

Clustering-Based Identification of Governance Risks in Campus Environments

Sumiadji¹, Usman Nurhasan^{2*}, Zainal Abdul Haris³

¹Department of Information Technology, State Polytechnic of Malang, Indonesia

^{2,3}Department of Accounting, State Polytechnic of Malang, Indonesia

¹sumiadji@polinema.ac.id, ²usmannurhasan@polinema.ac.id, ³zainaloni999@gmail.com



*Corresponding Author

Article History:

Submitted: 04-11-2025

Accepted: 07-11-2025

Published: 21-11-2025

Keywords:

Governance Risk; Higher education; Data mining; Clustering Algorithms; K-Means; Risk Visualization System.

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

The growing implementation of governance models in higher education institutions has raised new challenges in accountability and performance. Although risk management and analysis in governance has received some attention in the literature, there is a lack of studies focusing on the use data-driven clustering algorithms. The current research focuses on the development and assessment of a web-based risk management information system that incorporates clustering principles to identify and map non-academic governance risks. The study followed the CRISP-DM framework and analyzed 678 risks from the Internal Audit Unit (SPI) of Politeknik Negeri Malang. The system employs K-Means clustering analysis to classify the risks into tiers based on performance indicators, budget, and risk severity. The system is equipped with data upload, preprocessing, logging, and cluster visualization modules. A comparative analysis of K-Means, DBSCAN, and Hierarchical Clustering showed that K-Means yields the best cluster quality with a Silhouette Score of 0.48, in comparison to DBSCAN (-0.705) and Hierarchical clustering (-0.395). The developed system generated five distinct clusters corresponding to risks of varying settlement priority viz. very high, high, medium, low, and very low. The functional and usability assessments of the system confirmed that it provides automated and actionable insights on risks in a user-centric manner. The study has demonstrated the clustering of governance risks in higher education using K-Means is feasible. The incorporation of predictive analytics and real-time data would best support the research in active risk avoidance.

INTRODUCTION

The growing dependence of higher educational institutions on digital technologies has considerably modified how academic activities, operations, and administration activities are performed. There are certainly some benefits on the system efficiency and integration of varied processes. However, the digital technologies also pose risks that can affect the accountability, sustainability, and overall performance of an institution. As information and governance systems become complex, the need for well-structured, evidence-based, and threat-oriented approaches to mitigate governance quality risks escalates. As integral organizational functions, risk management cycles through five fundamental activities: identification, analysis, evaluation, mitigation, and monitoring. In higher education, risk is primarily classified into five categories: strategic, operational, financial, compliance, and fraud (Khaw & Teoh, 2022). Each risk entails a different level of unique analysis and response mechanisms (Valles Coral et al., 2022). Adopting risk management not only preserves valuable assets in an institution, but also facilitates decision making, enhances institutional agility, fulfills legal obligations, and promotes sustainability.

Notably, however, implementing risk management practices within higher educational institutions particularly within non-academic governance areas is still underdeveloped. Institutions continue to use manual and experience-based evaluations which, given the scope and complexity of risks universities face, may prove inefficient (Chao et al., 2024). This gap points to an urgent need for objective, scalable systems capable of understanding and prioritizing risks through pattern recognition. Data-driven techniques risk-based prioritization may use clustering algorithms within data mining to hone the identification and risk classification process around controllable variables, including performance metrics, budgets, and risk exposure (Wang, 2022). Still, the use of such techniques in the governance of educational institutions remains largely untapped, particularly in non-academic operational risks which tend to be overlooked yet pose significant challenges to an institution's overall effectiveness (Zein et al., 2024).

In response to this deficiency, the current research intends to design and assess a web-based information system that utilizes clustering techniques to uncover and portray governance-related risks in non-academic campus activities (Rhomadhoan et al., 2024). The research applies a quantitative methodology following the CRISP-DM framework and utilizes a set of 678 risk records acquired from the Internal Audit Unit (SPI) of Politeknik Negeri Malang. The system employs the K-Means clustering algorithm to analyze and categorize the risk data and features other modules for data harvesting and automatic pre-processing, periodical logging, and graphical display of the risk clusters. The system also



evaluates the performance of the K-Means algorithm relative to the other clustering methods, such as DBSCAN and Hierarchical Clustering, in determining clustering quality as measured by the Silhouette Score (Sinaga & Yang, 2020) (Bushra & Yi, 2021).

This research intends to investigate how clustering algorithms assist in the identification and classification of risk within the governance of higher education institutions. It will (1) create a governance risk assessment data-driven interventional information system, (2) use clustering techniques to stratify non-academic risk profiles, and (3) assess and contrast the clustering algorithms to guide more effective risk management within the institution. This study attempts to answer the following research question (RQ):

RQ1. In what ways can clustering-based data mining techniques be employed for the identification and mapping of non-academic governance risks within higher education institutions?

RQ2. Of the K-Means, DBSCAN, and Hierarchical Clustering algorithms, which one will produce the optimal risk classification results based on the governance risk patterns and the Silhouette Score?

