

Identification of Governance Risks in a Campus Environment Using Clustering Algorithms: A Comparative Study Based on Data Mining

Sumiadji¹, Usman Nurhasan², Anugrah Nur Rahmanto³, and Zainal Abdul Haris⁴

^{1,4}*Department of Accounting., State Polytechnic of Malang, Jl. Soekarno - Hatta 9, Malang-Indonesia*

^{2,3}*Department of Information Technology., State Polytechnic of Malang, Jl. Soekarno - Hatta 9, Malang-Indonesia*

Corresponding Author: usmannurhasan@polinema.ac.id

Received Date: 24-03-2026

Revised Date: 13-05-2026

Accepted Date: 21-06-2006

Abstract

Governance risks in higher education remain a critical challenge that influences both performance and accountability. Despite increasing attention to risk analysis, the application of data-driven approaches—particularly clustering algorithms—for the identification of governance risk is still limited. This study proposes and evaluates a web-based information system integrating clustering algorithms to classify and visualize non-academic governance risks. Using the CRISP-DM framework, 678 risk records from the Internal Audit Unit (SPI) of Politeknik Negeri Malang were analyzed. The system used K-Means clustering on three attributes: performance indicators, budget allocation, and risk levels. Additional modules supported data upload, automated preprocessing, historical logging, and visualization. Comparative evaluation with DBSCAN and Hierarchical Clustering was conducted using Silhouette Scores. The results indicated that K-Means achieved the highest score (0.48), outperforming DBSCAN (-0.705) and Hierarchical Clustering (-0.395). K-Means produced five interpretable clusters that represent very high, high, medium, low, and very low risk priorities. Functional and usability tests confirmed the reliability and user-friendliness of the system in supporting automated governance risk assessment. The findings demonstrate the effectiveness of K-Means clustering in segmenting governance risks and providing actionable insights for audit teams. Future work should integrate predictive models and real-time data to strengthen proactive governance risk mitigation.

Keywords: Governance Risk, Higher education, Data mining, Clustering Algorithms, K-Means, Risk Visualization System

1 Introduction

The increasing reliance on digital systems in higher education institutions has significantly transformed the way academic, administrative, and operational activities are conducted. While these systems offer numerous benefits in terms of efficiency and process integration, they also introduce various forms of risk that can directly affect institutional performance, accountability, and sustainability. As the complexity of information systems grows, so too does the need for structured and evidence-based risk management approaches that can effectively mitigate threats to governance quality. Risk management is a fundamental organizational function that encompasses five key stages: identification, analysis, evaluation, mitigation, and monitoring. Within the higher education context, risks are typically categorized into strategic, operational, financial, compliance, and fraud-related risks [1], [2]. Each category varies in terms of its likelihood and impact, thereby necessitating differentiated approaches for analysis and response. The implementation of effective risk management practices in universities not only serves to protect critical institutional assets but also enhances decision-making, strengthens institutional resilience, ensures regulatory compliance, and supports long-term sustainability.

Despite its importance, the practical implementation of risk management strategies in higher education particularly in non-academic governance domains remains limited. Many institutions continue to rely on manual or experience-based assessments, which may be inadequate for handling the volume and complexity of risk-related data generated in contemporary university environments [3]. This gap highlights the urgent need for more objective and scalable approaches capable of identifying patterns and prioritizing risks using data-driven methods. In this regard, data mining techniques specifically clustering algorithms offer promising potential for improving the identification and classification of risks based on key variables such as performance indicators, budget allocations, and risk levels [4], [5]. However, the application of such methods within the governance structures of academic institutions is still underexplored, especially concerning non-academic operational risks that are often underestimated but may carry significant consequences for institutional effectiveness [6].

To address this gap, the present study aims to develop and evaluate a web-based information system that integrates clustering algorithms to identify and visualize governance-related risks in non-academic campus activities. The study utilizes a quantitative approach based on the CRISP-DM framework and analyzes a dataset comprising 678 risk records obtained from the Internal Audit Unit (SPI) of Politeknik Negeri Malang. The system incorporates the K-Means clustering algorithm to segment risk data and includes additional modules for data upload, automatic preprocessing, historical logging, and visual presentation of risk clusters. Furthermore, the performance of the K-Means algorithm is compared with DBSCAN and Hierarchical Clustering using Silhouette Score metrics to determine clustering quality [7], [8].

This study seeks to explore the effectiveness of clustering algorithms in supporting risk identification and classification in the governance processes of higher education institutions. Specifically, it aims to (1) design a data-driven information system to analyze governance risk data, (2) segment non-academic risk profiles using clustering techniques, and (3) evaluate and compare the clustering performance of different algorithms to inform better decision-making in institutional risk management. The research question (RQ) in this study is as follows:

RQ1: How can clustering-based data mining techniques be effectively applied to identify and map non-academic governance risks within higher education institutions?

