World Journal of Information Systems 3(4) 84-103
DOI: https://doi.org/10.17013/wjis.v3i4.74

Article

Imputation Sensitivity and Robustness of Student Clustering in Higher Education Data

Bojan Radišić [1], Ivan Dunđer [2,*] and Sanja Seljan [2]

[1] Faculty of Tourism and Rural Development in Požega, Josip Juraj Strossmayer University of Osijek, Vukovarska 17, 34000 Požega, Croatia

[2] Faculty of Humanities and Social Sciences, University of Zagreb, Department of Information and Communication Sciences, Ivana Lučića 3, 10000 Zagreb, Croatia

***** Correspondence: idundjer@ffzg.unizg.hr

(Article history: Received 12 September 2026; Received in revised form 10 October 2026; Accepted 20 November 2026; Available online 15 December 2026)

Abstract: This paper investigates how different approaches to handling missing data shape the outcomes of unsupervised clustering in educational datasets. Three commonly used imputation algorithms – mean substitution, Random Forest and Multiple Imputation with Chain Equations (MICE) – were applied to student records from a Croatian higher education institution. The analysis included 461 students, with 46 (9.98%) missing values for study duration. Each dataset was analyzed individually using three clustering algorithms: K-means, Ward's hierarchical clustering and Gaussian mixture model (GMM). The three imputation methods and three clustering algorithms produced nine algorithm-dataset combinations. Solutions containing between two and ten clusters were evaluated using the silhouette coefficient. It was also examined whether the different datasets produced numerically and substantively similar student clusters. The results show that different imputation choices have a pronounced impact on Ward's hierarchical clustering, which selected ten clusters for MICE, six for Random Forest and four for Mean Replacement. On the other hand, K-means and GMM showed a high degree of robustness, consistently identifying the same number of clusters in all completed datasets – eight clusters for K-means and two for GMM. The analysis showed that when imputation was done with the mean value, all missing study durations were replaced with the same value. In contrast, Random Forest and Multiple Imputation by Chained Equations (MICE) estimated different study durations for each student, which preserved larger differences among the imputed values. Despite these methodological differences, several interpretable student profiles emerged repeatedly, including high-achieving students, long-studying older students, low-achieving early-leavers and students with moderate academic success. These results show that the effect of missing data processing on student grouping is algorithm-specific. Solutions based on centroids and

© 2026 ITMA. All rights reserved.

mixtures were comparatively robust in this data set, while Ward's hierarchical clustering produced different results since the imputation methods changed the calculated similarities between students. These results may benefit researchers working with unsupervised machine learning and incomplete educational data sets. Therefore, educational clustering analyses should examine whether different imputation methods produce the same number of clusters, and whether the resulting clusters represent similar groups of students.

Keywords: Missing data imputation; Imputation algorithms; Educational data mining; Clustering robustness; Machine learning algorithms; K-means; Ward's method; Gaussian mixture model

1. Introduction

Clustering plays a significant role in the processing of educational data, enabling researchers to discover related groups of students without predefined labels. Clustering techniques are widely used in educational analytics for student profiling, performance monitoring and decision support. Understanding the robustness of these techniques under incomplete data conditions is therefore also of practical importance. The K-means algorithm is often used for its simplicity and interpretability [1], though its effectiveness is often limited in heterogeneous educational datasets. However, models such as the Gaussian mixture model (GMM) have proven to be better for dealing with overlapping or irregularly shaped clusters, especially when variability among student subpopulations is significant [2]. Ward's hierarchical clustering has also been successfully applied to distinguish groups of students who exhibit similar behavioral patterns [3].

A common problem in datasets, including educational datasets, is the presence of missing values, which can arise from administrative errors, incomplete or partial records and technical problems. Research using predictive or statistical models is largely dependent on preprocessing of missing data, because an inadequate choice of imputation can lead to incorrect calculations and misinterpretations. For these reasons, analyzing the impact of different imputation techniques on clustering outcomes is crucial to the research methodology. Assumptions about missing data and the procedures used to process them can affect the results of statistical analyses. This issue is particularly important in clustering because there is no known outcome against which to check the effects of preprocessing.