LITERATURE REVIEW

Risk management is now an important field of study because of the rapid rise of complex digital governance systems and increasing demands for transparency and accountability within and outside the institution (Putra et al., 2023). Although routine methods such as manual audits and qualitative assessments continue to be implemented, they tend to be subjective, less precise, and less effective in large-scale heterogeneous datasets (Nugraha et al., 2023). More recent literature describes the adoption of data-driven quantitative approaches aimed at advancing the consolidation of risk management within newly emerging data paradigms (Khaw & Teoh, 2022). Previous research literature advocates a number of standardised governance frameworks like COBIT, ISO/IEC 27002, and ITIL while the research within the frameworks has been limited to traditional methods and has not attempted to incorporate exploratory techniques such as data mining (Ulya et al., 2025). Furthermore, the information systems governance literature pertaining to risk assessment seems to be less developed and has not widely applied clustering techniques as a holistic approach to analysis (Engku Abdullah et al., 2024).

Clustering techniques, including K-Means, DBSCAN, and Hierarchical Clustering, have many successful applications outside of educational fields. Clustering techniques have been applied successfully to fraud detection credit risk evaluation, and cybersecurity (Jayaprakasam, 2024). They are particularly useful for identifying hidden insights in large, unlabelled datasets since they require no labeled data to identify patterns and group data. Yet there remains a lack of research on the application of clustering techniques to governance in higher education, particularly to identify and analyze non-academic risks such as budgetary misuse, procedural abuse, and fragile internal control systems (Zogara & Ningrum, 2025). Additionally, the lack of research in higher education on the comparative analysis of clustering techniques and their application remains a gap. Within clustering analysis, techniques for internal validation, including the Silhouette Score, Davies-Bouldin Index, and Dunn Index, are well-established for assessing the quality of clusters; however, the application of such well-known techniques to governance of educational institutions must be further advanced (Riza et al., 2025). There appears to be a gap in the integration of clustering techniques and information systems that may be used by internal audit or quality assurance departments to identify and depict risks visually (Bhavani et al., 2025).

The literature gap research on the use of clustering algorithms and non-academic governance risk mapping—presents a considerable opportunity for contribution. Much of the work tends to concentrate on strategic-level risk frameworks or the technical components of information systems, neglecting the operational governance aspects that pertain to the academic context. Hence, this study seeks to address this gap by integrating a methodological component that entails the comparative study of clustering algorithms with a practical component of creating a web-based risk information system. By integrating risk management, data mining, and governance in higher education, this study is designed to provide a scalable and contextually appropriate data-driven solution to enhance non-academic risk identification and visualization on campuses in a more effective manner.

METHOD

This study takes a data-focused approach to pinpoint and evaluate risks associated with the governance of non-academic activities on campus, utilizing clustering techniques as outlined in the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Fuseini & Missah, 2024). Consequently, the approach was constructed utilizing six phases of this framework: (1) in the first phase, business understanding, the governance risk problem pertaining to non-academic activities is specified; (2) in the next phase, data understanding, data is gathered from institutional records and risk indicators; (3) in data preparation, preprocessing and selection of features is conducted; (4) in the modeling step, the clustering algorithms are created and applied; (5) in evaluation, the model's performance is reviewed, and the quality of the clustering is measured; and (6) in deployment, the system is set up, and the visualization of the risks in patterns is described (Lin et al., 2024). Figure 1 shows the stages of the CRISP-DM model employed in this study.