RQ2: Among K-Means, DBSCAN, and Hierarchical Clustering, which algorithm demonstrates the best performance in classifying governance risk patterns based on Silhouette Score evaluation?

2 Method

Risk management in higher education institutions has become an increasingly important area of study, driven by the growing complexity of digital governance systems and rising demands for institutional transparency and accountability [9]. Although conventional approaches such as manual audits and qualitative assessments remain widely practiced, their effectiveness in managing large-scale, heterogeneous datasets is often limited, particularly in terms of objectivity and analytical precision [10]. To address these limitations, recent studies have increasingly adopted data-driven quantitative approaches to strengthen risk management practices [11]. Earlier works have highlighted the importance of standardized frameworks such as COBIT, ISO/IEC 27002, and ITIL in IT governance. However, these approaches remain constrained within conventional framework-based evaluations and have not fully extended to exploratory data-driven methods such as data mining. Moreover, research on risk assessment in information systems governance often exhibits limited scope and has not extensively explored clustering algorithms as a comprehensive analytical alternative [12], [13], [14].

Clustering algorithms such as K-Means, DBSCAN, and Hierarchical Clustering have demonstrated promising applications in various non-educational domains. For example, these techniques have been successfully employed in fraud detection [15], [16], credit risk management [17], [18], and cybersecurity [19], [20]. Their ability to group data based on similarity without requiring pre-labeled datasets makes them particularly effective for uncovering hidden patterns in large data collections. Nevertheless, the use of clustering algorithms within the context of governance in higher education institutions particularly in mapping non-academic risks such as budget misuse, procedural violations, and weaknesses in internal controls remains underexplored [21].

Furthermore, previous studies have rarely conducted comparative evaluations of clustering algorithm performance in higher education settings. Although internal validation metrics such as the Silhouette Score, Davies-Bouldin Index, and Dunn Index are well-established for assessing clustering quality, these tools have yet to be widely applied to institutional governance contexts. Another limitation is the lack of integration between clustering approaches and web-based information systems that can be readily utilized by internal audit units or quality assurance divisions for risk identification and visualization [22].

This gap in the literature particularly the lack of research linking clustering algorithms with non-academic governance risk mapping signals a significant opportunity for contribution. Most existing work tends to focus on strategic-level risk frameworks or technical aspects of information systems, overlooking the operational governance needs specific to academic institutions. Therefore, the present study aims to bridge this gap by combining a methodological approach through the comparative evaluation of clustering algorithms with a practical application via the development of a web-based risk information system. By synthesizing insights from risk management theory, data mining practices, and higher education governance, this study seeks to offer a scalable and contextually relevant data-driven approach to improve the effectiveness of non-academic risk identification and visualization in campus environments.