Mean imputation assigns the same value to each incomplete case and can therefore reduce data variability [4]. The Random Forest method uses decision tree models to estimate missing values from relationships between observed variables, including nonlinear relationships and interactions [5]. MICE is an iterative procedure in which missing values are estimated using a series of conditional models [6]. Unlike mean imputation, Random Forest and MICE can provide different estimates for different students because they use information from other variables. Clustering algorithms use different criteria and assumptions when forming groups, and can therefore respond differently to changes in the

© 2026 ITMA. All rights reserved.

structure of the data [7]. Because imputation changes the calculated similarities and distances between students, it can affect both the number and composition of the resulting clusters.

This issue is particularly relevant in higher education, where administrative records can be used to identify students with different patterns of study progress and academic achievement. Therefore, robustness should be assessed by comparing a selected number of clusters, and examining the similarity of student profiles obtained from differently imputed data sets. The contribution of this research does not refer to proposing a new clustering technique, but rather to providing practical guidelines for educational data mining based on robustness analysis of selected algorithms.

This study compares three clustering algorithms in datasets that were imputed using mean substitution, Random Forest and Multiple Imputation by Chained Equations (MICE). Although there are several techniques for determining the number of clusters, in this paper only the silhouette coefficient [8] was used, which ensures objectivity in selecting the optimal number of clusters across all datasets and across all algorithms. The structure of the resulting clusters, their robustness and their interpretability in the context of study success were analyzed. The study makes several additional contributions.

First, it provides a controlled comparison in which the three clustering algorithms are applied to three versions of the same data set, differing only in the method used to impute missing study duration values. Second, silhouette-based cluster selection is complemented by comparisons of convergence behavior, cluster centroids and distributions of imputed values. Third, the study examines whether different imputation methods produce the same number of clusters and similar groups of students. The algorithm can select a different number of clusters while still identifying similar student groups, or it can select the same number of clusters even though the characteristics of those groups have changed.

This article extends the authors' previously published conference study [9] by providing a broader analysis of missing-data assumptions, distributional diagnostics, cluster-profile stability, practical implications and limitations. In this paper the following research questions (RQ1 and RQ2) were raised:

- *(RQ1) How do different imputation techniques (mean substitution, Random Forest and MICE) impact the determination of the optimal number of clusters in educational datasets?*
- *(RQ2) To what extent do the results of each clustering algorithm change when applied to datasets generated through different imputation algorithms?*

2. Literature review

Recent work in educational data mining increasingly relies on unsupervised machine learning to identify latent structures in student behavior and academic performance. Clustering algorithms such as K-means, Ward's hierarchical clustering and Gaussian Mixture Model (GMM) are frequently applied in this context since they are able to capture heterogeneity within educational datasets, and reveal patterns that are not apparent through descriptive analysis. Several studies emphasize the methodological relevance of comparing multiple clustering algorithms when analyzing student-related data. For example, one study highlights the use of the silhouette coefficient as a reliable criterion for selecting the optimal

© 2026 ITMA. All rights reserved.

number of clusters, and demonstrates that different algorithms can produce complementary insights [10].

This stresses the value of methodological triangulation, and directly impacts the design of the present study. A similar conclusion arises from recent analytical work outside the educational domain, particularly in predictive maintenance, where clustering algorithms are compared across multi-feature datasets. The findings show that K-means, GMM and hierarchical techniques often generate substantially different partitions, underscoring the importance of evaluating several algorithms rather than relying on a single model [11].

These results also confirm that the silhouette coefficient performs robustly when cluster geometry varies across algorithms. Clustering has long been used to study student performance and learning behaviors.

A study based on institutional information systems data [12] applies K-means to group students according to academic indicators, behavioral patterns and participation metrics, demonstrating that unsupervised machine learning can uncover profiles relevant for early-warning systems and personalized academic support.