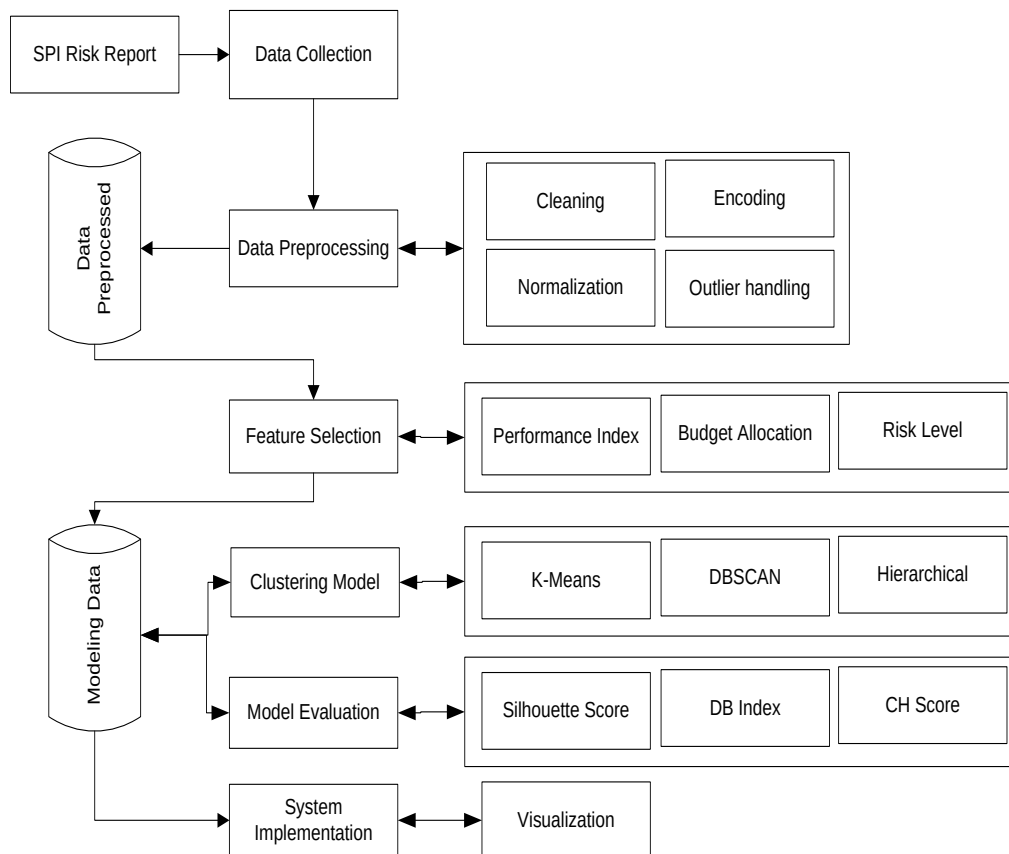


Fig. 1 Clustering-based Risk Classification Framework for SPI Risk Reports

Data Collection

In this study, data collection constitutes a foundational step which entails the acquisition of Campus Governance Risk data. The source of data was the internal SPI (Satuan Pengawasan Internal) Risk Reports. These reports preserve the historical account of non-academic governance risks and provide the primary data for further analysis. Recognizing the need for a systematic approach, I created stratified layers of data and organized them in a hierarchy. Historical risk reports were digitized into structured formats for computers to read, and stored in CSV files, Excel databases, or as file systems in relational databases(Airlangga, 2023). Some key parameters for analysis which included performance index, allocated budget, and risk grade were collated and checked for availability in all data points. Metadata was captured to log the data source and acquisition timestamp for each record in order to preserve data integrity. This effort was made in order to support the analytical framework with traceable and reproducible evidence.

The first step in this phase should involve designing a complete and representative dataset that captures the governance risk profile in higher educational institutions. This dataset will be referred to in the next stages pertaining to the data preprocessing steps, followed by feature selection, the building of a clustering model, and the evaluation of the model's performance. Since the performance of clustering algorithms is dependent on the quality of the input data, this phase is extremely important in shaping the precision and dependability of the identification of risk categories. The acquisition of high-quality data will positively impact the clarity of the clusters that will be produced and the tactical recommendations that will be provided to improve governance.

Data Preprocessing

Immediately after the acquisition of data, data preprocessing represents an important step within the data analytics pipeline. The main goal of this stage is to enhance the overall quality, integrity, and usability of a dataset to guarantee its compatibility to downstream machine learning, and more specifically, clustering analysis, tasks. For this study, a dataset based on the risk reports of the institution's Internal Audit Unit (Satuan Pengawasan Internal/SPI) required a series of systemic transformations in order to be analytical ready. The preprocessing workflow includes the following steps:

1. **Data Cleaning:** Duplicate records were eliminated, missing values were addressed, and inconsistencies or input errors were corrected. These procedures are essential to reduce noise and enhance the dataset's reliability.

2. **Categorical Encoding:** Non-numerical features were transformed into numerical representations to facilitate compatibility with machine learning algorithms. Encoding methods, such as one-hot encoding or label encoding, were applied based on the nature and distribution of the categorical variables.
3. **Normalization:** Numerical attributes were rescaled to a uniform range to avoid feature dominance during model training. Normalization also contributes to improved model convergence and overall algorithmic stability.
4. **Outlier Treatment:** Extreme values were identified and handled appropriately to prevent skewing of the clustering results. The treatment strategy—whether removal, transformation, or adjustment—was determined in accordance with the statistical context and influence of the outliers.

The data preprocessing steps followed ensure that the input data are analytically sound, validated statistically, and devoid of any misleading artifacts. Consequently, this stage plays a crucial role in efficient feature selection, model building, and subsequent evaluation of the model's performance. Furthermore, the rigorous pre-processing of data relates to the precision and the clustering results' interpretability, which are fundamental in making the relevant governance risk profile insights in higher education context. The number of data records utilized for the development of clustering models is presented in Table 1.

Table 1. Summary of Data Volume Before and After Preprocessing

Stage	Number of Records	Description
Raw Data Collection	678	Risk records collected from the Internal Audit Unit (SPI) of Polinema, containing unprocessed information such as risk descriptions, responsible units, and related attributes.
Post-Preprocessing	612	The dataset was cleaned by removing duplicate records, incomplete entries, and outliers. Only relevant and consistent attributes were retained for further analysis

Feature Selection

Feature selection constitutes another important phase of the data analytics pipeline. The objective of this phase is to select the most informative features that contribute the most to the general structure of the dataset. In this study, this was applied to the preprocessed governance risk data to optimize computational efficiency and improve the clustering results during the identification of meaningful risk groupings within the campus environment. The combination of theoretical and statistical domains guided the selection of features. Three features were retained because of their theoretical importance and analytical relevance to governance risk:

1. **Performance Index**, which measures the operational efficiency of organizational units in carrying out non-academic governance functions;
2. **Budget Allocation**, which records the distribution and use of budgetary funds for various programs and activities;
3. **Risk Level**, which describes the likely value and the probability of the described governance risks.

The features displayed enough variability and discriminative capacity to justify unsupervised learning, and their inclusion is consistent with the empirical risk assessment literature within the governance of higher education institutions. The goal of this phase was, primarily, to reduce the dimensions of the dataset, eliminate redundant or irrelevant variables, and improve the interpretability of the clustering results. The retention of only the most important dimensions allows the model to improve its capacity to identify hidden patterns within the risk data while extraneous dimensions provide the data noiser. In this sense, feature selection acts as a methodological filter and a conceptual simplification of the data. In the former sense, it contributes to the conceptual, the clarity, coherence, and analytical focus of the risk segmentation, while the latter enhances the precision of the identification and the understanding of the governance risk gaps of the institutions.