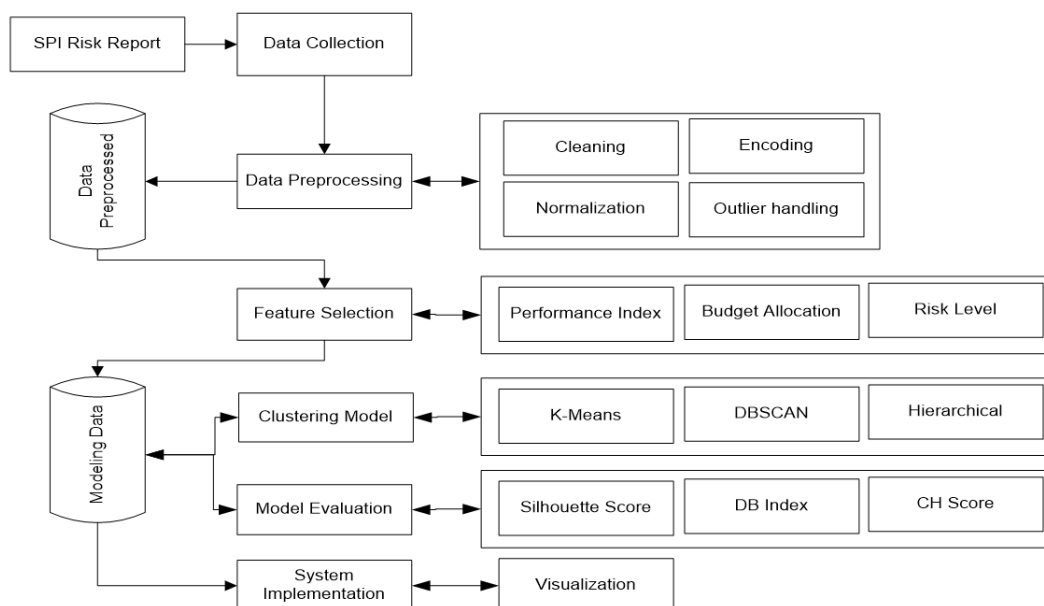


Figure 1: Clustering-based Risk Classification Framework for SPI Risk Reports

This study adopts a data-driven approach to identify and analyze governance risks in non-academic campus activities by applying clustering algorithms within the framework of the Cross-Industry Standard Process for Data Mining (CRISP-DM) [23], [24]. Following this framework, the methodology is organized into six stages: (1) business understanding, in which governance risk issues in non-academic activities are defined; (2) data understanding, including data collection from institutional records and risk indicators; (3) data preparation through preprocessing and feature selection; (4) modeling, which involves the development and application of clustering algorithms; (5) evaluation, where model performance and clustering quality are assessed; and (6) deployment, which covers the system implementation and visualization of identified risk patterns. The stages of the CRISP-DM model as implemented in this study are illustrated in Figure 1.

2.1 Data Collection

The data collection phase constitutes the foundational step of this study, concentrating on the acquisition of governance risk records within the campus environment. The data were specifically obtained from official risk reports issued by the institution's Internal Audit Unit (Satuan Pengawasan Internal/SPI). These reports contain historical information pertaining to non-academic governance risks and serve as the empirical basis for subsequent analytical procedures.

A structured approach was adopted to systematically extract and organize the relevant data. Historical risk records were compiled and converted into standardized, machine-readable formats, such as CSV files, Excel spreadsheets, or relational databases. Essential attributes required for robust analysis namely, performance index, budget allocation, and risk level were identified and verified to be consistently available across all data entries [24], [25]. To preserve data integrity and ensure analytical validity, metadata such as data source identifiers and timestamps of acquisition were recorded. This step was critical for ensuring the traceability

and reproducibility of the entire data-driven process.

The main objective of this phase was to develop a comprehensive and representative dataset that accurately reflects the governance risk landscape in higher education institutions. The resulting dataset serves as the input for subsequent stages, including data preprocessing, feature selection, clustering model development, and performance evaluation. Given that the effectiveness of clustering algorithms is highly contingent upon the quality of the input data, this phase plays a pivotal role in determining the accuracy and reliability of risk group identification. High-quality data acquisition significantly influences the interpretability of the resulting clusters and the strategic insights generated for enhancing governance practices.

2.2 Data Preprocessing

Data preprocessing represents a crucial phase within the data analytics pipeline, performed immediately after data acquisition. The principal aim of this stage is to improve the overall quality, consistency, and usability of the dataset, thereby ensuring its suitability for downstream machine learning tasks, particularly clustering analysis [26], [27]. In this study, the dataset derived from the risk reports of the institution’s Internal Audit Unit (Satuan Pengawasan Internal/SPI) was subjected to a series of systematic transformations to ensure analytical readiness. The preprocessing workflow consists of the following key operations:

- a. **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.
- b. **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.
- c. **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.
- d. **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 systematic execution of these preprocessing steps ensures that the input data are analytically robust, statistically valid, and free from distortive artifacts. As such, this phase is instrumental in facilitating effective feature selection, model construction, and performance evaluation. Moreover, rigorous data preprocessing contributes directly to the accuracy and interpretability of clustering outcomes, which are critical for deriving meaningful insights into governance risk profiles within the higher education context. Table 1 presents the number of data records used for clustering model development.

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.

2.3 Feature Selection

Feature selection represents a pivotal stage in the data analytics pipeline, aimed at isolating the most informative attributes that contribute significantly to the underlying structure of the dataset. Within the context of this study, the process was applied to preprocessed governance risk data to improve the computational

efficiency and clustering performance in identifying meaningful risk groupings across the campus environment [28]. The selection of features was driven by a combination of domain-specific reasoning and statistical assessment. Three core features were retained based on their theoretical relevance to governance risk and their analytical utility:

- a. **Data Cleaning:** Performance Index, which reflects the operational effectiveness of organizational units in executing non-academic governance functions;
- b. **Categorical Encoding:** Budget Allocation, which captures the distribution and utilization of financial resources across programs and activities;
- c. **Normalization:** c. Risk Level, which denotes the estimated severity and likelihood of identified governance risks.

