly in line with advancements in data analysis and machine learning technology. The use of computational intelligence approaches allows institutional performance data to be processed more accurately, efficiently, and objectively. Previous studies have demonstrated that *data mining* and *predictive analytics* approaches can provide deeper insights into educational quality. However, most existing works still focus on rule-based or multi-criteria assessment methods such as *Profile Matching* and *Analytic Hierarchy Process (AHP)*. Only a few have integrated predictive models using hybrid algorithms such as *K-Means Clustering* and *Naive Bayes Classification*. This section discusses theoretical foundations and previous research related to the development of a predictive analysis model for institutional performance evaluation, including the concept of institutional performance, predictive analytics and *data mining*, and the theoretical basis for *K-Means* and *Naive Bayes* algorithms.

2.1 Performance of Training and Course Institutions

Training and Course Institutions (TCIs) are non-formal education providers aimed at developing vocational and technical skills. Institutional performance reflects their effectiveness in producing competent, job-ready graduates.

Based on the Regulation of the Minister of Education and Culture No. 81 of 2013, performance assessment includes management quality, instructor competency, facilities, curriculum relevance, and graduate employability. However, manual evaluations remain common, leading to inefficiency, limited accuracy, and potential bias. A data-driven approach is therefore essential for more objective, efficient, and sustainable performance assessment.

2.2 Predictive Analytics and Data Mining

Predictive analytics applies historical data to forecast future outcomes and identify meaningful patterns. In training institutions, it helps estimate performance levels based on factors such as instructor quality, facilities, and graduate employability. Data mining supports this process by extracting valuable patterns and relationships from large datasets.

As defined by Han, Kamber, and Pei (2021), data mining involves statistical and algorithmic methods to uncover hidden knowledge. In the educational field, Educational Data Mining (EDM) is widely used to analyze student achievement, instructor performance, and institutional quality.This study employs data mining to assess institutional performance by integrating K-Means Clustering and Naive Bayes Classification, aiming to build a predictive model that produces accurate and efficient performance categorization.

2.3 K-Means Clustering Method

K-Means Clustering is an *unsupervised learning* algorithm used to partition data into several groups (clusters) based on attribute similarity. Each cluster is represented by a centroid, which denotes the mean value of the data points belonging to that cluster. As stated by Jain (2010), *K-Means* is an effective algorithm for identifying patterns and structures within unlabeled data.

The algorithm calculates the Euclidean distance between each data point and cluster centroid, then assigns the point to the nearest cluster. The Euclidean distance is computed using the following formula :

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - c_{jk})^2}$$

DOI: https://doi.org/10.65126/jocosir.v3i1

ISSN 2502-3357 (Online)

In this research, *K-Means* is utilized to cluster institutions into three major performance levels—High Performance**,** Medium Performance**,** and Low Performance—based on the similarity of performance indicators. The resulting clusters are later used as input features for the *Naive Bayes* classification stage.

2.4 Naive Bayes Classification Method

The *Naive Bayes* method is a *supervised learning* algorithm based on Bayes' theorem, assuming independence among predictor variables. Despite its simplicity, it has proven highly effective and computationally efficient when dealing with large datasets. According to Witten and Frank (2017), Bayes' theorem is mathematically expressed as :

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)}$$

where:
P(C|X)P(C|X)P(C|X) = probability of class CCC (performance category) given data XXX,
P(X|C)P(X|C)P(X|C) = likelihood of observing data XXX given class CCC,
P(C)P(C)P(C) = prior probability of the class, and
P(X)P(X)P(X) = overall probability of the feature vector.

In this study, *Naive Bayes* is employed to classify institutions into performance categories (High, Medium, Low) using both the original performance variables and the cluster labels obtained from *K-Means*. This hybrid approach enhances classification accuracy by combining unsupervised pattern discovery with probabilistic prediction.