Development And Evaluation of Clustering Models

The focus of this phase is on the development and evaluation of unsupervised clustering models designed to identify hidden patterns within the governance risk data. The goal is to classify risk records into separate groups that share similar traits with respect to performance, budget distribution, and risk degree.

1. Risk Classification Using a Clustering Algorithm

The effectiveness of each clustering algorithm in grouping risk was assessed using three different clustering algorithms as follows:

Risk Classification Using K-Means Clustering

The objective of K-Means clustering is to minimize the intra-cluster variance to a certain degree by assigning each data point to the nearest cluster centroid. The minimization is mathematically expressed as:

$$\arg \min_c \sum_{i=1}^k \sum_{x \in C_i} ||x - \mu_i||^2 \quad (1)$$

where:



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

k	= number of clusters
C_i	= set of data point assigned to the i^{th} cluster
μ_i	= centroid (mean vector) of the i^{th} cluster
$ x - \mu_i ^2$	= squared Euclidean distance between a data point x and the cluster centroid μ_i

The algorithm will update the centroid positions iteratively until convergence is obtained.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN identifies clusters based on data point density and is particularly robust to noise and outliers. It requires two parameters:

ε (epsilon) = the adjacent radius

MinPts = the minimum number of point required to form a dense region

ε – (Neighborhood) for a data point x , its neighborhood is defined as :

$$N_\varepsilon(x) = \{y \in D \mid \text{dist}(x, y) \leq \varepsilon\} \quad (2)$$

Where D is a data set and dist is usually a Euclidean distance. A data point x is designated as a core point provided that it lies within a neighborhood radius ε (epsilon) containing at least a minimum number of points, denoted as MinPts. Formally, if the number of data points within the ε -neighborhood of x is greater than or equal to MinPts, then x qualifies as a core point. This condition enables that x occupies a region of ample density and acts as the probable point of a cluster

$$|N_\varepsilon(x)| \geq \text{MinPts} \quad (3)$$

In terms of clustering based on density, point y is said to be directly density-reachable from a core point x , if y is in the ε -neighborhood of x . This holds if x is a core point, meaning that direct density-reachability captures the density properties within the local area. Let x be a core point and y belong to the neighborhood ε , so $y \in N_\varepsilon(x)$.

A value y is said to be density-reachable from a value x if there exists a sequence of points $x_1, x_2, \dots, x_n$ where $x_1 = x$ and $x_n = y$ and every value in the sequence is directly density-reachable from the preceding value. This is a form of a transitive relationship and, through a continuity of a chain of dense neighborhoods, will allow for the fracturing of points into clusters. A point is described as noise or an outlier if it is not density-reachable from any other point in the dataset, and therefore, does not meet the classification of core points. These points sit in low density spatial data and will be excluded from all clusters. This is indicative of anomalous, or sparsely, distributed data points.

2. Clustering Model Validation

Three well-known internal validation metrics were used to measure the internal quality and structural validity of the clustering outcomes: the Silhouette Score (S), the Davies–Bouldin Index (DBI), and the Calinski Harabasz Index (CH Score). Each of these metrics captures a different aspect of the cohesion and separability of the clusters generated. The Silhouette Score measures how well a data point lies within a cluster and how well it does relative to the neighboring clusters. In equation form:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (4)$$

where:

a_i denotes the average intra-cluster distance for point i (i.e., the average distance to all other points within the same cluster),

b_i represents the minimum average distance between point i and all points in the nearest adjacent cluster

The Silhouette Score is between -1 and 1. The higher the score is, the better the clusters are, and the more distinct and well separated they are. The Davies–Bouldin Index computes how similar each cluster is to their most similar cluster by measuring the within-cluster compactness and between-cluster separations:

$$BI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{s_i + s_j}{d_{ij}} \right) \quad (5)$$

where:

s_i is the average distance between all points in cluster i and the corresponding cluster centroid μ_i

d_{ij} is the distance between centroids μ_i and μ_j

Smaller values of the DBI suggest more compact and well-separated clusters, thus illustrating the within-cluster compactness and between-cluster separations are improved.

The Calinski–Harabasz Index, or Variance Ratio Criterion, describes the ratio between the within-cluster dispersion and the between-cluster dispersion:



$$CH = \frac{tr(B_k)}{tr(W_k)} \times \frac{N-k}{k-1} \quad (6)$$

where:

$tr(B_k)$ and $tr(WB_k)$ denote the traces of the between-cluster and within-cluster dispersion matrices, respectively,
 N is the total number of observations,
 k is the number of clusters.

The greater the value of the Calinski score, the more separation between centroids of the clusters and the lower the variance within clusters.

System Implementation and Risk Visualization

After developing and evaluating the model, the last remaining task is to build a system prototype that combines the best performing clustering algorithm and encapsulates risk groupings in an easy to understand way. This aims to provide visual assistance to institutional decision-makers on the governance-related risks associated with the different non-academic units in the campus setting. The system consists of data loading, automated preprocessing, and real-time execution of clustering model, of which the chosen parameters are performance index, budget, and risk level. Once clustering is done, the system outputs a risk map or chart showing the distribution and severity of the risk groups that have been identified.

At this point, effective visualization is vital in comprehending intricate risk patterns. Stakeholders can understand the severity of risk areas, assess the risk trends of an organization, and understand the proximity of various organizational units in the multidimensional space of risk by using color-coded risk blocks, risk cluster centroids, and distribution plots. The strategic risk visualization results in enhanced situational awareness and advanced, rational, and data-supported risk management, incorporating the principles of rational risk control. In systems that classify risk, features aimed at temporal risk analysis and audit trails supporting trace analysis of risk classification over time enhance the value of predictive temporal visualization. Coupling oversight risk management with modeling and visualization creates an extensible and sustainable framework for the integration of advanced features onto the platform, such as predictive analytics and anomaly detection.