These features exhibited sufficient variability and discriminative power to support unsupervised learning, and their inclusion aligns with established risk assessment frameworks in higher education governance. The primary goal of this phase was to reduce dimensionality, eliminate redundant or irrelevant variables, and enhance the interpretability of clustering outcomes. By retaining only the most salient attributes, the model becomes more robust and capable of uncovering latent structures within the risk data without the noise introduced by extraneous features. Thus, feature selection serves not only as a methodological filter but also as a conceptual refinement of the data representation. This step directly influences the clarity, coherence, and analytical value of the resulting risk segmentation, thereby contributing to a more precise understanding of governance vulnerabilities in institutional settings.

2.4 Development and Evaluation Of Clustering Models

This stage focuses on the development and evaluation of unsupervised clustering models to uncover meaningful patterns within the governance risk dataset. The objective is to segment risk records into distinct groups that exhibit similar characteristics in terms of performance, budget allocation, and risk level [29]. Three clustering algorithms were employed to compare their effectiveness in risk grouping:

2.4.1 K-Means Clustering

K-Means is a partition-based clustering algorithm that aims to minimize the intra-cluster variance by assigning each data point to the nearest cluster centroid. The objective function is defined as:

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

where:

- 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 iteratively updates the centroid positions until convergence is reached.

2.4.2 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 ensures that x resides in a region of sufficiently high density, thereby serving as a potential anchor for cluster formation.

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

In the context of density-based clustering, a point y is defined as directly density-reachable from a core point x if y resides within the ε -neighborhood of x . This relationship holds only when x satisfies the core point criterion, thereby ensuring that direct density-reachability reflects local density characteristics. Point x is a core point and point y belongs to the ε -neighborhood of x , denoted as $y \in N_\varepsilon(x)$.

A point y is said to be density-reachable from a point x if there exists a sequence of points $x_1, x_2, \dots, x_n$ such that $x_1 = x$ and $x_n = y$, where each point in the sequence is directly density-reachable from its predecessor. This transitive relationship allows the identification of clusters by connecting points through a continuous chain of dense neighborhoods, even if the individual points are not directly reachable from the original core point. A point is considered a noise point, or outlier, if it is not density-reachable from any other point in the dataset and does not satisfy the criteria to be classified as a core point. Such points reside in regions of low data density and are excluded from all clusters, reflecting anomalous or sparsely distributed observations within the dataset.

2.5 Clustering Model Evaluation

To assess the internal quality and structural validity of the clustering results, three widely recognized internal validation metrics were employed: the Silhouette Score (S), the Davies–Bouldin Index (DBI), and the Calinski Harabasz Index (CH Score). Each metric offers a distinct perspective on the cohesiveness and separability of the formed clusters. The Silhouette Score evaluates how well each data point fits within its assigned cluster in comparison to neighboring clusters. It is formally defined as:

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

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 ranges from -1 to 1 , with higher values indicating more distinct and well-separated clusters. The Davies–Bouldin Index quantifies the average similarity between each cluster and its most similar cluster, taking into account both intra-cluster dispersion and inter-cluster separation:

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

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

Lower DBI values reflect more distinct clustering structures, indicating better overall cluster separation and compactness. The Calinski–Harabasz Index, also referred to as the Variance Ratio Criterion, measures the ratio of between-cluster dispersion to within-cluster dispersion:

$$CH = \frac{tr(B_k)}{tr(W_k)} \times \frac{N - k}{k - 1} \quad (4)$$

where:

- $tr(B_k)$ and $tr(W_k)$ 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.

A higher CH Score indicates well-defined clusters with greater separation between group centroids and low intra-cluster variance.

2.6 System Implementation and Risk Visualization

Following the model development and evaluation stage, the final step involves the implementation of a system prototype that integrates the optimal clustering algorithm and presents risk groupings in an interpretable format. The objective is to support institutional decision-makers by providing a visual representation of governance-related risks across various non-academic units within the campus environment [30]. The system includes modules for data loading, automated preprocessing, and real-time execution of the clustering model using preselected features namely, performance index, budget allocation, and risk level. Once clustered, the system generates a risk map or chart that visualizes the distribution and intensity of identified risk groups.

Visualization plays a pivotal role in this stage, as it facilitates intuitive understanding of complex risk patterns. Through the use of techniques such as color-coded clusters, cluster centroids, and distribution plots, stakeholders can quickly identify high-risk zones, monitor risk trends over time, and assess the relative positioning of organizational units within the multidimensional risk space. This visualization output not only enhances situational awareness but also enables a more informed, data-driven risk management strategy. In addition, the system may log historical data to support temporal analysis and audit trail documentation, ensuring traceability of risk classification over multiple periods. The integration of modelling and visualization into a single system provides a scalable and extensible platform for governance risk monitoring, laying the groundwork for further development such as predictive analytics or anomaly detection in future iterations.