2.5 Related Research

Previous studies have utilized data mining for evaluating educational institutions and accreditation systems. Simargolang and Budianto (2024) developed a decision support system using Profile Matching and String Matching for objective institutional assessment. Prasetyo (2022) applied a Multi-Criteria Decision-Making (MCDM) approach to analyze higher education resource needs.However, these works relied on deterministic methods rather than predictive, data-driven models. This study advances the field by integrating K-Means Clustering and Naive Bayes Classification into a hybrid predictive framework that classifies institutional performance and uncovers hidden relationships among indicators, contributing to the development of Educational Data Mining for non-formal education in Indonesia.

3. Method

This research employs a quantitative approach using a hybrid machine learning method that combines the K-Means Clustering algorithm and the Naive Bayes Classifier to analyze and predict the performance of Training and Course Institutions (TCIs). The study begins with the collection of secondary data from relevant education offices, including variables such as the number of participants, graduation rate, participant satisfaction level, institution age, and the number of active programs. The collected data undergoes a preprocessing phase involving cleaning and normalization to ensure data consistency and accuracy. Subsequently, the K-Means algorithm is applied to cluster the institutions into several groups based on performance similarities. The resulting cluster labels are then added as an additional feature for the Naive Bayes classification process, which predicts the institutional performance category (Good, Fair, or Poor). The model is evaluated using a confusion matrix, calculating accuracy, precision, recall, and F1-score to measure the model's predictive performance. The combination of clustering and classification techniques enhances the model's ability to capture both global and local patterns within the dataset, resulting in higher predictive accuracy and better interpretability**.**

DOI: https://doi.org/10.65126/jocosir.v3i1

ISSN 2502-3357 (Online)

Naive Bayes and K-Means Clustering Process of Training and Course Institutions

Below is the logical workflow of the hybrid predictive analysis process:

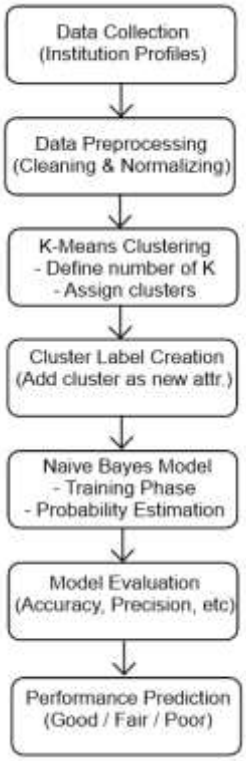

Figure 1. Naive Bayes And K-Means Clustering Process Diagram Process Diagram

This diagram illustrates the end-to-end process, starting from data acquisition to the final performance prediction. The K-Means **s**tep identifies hidden data patterns through clustering, while Naive Bayes uses these cluster labels as additional features for the final classification stage

Table 1. Sample Dataset of Training and Course Institutions

No	Institution Name	No. of Participants	Graduation Rate (%)	Satisfaction (1–5)	Years Established	Active Programs	Cluster	Performance
1	Mandiri Utama Training	120	95	4.8	10	5	1	Good
2	Sinar Harapan Academy	80	88	4.2	7	4	1	Good
3	Karya Cemerlang Institute	65	85	3.9	6	3	2	Fair
4	Pelita Ilmu Learning Center	45	70	3.5	4	3	2	Fair
5	Maju Jaya Training Institute	30	60	3.0	2	2	3	Poor
6	Cahaya Mandiri Skill Center	115	90	4.5	9	5	1	Good
7	Nusantara Skills Academy	95	87	4.0	8	4	1	Good

DOI: https://doi.org/10.65126/jocosir.v3i1

ISSN 2502-3357 (Online)