RESULT

Modelling

In Fig. 2(a), the different colors represent the various clusters created from the risk profiling dataset. Using the DBSCAN algorithm, the data was successfully clustered into four primary groups, also identifying some ungrouped outliers. This demonstrates the model's ability to distinguish various risk themes while retaining the outlying anomalies. Specifically, Cluster 1 was referred to as Performance Management Risk, and out of 158 records, it was characterized predominately by a low budget allocation alongside a considerable performance deviation. The risk level distribution within this cluster was also notable, as it comprised 28% Moderate, 49% High, and 23% Extreme, with issues including delays in reporting and KPIs not being met. Cluster 2 included 142 records that were marked as Operational Inefficiency Risk, with the main features being irregular spending of budgets and moderate performance challenges. Low, Moderate, and High Risk were observed at 21, 47, and 32 percent respectively, with examples being uncoordinated procurement and activities that overlapped. In relation to Cluster 3, it contained 167 records that were budget misallocation risk, identified as having high budget consumption and low performance results. High, Extreme, and Moderate Risk were distributed at 39, 45, and 16 percent respectively, typically, the manifestation of this scenario comprised ineffective large scale projects and underperforming vendors. In the meantime, Cluster 4, with 145 records, encompassed Low Risk Routine Operations that were largely consistent with both budgetary and performance metrics, having Very Low Risk, Low Risk and Moderate Risk predominantly at 42 percent, 38 percent, and 20 percent in that order of magnitude. These were exemplified by routine activities such as scheduled audits and preventive maintenance.

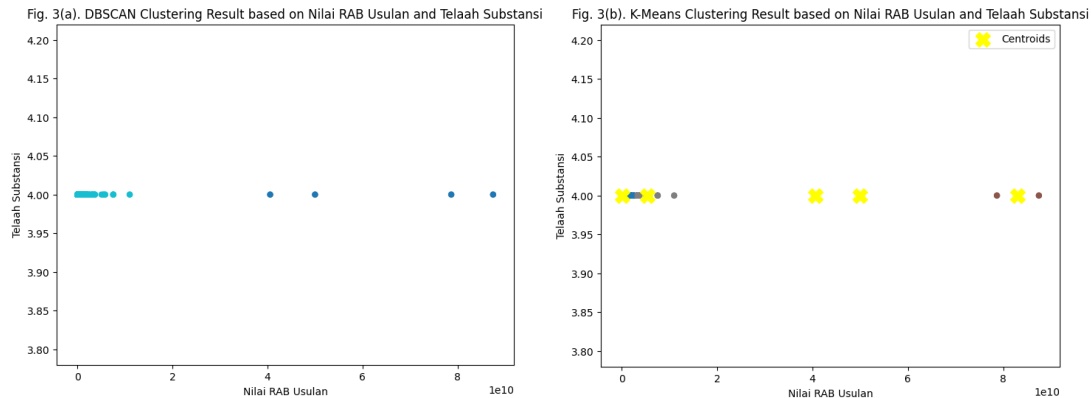


Fig. 2 Governance risk clustering results differentiated based on (a) DBSCAN; (b) K-Means

As presented in Fig. 2(b), the outcomes generated through the K-Means method are depicted. Each cluster is represented in the diagram using a unique color, while the yellow “X” markers signify the cluster centroids. The distribution of clusters indicated that, in contrast to DBSCAN, K-Means produced a greater degree of balanced clusters, which included 4279 instances in Cluster 0, 90 in Cluster 1, 1578 in Cluster 2, 305 in Cluster 3, and 5 in Cluster 4. Although K-Means produced compact clusters, unlike DBSCAN, it did not clearly state which clusters were outliers. In Fig. 3, a comparison of the DBI scores for the two methods is presented. The results demonstrated that K-Means produced a lower score, indicating that it produced higher quality clusters in comparison to DBSCAN. The disadvantage of K-Means, as stated earlier, is that it does not isolate outliers. This is a valuable trait in the context of governance risk, as the anomalies within a clustered group may represent significant, yet invisible, risk elements.

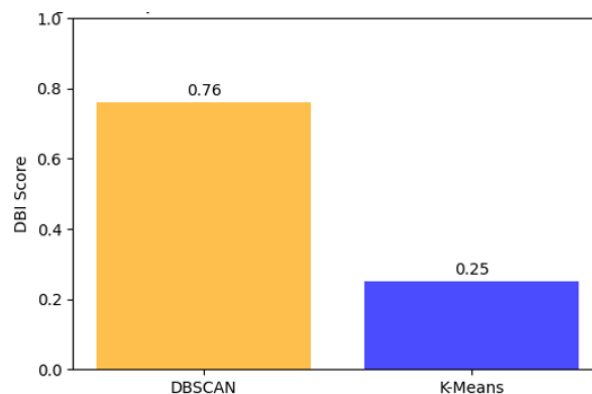


Fig. 3 Comparison of DBI Score between DBSCAN and K-Means

In general, the results suggest that while K-Means delivers the best performance in clustering quality and cohesion, DBSCAN makes an important contribution in the area of anomaly detection in the dataset. The extent to which these features enhance the Internal Audit Unit’s efforts in clustering governance risks and determining possible important outliers is remarkable. Table 2 provides a summary of the clustering results and the DBI scores for DBSCAN and K-Means.