3 Result and Discussion

3.1 Modelling

In Figure. 2(a), each colour represents a specific cluster formed from the risk profiling dataset. The DBSCAN algorithm successfully grouped data into four major clusters while also detecting several outliers that were not assigned to any cluster. This indicates that the algorithm was able to differentiate risk themes while preserving anomalies that deviate from general patterns. Specifically, Cluster 1 consisted of 158 records and was characterized as Performance Management Risk, where the dominant features were low budget allocation accompanied by significant performance deviation. The distribution of risk levels within this cluster showed Moderate (28%), High (49%), and Extreme (23%), with representative issues such as delays in reporting and underachievement of KPIs.

Cluster 2 comprised 142 records, labelled as Operational Inefficiency Risk, with irregular budget execution and moderate performance issues as the dominant characteristics. The risk distribution was Low (21%), Moderate (47%), and High (32%), represented by cases such as uncoordinated procurement and overlapping activities. Cluster 3 contained 167 records that were categorized as Budget Misallocation Risk. This cluster was identified by high budget utilization but poor performance outcomes, with a distribution of High (39%), Extreme (45%), and Moderate (16%), typically reflected in ineffective large-scale projects and failures in vendor performance. Meanwhile, Cluster 4, with 145 records, represented Low-Risk Routine Operations that were relatively stable in both budget allocation and performance indicators. The risk levels in this cluster were dominated by Very Low (42%), Low (38%), and Moderate (20%), exemplified by routine activities such as scheduled audits and preventive maintenance.

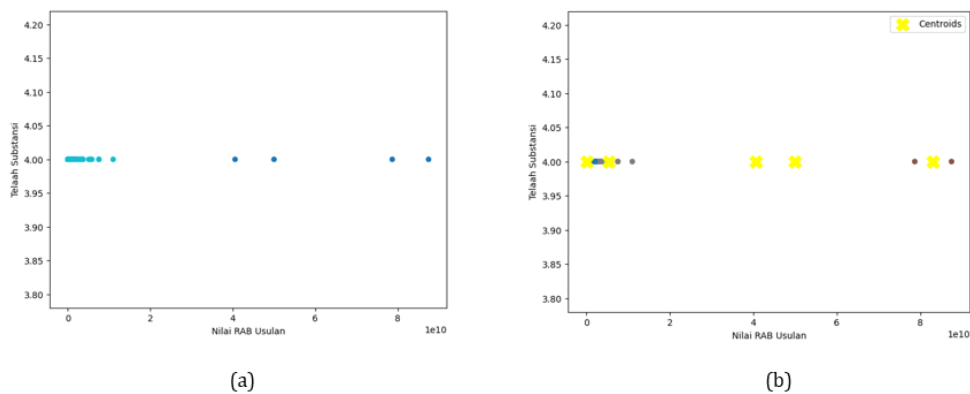


Figure 2: Governance risk clustering results differentiated based on (a) DBSCAN; (b) K-Means

In Figure. 2(b), the results of the K-Means algorithm are illustrated. Each colour represents a cluster, and the yellow ‘X’ marker indicates the centroid of each cluster. The clustering distribution showed that K-Means produced a higher number of balanced groups compared to DBSCAN, with 4279 records in Cluster 0, 90 in Cluster 1, 1578 in Cluster 2, 305 in Cluster 3, and 5 in Cluster 4. While K-Means yielded compact clusters, it did not explicitly identify outliers, in contrast to DBSCAN. Figure. 3 provides a comparative analysis of the Davies–Bouldin Index (DBI) scores between the two algorithms. The results revealed that the K-Means algorithm achieved a lower DBI score compared to DBSCAN, indicating superior clustering quality. However, DBSCAN offered an advantage by effectively isolating outliers, which is crucial in governance risk profiling where anomalous cases may represent hidden vulnerabilities.