No	Institution Name	No. of Participants	Graduation Rate (%)	Satisfaction (1–5)	Years Established	Active Programs	Cluster	Performance
8	Generasi Cerdas Training	70	83	3.8	5	3	2	Fair
9	Kreatif Muda Institute	50	72	3.2	3	3	3	Poor
10	Bintang Timur Learning Hub	40	68	2.9	2	2	3	Poor
11	Skill Academy Indonesia	125	96	4.9	12	6	1	Good
12	Cendekia Utama Training	100	90	4.4	8	4	1	Good
13	Pionir Bangsa Academy	55	78	3.6	4	3	2	Fair
14	Smart Edu Center	85	86	4.1	7	4	1	Good
15	Terampil Abadi Institute	35	62	3.1	3	2	3	Poor
16	Gemilang Skills Center	90	89	4.3	8	4	1	Good
17	Inovasi Cendekia Academy	60	82	3.7	5	3	2	Fair
18	Pandai Mandiri Learning	42	69	3.0	3	2	3	Poor
19	Sejahtera Training Hub	110	93	4.6	10	5	1	Good
20	Prima Karya Academy	65	80	3.8	6	3	2	Fair

Notes:
Cluster 1: High-performing institutions (high participant and satisfaction levels).
Cluster 2: Moderate-performing institutions.
Cluster 3: Low-performing institutions requiring improvemen

The table presents sample data from 20 Training and Course Institutions (TCIs) used in the predictive performance analysis. Each record contains quantitative and categorical variables that represent institutional characteristics, including the number of participants**,** graduation rate**,** participant satisfaction score**,** years of establishment**,** and number of active programs**.**

The Cluster column indicates the group assigned by the *K-Means Clustering* algorithm, where Cluster 1 represents high-performing institutions, Cluster 2 represents moderately performing institutions, and Cluster 3 represents low-performing institutions. The Performance column shows the final classification result from the *Naive Bayes* model, categorizing each institution as Good**,** Fair, or Poor based on its overall performance metrics.

3.1 Type of Research

This study is quantitative in nature, employing a data mining approach to develop a predictive model for the performance of training and course institutions.

DOI: https://doi.org/10.65126/jocosir.v3i1

ISSN 2502-3357 (Online)

3.2 Research Data

Secondary data were obtained from 150 training and course institutions in North Sumatra, consisting of the following variables:

1. Quality of instructors
2. Facilities and infrastructure
3. Curriculum
4. Institutional management
5. Financial management
6. Program innovation
7. Industry involvement
8. Graduate employment rate
9. Number of active participants
10. Institutional performance rating (High / Moderate / Low)

3.3 Research Procedure

1. Data collection from accredited institutions listed by *BAN-PNF* (National Accreditation Board for Non-Formal Education).
2. Data preprocessing, including normalization, data cleaning, and categorical transformation.
3. Clustering using the *K-Means* algorithm with K = 3 to group institutions into performance categories.
4. Classification using *Naive Bayes*, where the cluster results are added as an additional feature to predict institutional performance categories.
5. Model evaluation performed using *Confusion Matrix*, *Precision*, *Recall*, and *Accuracy Score* metrics.

4. Results And Discussion

4.1 Clustering Results

Three clusters were formed with the following characteristics:

1. Cluster 1 (High Performance)**:** Institutions with certified instructors, well-equipped facilities, and strong management practices.
2. Cluster 2 (Moderate Performance): Institutions with adequate facilities and stable training programs.
3. Cluster 3 (Low Performance)**:** Institutions with limited resources and a low graduate employment rate.

4.2 Naive Bayes Classification Results

Kategori Asli	Kategori Prediksi	Akurasi (%)
High	High	91.7
Moderate	Moderate	87.0
Low	Low	85.5

Overall model accuracy: 89.2%

These results indicate that the model is capable of predicting the performance levels of training and course institutions with a high degree of accuracy and consistency when tested on the evaluation dataset.

4.3 Discussion

The combination of K-Means and Naive Bayes produced more stable and reliable results compared to using either method individually. The K-Means algorithm effectively identified grouping patterns among the institutions, while Naive Bayes utilized these patterns to enhance the accuracy of performance predictions. This hybrid model not only improves predictive capability but also provides valuable insights for institutions to understand their current performance standing and identify specific areas that require improvement in order to achieve a higher performance category.