Table 2. Summary of Data Clustering Result And DBI Scores

Algorithms	Number of Clusters	Outlier Detection	Largest Cluster Size	Smallest Cluster Size	DBI Score	Remarks
DBSCAN	3 + outliers	yes	167 (Cluster 3)	142 (Cluster 2)	0.76	Able to separate anomalies but produced less compact clusters
K-Means	5	no	4279 (Cluster 0)	5 (Cluster 4)	0.25	Produced more compact clusters, but failed to detect anomalies

Evaluation

Applying the K-Means clustering method to classify non-academic governance risks showed a high capability of creating organized and interpretable classifications. By relying on the three core attributes of Key Performance Indicators (KPI), budget, and risk, the clustering result effectively configured the data into 5 groups, each representing a different prioritized category. This classification is necessary for the Internal Audit Unit (SPI) of Politeknik Negeri



Malang, as it offers a feasible framework for decision support for monitoring risks and governance.

According to Table 2, Custer 0 includes 97 data points and is classified as "very high" priority, meaning the potential impact on goals for the institution are high and the associated risks need mitigation and attention as soon as possible. Custer 1 is classified as "high" priority, as the 126 data points it includes are still urgent, albeit a little less than those of Custer 0. Risks that Custer 2 includes with 160 data points are classified as "medium." This suggests that those risks can be handled with reasonable supervision and the adequate allocation of resources. Custer 3 contains the most data points, 178, and is classified as "low" priority which suggests that those risks are unlikely to affect the institution to a significant degree. Custer 4 includes 111 data points and is classified as "very low" priority, which captures the residual risks that are unlikely to affect the institution at all.

The relationship of hierarchy within these clusters illustrates K-Means capacity to deliver classification and actionable insights. The prioritization within the clusters aids decision-makers' resource distribution by directing attention to the most critical clusters while still monitoring the less critical clusters. The K-Means clustering results are displayed in Table 3 where governance risks are classified into five different priority levels. The table details the volume of data records, the KPI achievements, and the corresponding assigned priority levels.

Table 3. Results of K-Means Clustering

Cluster	Number of Data	KPI Achieved	Priority Level
0	97	97	Very High Priority
1	126	126	High Priority
2	160	160	Medium Priority
3	178	178	Low Priority
4	111	111	Very Low Priority

To validate the effectiveness of the clustering model, K-Means performance was compared to DBSCAN and Hierarchical Clustering using the Silhouette Score as the primary evaluation metric. Results in Table 3 indicate that K-Means scored 0.48, the highest among the methods. This score suggests that the clusters formed were of reasonable compactness and were relatively well separated. On the other hand, DBSCAN and Hierarchical Clustering obtained Silhouette Scores that were negative, -0.705 and -0.395, respectively. These negative Scores indicate that overlapping or poorly defined clusters were formed. Therefore, in these particular cases, the methods appear to have limitations in addressing the structure of the dataset. Table 4 shows the results of clustering evaluations based on the Silhouette Score for K-Means, DBSCAN, and Hierarchical Clustering, where K-Means shows the best performance. This correlates to the formation of more compact and well separated clusters.

Table 4. Clustering Evaluation Result

Algorithm	Silhouette Score
K-Means	0.48
DBSCAN	-0.705
Hierarchical	-0.395

The comparative analysis provides strong evidence that K-Means is more suitable to classify non-academic governance risk data when compared to DBSCAN and Hierarchical Clustering. This is mainly due to K-Means having the right balance of simple computations and effective cluster separation which is ideal for governance datasets where interpretability and stability are of the essence. In this case, K-Means has 0.48 as the most ideal Score. Figures 3 and 4 supplement this narrative by providing a visual representation of clustering distribution and a comparative analysis of algorithm performance, respectively. The stability of K-Means clusters is reinforced by visualization and illustrates, once more, K-Means dominance when compared to other clustering strategies.

To sum up the analysis, the K-Means technique provides valuable insights regarding actionable steps pertaining to institutional governance and the analytical approach demonstrates the discernment of actionable and defensible classifications of governance tiered risks. Moreover, the method aids the construction of the elemental and hierarchical order of risks. This arrangement is determinative of the information and priority order of the decision frameworks used by the SPI. This clarity reinforces institutional governance because resources can be allocated to the risks in a targeted manner and risks can be mitigated strategically. This is a table reference 5 which illustrates and explains the significance of relationships within the clustering model. The table demonstrates the span of governance related to the records within 4 overarching risk themes and the 4 disparate analytical clusters while providing insight into the governance risk profile.