Table 2: Summary of Data Clustering Result and DBI Scores

Algorithm	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

3.2 Evaluation

The application of the K-Means clustering method to classify non-academic governance risks demonstrated a significant ability to generate structured and interpretable groupings. By utilizing three critical attributes—Key Performance Indicators (KPI), budget allocation, and risk level—the clustering process effectively divided the dataset into five distinct clusters, each corresponding to a unique level of priority. This systematic classification is essential for the Internal Audit Unit (SPI) of Politeknik Negeri Malang, as it provides a practical decision-support framework for risk monitoring and governance. As presented in Table 2, Cluster 0 was categorized as very high priority with 97 data points. The risks within this cluster demand immediate attention and mitigation due to their potential high impact on institutional objectives. Cluster 1, comprising 126 data points, was classified as high priority, representing issues that, although less urgent than those in Cluster 0, still require strategic intervention. Cluster 2, with 160 data points, was designated as medium priority, suggesting risks that can be managed with moderate oversight and resource allocation. Cluster 3 contained the largest number of data points, 178, and was labeled low priority, which implies limited implications for institutional performance. Finally, Cluster 4 consisted of 111 data points, categorized as very low priority and representing residual risks with minor institutional consequences. The hierarchical nature of these clusters demonstrates the capability of K-Means to provide not only classification but also actionable insights. The priority stratification ensures that decision-makers can allocate resources efficiently, focusing on clusters with the highest urgency while maintaining oversight over less critical risks. Table 3 presents the K-Means clustering results, which classify governance risks into five distinct priority levels. The table shows the number of data records, corresponding KPI achievements, and their assigned priority categories.

Table 3: K-Means Clustering Result

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 ensure the robustness of the clustering model, the performance of K-Means was benchmarked against DBSCAN and Hierarchical Clustering using the Silhouette Score as the primary evaluation metric. The results,

presented in Table 3, revealed that K-Means achieved a score of 0.48, the highest among the tested methods. This value indicates that the clusters produced were relatively compact and well separated. In contrast, DBSCAN and Hierarchical Clustering yielded negative Silhouette Scores of -0.705 and -0.395, respectively. These negative values suggest overlapping or poorly defined clusters, highlighting the limitations of these methods in handling the structure of the dataset. Table 4 presents the clustering evaluation results based on the Silhouette Score, comparing the performance of K-Means, DBSCAN, and Hierarchical Clustering. The results indicate that K-Means achieved the highest score, reflecting 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 for classifying non-academic governance risk data compared to DBSCAN and Hierarchical Clustering. The superiority of K-Means is attributed to its ability to balance computational simplicity with effective cluster separation, making it particularly appropriate for governance datasets where interpretability and stability are critical.

Figures 3 and 4 complement these findings by providing a visual overview of the clustering distribution and a comparative analysis of algorithm performance, respectively. The visualization confirms the stable structure of K-Means clusters and further emphasizes its effectiveness over alternative clustering techniques. In conclusion, the evaluation underscores the robustness of the K-Means approach in delivering meaningful and actionable insights for institutional governance. The results highlight not only the ability of the method to classify risks into interpretable categories but also its practical utility in establishing a clear priority hierarchy. This hierarchy directly informs SPI's decision-making process, enabling evidence-based allocation of resources, targeted risk mitigation strategies, and enhanced institutional governance outcomes. Table 5 presents the significance of the relationships in the clustering model, detailing the distribution of records across four risk themes. Each cluster is characterized by its dominant attributes, associated risk level distribution, and representative examples, providing deeper insights into governance risk profiling.

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				

The comparative assessment of DBSCAN and K-Means reveals a trade-off: the former excels at anomaly detection, while the latter offers more compact and interpretable clustering. This distinction is critical in governance risk profiling, where identifying rare but impactful outliers is as important as understanding risk patterns systematically. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) demonstrated pronounced strength in isolating outliers. As shown in Figure 3(a), it grouped data into three coherent

clusters while flagging atypical records often suggestive of fraudulent reporting, unreported expenditures, or extreme underperformance. It categorized the data into four risk themes—performance management, operational inefficiency, budget misallocation, and low-risk routine operations—without masking rare but significant anomalies. However, DBSCAN yielded a relatively high Davies–Bouldin Index ($DBI = 0.76$), indicating less compact clusters, and its reliance on epsilon (ϵ) and MinPts parameters often resulted in unstable outcomes under suboptimal tuning. Its sensitivity diminishes its reproducibility in governance auditing applications. These observations align with extant literature highlighting DBSCAN’s effectiveness for outlier detection in irregularly distributed datasets, albeit at the cost of cluster cohesion and parameter sensitivity.

Conversely, K-Means produced more compact and interpretable clusters ($DBI = 0.25$, Figure 3(b)), yielding a clear five-tiered risk hierarchy (Table 2). This facilitated resource allocation and targeted audit strategies for SPI. Nonetheless, K-Means lacks explicit anomaly detection: outliers are absorbed into the nearest centroid, potentially obscuring hidden vulnerabilities. The resulting clusters were markedly imbalanced—Cluster 0 included 4,279 records compared to only 5 in Cluster 4 reflecting its bias toward denser groups and underrepresentation of smaller, yet meaningful, subsets. This is consistent with known limitations of K-Means, particularly its assumption of spherical, uniformly sized clusters and dependency on the predetermined number of clusters.