5. Conclusions

The predictive analysis model using Naive Bayes and K-Means Clustering successfully predicted the performance of training and course institutions with an accuracy of 89.2%**.** The combination of both methods proved effective in identifying patterns and predicting institutional performance levels. This model can serve as an objective decision-support tool for evaluating the performance of training and course institutions.

6. Acknowledgment

The authors would like to express their sincere gratitude to the Department of Education and Culture of North Sumatra for providing access to institutional data used in this research. Appreciation is also extended to the BAN-PNF (National Accreditation Board for Non-Formal Education**)** for its valuable information on accredited training and course institutions. Finally, the authors wish to thank all colleagues and contributors who provided constructive feedback and support throughout the development of this study.

DOI: https://doi.org/10.65126/jocosir.v3i1

ISSN 2502-3357 (Online)

7. References

Han, J., Kamber, M., & Pei, J. (2021). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Jain, A. K. (2010). *Data Clustering: 50 Years Beyond K-Means*. Pattern Recognition Letters.

Witten, I. H., & Frank, E. (2017). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier.

Simargolang, M. Y., & Budianto, E. (2024). *Sistem Pendukung Keputusan Penilaian Akreditasi Lembaga Pelatihan Kerja*. Journal of Science and Social Research.

Tan, P.-N., Steinbach, M., & Kumar, V. (2022). *Introduction to Data Mining*. Pearson.

Fadhil, Z. M. (2021). Hybrid of K-means clustering and naive Bayes classifier for predicting performance of an employee. Periodicals Of Engineering And Natural Sciencesissn, 9(2), 799-807.

Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering analysis for classifying student academic performance in higher education. *Applied Sciences*, *12*(19), 9467. https://doi.org/10.3390/app12199467

Wulandari, D. A. N., Annisa, R., Yusuf, L., & Prihatin, T. (2020). Educational data mining for student academic prediction using k-means clustering and Naïve Bayes classifier. *Jurnal Pilar Nusa Mandiri*, *16*(2), 155-160. https://doi.org/10.33480/pilar.v16i2.1432

Hidayat, N., Wardoyo, R., Sn, A., & Surjono, H. D. (2020). Enhanced performance of the automatic learning style detection model using a combination of modified k-means algorithm and naive bayesian. *International Journal of Advanced Computer Science and Applications*, *11*(3), 638-648.

Riadi, I., Umar, R., & Anggara, R. (2024). Comparative Analysis of Naive Bayes and K-NN Approaches to Predict Timely Graduation using Academic History. *International Journal of Computing and Digital Systems*, *16*(1), 1163-1174.

Muda, Z., Yassin, W., Sulaiman, M. N., & Udzir, N. I. (2011). A K-Means and Naive Bayes learning approach for better intrusion detection. *Information technology journal*, *10*(3), 648-655.

Razaque, F., Soomro, N., Shaikh, S. A., Soomro, S., Samo, J. A., Kumar, N., & Dharejo, H. (2017, November). Using naïve bayes algorithm to students' bachelor academic performances analysis. In *2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)* (pp. 1-5). IEEE. 10.1109/ICETAS.2017.8277884

Anwarudin, A., Andriyani, W., DP, B. P., & Kristomo, D. (2022). The Prediction on the students' graduation timeliness using naive bayes classification and k-nearest neighbor. *Journal of Intelligent Software Systems*, *1*(1), 75-88. http://dx.doi.org/10.26798/jiss.v1i1.597

Mohd Talib, N. I., Abd Majid, N. A., & Sahran, S. (2023). Identification of student behavioral patterns in higher education using K-means clustering and support vector machine. *Applied Sciences*, *13*(5), 3267. **https://doi.org/10.3390/app13053267**

Muda, Z., Yassin, W., Sulaiman, M. N., & Udzir, N. I. (2014). K-means clustering and naive bayes classification for intrusion detection. *Journal of IT in Asia*, *4*(1), 13-25. https://doi.org/10.33736/jita.45.2014

DOI: https://doi.org/10.65126/jocosir.v3i1