Table 5. The Significance of the Relationships in the Model

Cluster ID	Number of Records	Risk Theme	Dominant Characteristics	Risk Level Distribution	Representative Examples
Cluster 1	158	Performance Management Risk	Low budget allocation with significant performance deviation	Moderate (28%), High (49%), Extreme (23%)	Delays in reporting; underachievement of KPIs
Cluster 2	142	Operational Inefficiency Risk	Irregular budget execution and moderate performance issues	Low (21%), Moderate (47%), High (32%)	Uncoordinated procurement; overlapping activities
Cluster 3	167	Budget Misallocation Risk	High budget usage with poor performance outcomes	High (39%), Extreme (45%), Moderate (16%)	Ineffective large-scale projects; failure in vendor performance
Cluster 4	145	Low-Risk Routine Operations	Stable in both budget allocation and performance indicators	Very Low (42%), Low (38%), Moderate (20%)	Scheduled audits; preventive maintenance
Total	612				

DISCUSSION

The comparative evaluation of DBSCAN and K-Means indicates a trade-off between both approaches: the former is more proficient in detecting anomalies while the latter is more compact and interpretable in its clustering. This difference is pivotal in governance risk profiling, where the need to detect potentially catastrophic outlier risks is as important as the systematic analysis of core risk patterns. DBSCAN showed marked strengths in risk outlier identification. As shown in Figure 3(a), DBSCAN achieved the specification of three centred clusters while also flagging atypical records which may be candidates for fraudulent reporting, unreported expenditures, or extreme underperformance. Four risk themes were identified and labelled: performance management, operational inefficiency, budget misallocation, and a low-risk routine operational cluster. The model did not obscure rare but qualitatively important anomalies. The DBSCAN algorithm did produce a relatively high DB index (DBI = 0.76) which points to lack of compactness in the clustered data. Furthermore, its performance on epsilon (ϵ) and MinPts parameters is weak, leading to unstable and unpredictable models which may hinder general governance and auditing applications. This is consistent with other work on the topic describing DBSCAN in outlier detection applications on unevenly distributed data. This is observed at the expense of cohesive clustering and parameter over-sensitivity.

K-Means clustering is preferred. It indicates more compact/interpret clusters (DBI = 0.25, Fig 3(b), Quality4), showing a clear five tiered risk hierarchy (Table 2), which helps with targeted audits and resource allocation for SPI. The K-Means method, though, is unable to explicitly describe intervene K-Means ‘anomaly detection’ is outlier absorption by centroids. The unequal distribution by clusters may also provoke concern; 4, 279 records in Cluster 0 compared to 5 in Cluster 4. K-Means also appears to describe Cluster 0 while Cluster 4’s ‘meaningful records’ may be indicative of more lightly clustered, yet relevant, data. These issues are documented biases of K-Means as described in section 4.3, K-Means Clustering Relies on spherical, equally sized data clusters and, a pre-decided number of clusters, K-Means has also documented biases of focusing on ‘densergroups’.

K-Means clustering is confirmed as the baseline clustering method for validation. The highest Silhouette Score of 0.48 proves a K-Means clustering model provides well differentiated models (aka clusters). In contrast to K-Means, DBSCAN and Hierarchical Clustering provide poorly differentiated clusters with -0.705 and -0.395 DB Scores respectively. These results are consistent with previously documented research, where K-Means provides expected results in controlled settings while DBSCAN is utilized in the overrun controlled settings. eg. fraud detection. Neither of the algorithms individually is sufficient for complete governance risk profiling. A hybrid or ensemble method, combining K-Means for building a structured hierarchy and incorporating DBSCAN for detecting anomalies, is advisable. This two-prong approach provides a balance between formal systematic risk profiling and the identification of outlier, rare, and potentially detrimental risk factors. In practical governance oversight, K-means will support governance risk strategic planning at the conceptual level, providing the consistency and clarity needed for long-term positioning, while DBSCAN will serve in the functional role of operational risk oversight and flagging troublesome cases. The dual approach will build institutional governance robustness. Table 6 contains a juxtaposition of DBSCAN with K-Means, capturing a balanced view of the merits and demerits, scoring on the axis of evaluation, and providing practical applications for each. The figure demonstrates the compromise where DBSCAN gives anomaly detection while K-Means provides relative compact clusters of risk in a clear hierarchy.

Table 6. Outlines the Major Features of Clustering Algorithms.

Aspect	DBSCAN	K-Means
Strength	Detects anomalies; flags rare high-impact risks	Compact, interpretable clusters; risk hierarchy
Limitations	Parameter-sensitive, less compact	Cannot detect anomalies; cluster size skewed
DBI Score	0.76 (less compact)	0.25 (more compact)
Silhouette Score	-0.705 (poor separation)	0.48 (well-separated clusters)
Risk Profiling	4 categories + outlier detection	5 hierarchical priority levels
Output		
Use Case	Diagnostic anomaly detection	Strategic risk stratification

Recent findings continue to highlight the importance of hybrid clustering and ensemble techniques, especially those with Explainable AI (XAI), for enhanced transparency and accountability in decision-making. This is especially true for governance auditing.

CONCLUSION

This study adds to the strategic governance risk profiling literature through a comparison of DBSCAN and K-Means clustering within the scope of non-academic risk classification in higher education. The findings indicated the potential of DBSCAN anomaly detection for diagnostic purposes illuminating the strategic blind spots of governance risk, albeit it clustered less compactly and was less insensitively to tuning of parameters. K-Means, in contrast, produced more compact and interpretable risk clusters, enabling the defining of a clear five-tiered risk classification to assist in aligning strategic decisions and resource allocation. The main contribution of the study is the first demonstration of the value of an integrated DBSCAN-K-Means hybrid approach as a more comprehensive solution for risk governance ‘auditing’ in the higher education institution. The combination of the two offers significant improvements to the accountability of decision-making and the active monitoring of risk. The limitations of the current study arise from the insensitivity of parameters in DBSCAN and the narrow scope of the dataset, which could be ameliorated in future studies through the use of larger datasets, the incorporation of ensemble clustering, and the application of Explainable AI for interpretability and reproducibility purposes.