Model validation reinforced these findings: K-Means achieved the best Silhouette Score (0.48), indicating well-separated clusters; DBSCAN (-0.705) and Hierarchical Clustering (-0.395) both scored negatively, suggesting cluster overlap and weak separation. Such outcomes reflect prior observations that K-Means is favorable for structured, compact clustering in well-behaved datasets, while DBSCAN remains preferable for contexts rich in anomalies, such as fraud detection. Taken together, neither algorithm alone suffices for comprehensive governance risk profiling. A hybrid or ensemble approach—employing K-Means for structured hierarchy and DBSCAN for anomaly detection—is recommended. This dual framework ensures both systematic risk classification and sensitivity to rare, high-impact deviations. In real-world governance monitoring, K-Means can underpin long-term strategic planning, offering clarity and consistency. DBSCAN then functions operationally to flag cases for further scrutiny. This combined methodology enhances resilience and efficiency in institutional governance. Table 5 presents a comparative summary of DBSCAN and K-Means, highlighting their strengths, limitations, evaluation scores, and practical use cases. The table illustrates the trade-off between DBSCAN’s ability to detect anomalies and K-Means’ capacity to produce compact clusters with a clear risk hierarchy.

Table 6: Comparative Summary 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 Output	4 categories + outlier detection	5 hierarchical priority levels
Use Case	Diagnostic anomaly detection	Strategic risk stratification

These results echo recent findings which underscore the potential of hybrid clustering and ensemble techniques in advancing governance auditing—particularly when paired with Explainable AI (XAI) for enhanced transparency and decision accountability.

4 Conclusion

This study contributes to governance risk profiling by comparing DBSCAN and K-Means clustering for non-academic risk classification in higher education. The results revealed that DBSCAN effectively detected anomalies—such as irregular reporting and extreme underperformance—providing diagnostic value for identifying hidden vulnerabilities, albeit with less compact clusters ($DBI = 0.76$) and sensitivity to parameter tuning. In contrast, K-Means produced more compact and interpretable clusters ($DBI = 0.25$, Silhouette

Score = 0.48), enabling the establishment of a clear five-tiered risk hierarchy to support resource allocation and strategic decision-making. The main contribution lies in demonstrating that a hybrid approach integrating DBSCAN's anomaly detection with K-Means' structured stratification offers a more comprehensive solution for institutional governance auditing. This dual framework strengthens risk monitoring and enhances decision accountability. Limitations include DBSCAN's parameter sensitivity and dataset scope, which future research should address by employing larger datasets, ensemble clustering, and Explainable AI to improve interpretability and reproducibility.

5 Acknowledgements

The authors would like to acknowledge the Research and Community Service Unit (P3M) of Politeknik Negeri Malang for providing financial support through the institutional DIPA funding scheme. The authors also extend their gratitude to the Internal Audit Unit (SPI) of Politeknik Negeri Malang for their valuable cooperation and provision of governance-related data, which greatly contributed to the success of this research.

References

- [1] S. C. Lopes, "A risk taxonomy approach for higher education institutions," pp. 66–76, 2025.
- [2] M. A. Valles-Coral *et al.*, "Density-based unsupervised learning algorithm to categorize college students into dropout risk levels," *Data*, vol. 7, no. 11, 2022.
- [3] Y. Chao, N. F. Elias, Y. Yahya, and R. Jenal, "Technologies on intelligent financial risk early warning in higher education institutions: A systematic review," *International Journal on Informatics Visualization*, vol. 8, 2024. [Online]. Available: <https://www.joiv.org/index.php/joiv>
- [4] A. F. Mohamed Nafuri, N. S. Sani, N. F. A. Zainudin, A. H. A. Rahman, and M. Aliff, "Clustering analysis for classifying student academic performance in higher education," *Applied Sciences*, vol. 12, no. 19, 2022.
- [5] Z. Wang, "Higher education management and student achievement," vol. 2022, 2022.
- [6] A. Zein, S. Farizy, E. Suharyanto, C. Tristianto, and F. Marwati, "Information technology governance between iso 38500, risk it and val it in private university," *Journal of Information System, Technology and Engineering*, vol. 2, no. 3, pp. 280–286, 2024.
- [7] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80 716–80 727, 2020.
- [8] A. A. Bushra and G. Yi, "Comparative analysis review of pioneering dbscan and successive density-based clustering algorithms," *IEEE Access*, vol. 9, pp. 87 918–87 935, 2021.
- [9] V. S. Fatmawaty, I. Riadi, and H. Herman, "Higher education institution clustering based on key performance indicators using quartile binning method," *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 24, no. 1, pp. 141–154, 2024.
- [10] N. Nugraha *et al.*, "Quality assurance in higher educational institutions: Empirical evidence in indonesia," *SAGE Open*, vol. 13, no. 4, pp. 1–12, 2023.
- [11] T. Y. Khaw and A. P. Teoh, "Risk management in higher education research: a systematic literature review," *Quality Assurance in Education*, vol. 31, no. 2, pp. 296–312, aug 2022.
- [12] E. M. E. Abdullah, R. Yakob, and M. H. S. B. Abdullah, "A comprehensive review of enterprise risk management on higher education institutions performance," 2024.
- [13] E. M. Engku Abdullah, R. Yakob, and M. H. S. B. Abdullah, "A comprehensive review of information technology in enterprise risk management at higher education institutions," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 63, no. 2, pp. 64–81, 2025.
- [14] A. A. S. Perera, A. K. Rahmat, A. Khatibi, and S. M. F. Azam, "Review of literature: Implementation of enterprise risk management into higher education," 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:226306330>
- [15] B. S. Jayaprakasam, "Autoencoder-based cloud framework for digital banking: A deep learning approach to fraud detection, risk analysis, and data security," *International Research Journal of Educational Technology*, vol. 6, no. 11, pp. 51–61, 2024.

- [16] S. Devaguptam, S. S. Gorti, T. L. Akshaya, and S. S. Kamath, "Automated health insurance processing framework with intelligent fraud detection, risk classification and premium prediction," *SN Computer Science*, vol. 5, no. 5, p. 450, 2024.
- [17] W. A. Addy, C. E. Ugochukwu, A. T. Oyewole, O. C. Ofodile, O. B. Adeoye, and C. C. Okoye, "Predictive analytics in credit risk management for banks: A comprehensive review," *GSC Advanced Research and Reviews*, vol. 18, no. 2, pp. 434–449, 2024.
- [18] J. Alvi, I. Arif, and K. Nizam, "Advancing financial resilience: a systematic review of default prediction models and future directions in credit risk management," *Heliyon*, vol. 10, no. 21, nov 2024.
- [19] F. R. Ododo and R. R. Sadiq, "A review of outlier detection techniques in cybersecurity: A machine learning perspective," *Journal of Science, Innovation and Technology Research*, vol. 7, no. 9, pp. 349–366, 2025.
- [20] J. Baugher and Y. Qu, "Create the taxonomy for unintentional insider threat via text mining and hierarchical clustering analysis," *European Journal of Electrical Engineering and Computer Science*, vol. 8, no. 2, pp. 36–49, 2024.
- [21] L. U. Zogara and L. Ningrum, "A comparative review of clustering and classification algorithms for big data analytics," *Scientific Journal of Information Systems*, vol. 3, no. 1, pp. 36–41, 2025.
- [22] T. Bhavani, R. Mahesh, B. Sravya, B. Sravanthi, and K. Ravikanth, "Advanced clustering techniques: A comprehensive study," *Lloyd Business Review*, vol. 1, no. 4, pp. 104–113, 2025.
- [23] L. Lin, D. Zhou, J. Wang, and Y. Wang, "A systematic review of big data driven education evaluation," *SAGE Open*, vol. 14, no. 2, p. 21582440241242180, apr 2024.
- [24] I. Fuseini and Y. M. Missah, "A critical review of data mining in education on the levels and aspects of education," *Quality Education for All*, vol. 1, no. 2, pp. 41–59, 2024.
- [25] T. H. Shin and S. H. Kim, "Utility analysis about log data anomaly detection based on federated learning," *Applied Sciences*, vol. 13, no. 7, 2023.
- [26] Y. Zhang, M. Safdar, J. Xie, J. Li, M. Sage, and Y. F. Zhao, "A systematic review on data of additive manufacturing for machine learning applications: the data quality, type, preprocessing, and management," *Journal of Intelligent Manufacturing*, vol. 34, no. 8, pp. 3305–3340, 2023.
- [27] W. Fang *et al.*, "Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence," *Science Advances*, vol. 9, no. 40, pp. 1–116, 2023.
- [28] O. Cernisevs, Y. Popova, and D. Cernisevs, "Business kpis based on compliance risk estimation," *Journal of Tourism and Services*, vol. 14, no. 27, pp. 222–248, 2023.
- [29] M. Suyal and S. Sharma, "A review on analysis of k-means clustering machine learning algorithm based on unsupervised learning," *Journal of Artificial Intelligence Systems*, vol. 6, no. 1, pp. 85–95, 2024.
- [30] T. I. Azonuche and J. O. Enyejo, "Adaptive risk management in agile projects using predictive analytics and real-time velocity data visualization dashboard," *International Journal of Innovative Science and Research Technology*, vol. 10, no. 4, pp. 2032–2047, 2025.