ACKNOWLEDGMENT

This research was funded by the DIPA grant of Politeknik Negeri Malang (Polinema). The authors would like to acknowledge the Internal Audit Unit (SPI Polinema) for their valuable collaboration and support throughout the data acquisition and validation process

REFERENCES

- Airlangga, G. (2023). *Advanced Seismic Data Analysis : Comparative study of Machine Learning and Deep Learning for Data Prediction and Understanding*. 3(2), 456–465.
- Bhavani, T., Ramineni, M., Sravya, B., Sravanthi, B., & Ravikanth, K. (2025). Advanced Clustering Techniques: A Comprehensive Study. *Lloyd Business Review*, 104–113. <https://doi.org/10.56595/lbr.v4i1.54>
- Bushra, A. A., & Yi, G. (2021). Comparative Analysis Review of Pioneering DBSCAN and Successive Density-Based Clustering Algorithms. *IEEE Access*, 9, 87918–87935. <https://doi.org/10.1109/ACCESS.2021.3089036>
- Chao, Y., Fazidah Elias, N., Yahya, Y., & Jenal, R. (2024). *INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION journal homepage: www.joiv.org/index.php/joiv INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION Technologies on Intelligent Financial Risk Early Warning in Higher Education Institutions: A Systematic Re. 8*(September). www.joiv.org/index.php/joiv
- Engku Abdullah, E. M., Yakob, R., & B.A.M, H.-S. (2024). A Comprehensive Review of Enterprise Risk Management on Higher Education Institutions Performance. *Asia Proceedings of Social Sciences*, 12, 20–24. <https://doi.org/10.31580/rf6td074>
- Fuseini, I., & Missah, Y. M. (2024). A critical review of data mining in education on the levels and aspects of education. *Quality Education for All*, 1(2), 41–59. <https://doi.org/10.1108/QEA-01-2024-0006>
- Jayaprakasam, B. S. (2024). AUTOENCODER-BASED CLOUD FRAMEWORK FOR DIGITAL BANKING: A DEEP LEARNING APPROACH TO FRAUD DETECTION, RISK ANALYSIS, AND DATA SECURITY. *International Research Journal of Education and Technology*, 6, 51–61. <https://doi.org/10.70127/irjedt.vol.3.issue12.62>
- Khaw, T. Y., & Teoh, A. P. (2022). Risk management in higher education research: a systematic literature review. *Quality Assurance in Education*, 31(2), 296–312. <https://doi.org/10.1108/QAE-04-2022-0097>



- Lin, Lin, Zhou, Danhua, Wang, Jingying, & Wang, Yu. (2024). A Systematic Review of Big Data Driven Education Evaluation. *SAGE Open*, 14(2), 21582440241242180. <https://doi.org/10.1177/21582440241242180>
- Nugraha, Nugraha, Prasetyo, Yogi Tri, Sugiharti, Harpa, Lhutfi, Iqbal, Widyaningsih, Aristanti, Triantoro, Arvian, Ong, Ardvin Kester S, Young, Michael N, Persada, Satria Fadil, Kristamtomo Putra, Raden Aditya, & Nadlifatin, Reny. (2023). Quality Assurance in Higher Educational Institutions: Empirical Evidence in Indonesia. *Sage Open*, 13(4), 21582440231203060. <https://doi.org/10.1177/21582440231203060>
- Putra, P. T., Anggrawan, A., & Hairani, H. (2023). *Comparison of Machine Learning Methods for Classifying User Satisfaction Opinions of the PeduliLindungi Application*. 22(3), 431–440. <https://doi.org/10.30812/matrik.v22i3.2860>
- Rhomadhoan, H., Hayatie, M. N., & Pebriana, R. (2024). *Design and Development of a Web-Based Correspondence Management Information System at Politeknik Negeri Tanah Laut*. 4(2), 638–647.
- Riza, L. S., Novia, D., & Husaeni, A. (2025). *Machine Learning - Based Clustering for Program Learning Outcomes in Higher Education : A Systematic Review*. 5(1), 182–189.
- Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, 8, 80716–80727. <https://doi.org/10.1109/ACCESS.2020.2988796>
- Ulya, A., Karima, A., Sukiman, T. S. A., Zulfia, A., & Rahmawati, R. (2025). *Information Security Risk Analysis Using ISO 31000 : 2018 and ISO 27001 : 2022*. 5(2), 843–853.
- Valles Coral, M., Salazar-Ramírez, L., Injante, R., Hernández Torres, E., juarez-diaz, J., Navarro-Cabrera, J., Pinedo, L., & Vidaurre-Rojas, P. (2022). Density-Based Unsupervised Learning Algorithm to Categorize College Students into Dropout Risk Levels. *Data*, 7, 165. <https://doi.org/10.3390/data7110165>
- Wang, Z. (2022). Higher Education Management and Student Achievement Assessment Method Based on Clustering Algorithm. *Computational Intelligence and Neuroscience*, 2022, 1–10. <https://doi.org/10.1155/2022/4703975>
- Zein, A., Farizy, S., Suharyanto, E., Trisianto, C., & Marwati, F. (2024). Information Technology Governance Between Iso 38500, Risk It and Val It in Private University. *Journal of Information System, Technology and Engineering*, 2, 280–286. <https://doi.org/10.61487/jiste.v2i3.82>
- Zogara, L. U., & Ningrum, L. (2025). a Comparative Review of Clustering and Classification Algorithms for Big Data Analytics. *Scientific Journal of Information System*, 3(1), 36–41. <https://doi.org/10.70429/sjis.v3i1.179>