Similarly, one study employs K-means to analyze LMS (Learning Management System) interaction data, showing that the algorithm effectively differentiates students with distinct engagement patterns, particularly in high-dimensional datasets [13].

Beyond K-means, several studies explore the advantages of more flexible models. Another research [2] employs GMM to detect structural patterns within computer science educational data, arguing that the probabilistic framework is well suited for representing heterogeneous learning behaviors, and identifying subpopulations that traditional clustering may overlook.

Ward's hierarchical clustering has also been applied to educational datasets; for instance, one research examined learning behaviors across two student cohorts, and identified consistent clusters of low-, medium- and high-progress learners, suggesting that hierarchical structuring can meaningfully capture behavioral differences [3].

Broader machine learning approaches also contribute to understanding student heterogeneity. A multi-factor framework proposed in [14], integrating behavioral, financial and sensor-based variables, illustrates the importance of modelling student populations across multiple dimensions – an idea that aligns with the motivation for applying clustering in this study.

Studies of campus-integrated MOOCs (Massive Open Online Courses) have used a system that combines Ward's algorithm and K-means, resulting in different motivational profiles of students across clusters [15], thus bringing interpretive value to clustering in higher education. The literature distinguishes three possible reasons for missing data. If data are missing completely at random (MCAR), their absence is not related to the available data on the student or to the actual value of the missing data. For example, data on the study duration may be missing due to a random administrative or technical error. If data are missing at random (MAR), their absence can be explained by other available student characteristics,

© 2026 ITMA. All rights reserved.

such as age at enrolment, number of ECTS credits achieved or grade point average (GPA). If data are not missing at random (MNAR), the probability of a value being missing is related to the value itself that is not recorded. For example, data on the study duration may be more likely to be missing for students who study for longer periods or who have dropped out [4].

Three methods of imputation of missing data have been considered. The simplest imputation method is mean because it uses a single equal value for all missing data, but then the variability and correlation between variables are reduced. Replacing with a single constant value can create an illusion of greater precision than it actually is. The Random Forest method is more complex because it does not take equal values but interprets the relationships between variables, and can capture nonlinear relationships and interactions in data of different types [5]. The third method, MICE, uses a series of conditional models, and iteratively estimates missing values, taking into account their correlation with other characteristics [6]. Also, these more advanced methods require verification of the obtained values because they can produce unrealistic values if there is no range of possible values.

The silhouette coefficient assesses the compactness and separation of clusters by comparing how well a given data item fits into an assigned cluster, and how close it is to the nearest other cluster [8]. Stability evaluation checks whether similar clustering solutions are obtained after changes in the data or analysis conditions [16]. Therefore, a high silhouette coefficient value does not in itself mean that the obtained solution will remain stable after changes in data preparation. When analyzing data with missing values, the stability of clustering may depend on the chosen imputation method, which may also affect the determination of the number of clusters [17]. In this study, robustness is therefore assessed by comparing the silhouette coefficient values for different numbers of clusters, the profiles of the resulting clusters and the convergence of individual algorithms.

Related work on the same educational dataset examined the effect of imputation methods on supervised classification performance [18].

3. Methodology

The dataset was created from the ISVU system (Cro. Informacijski sustav visokih učilišta, Eng. Information System of Higher Education Institutions) based on student records (2010–2018), using a standardized sequence of preprocessing and clustering steps, applied consistently across all datasets.

To assess the impact of missing data in the initial dataset, the authors created three datasets in parallel, using (i) mean substitution, (ii) MICE and (iii) Random Forest imputation. Each of these datasets is clustered using three algorithms: (i) K-means, (ii) Ward's hierarchical clustering and (iii) Gaussian mixture model (GMM). The number of clusters was selected using the silhouette coefficient only.

The individual cluster profiles were then grouped according to common features, and visualized to support the interpretation of the obtained results. From an optimization perspective, K-means and Gaussian mixture models can be formulated as unconstrained

© 2026 ITMA. All rights reserved.

optimization problems. No additional constraints are introduced in this study, as the focus is on algorithmic robustness rather than modelling constrained optimization.

The analytical procedure included data review, imputation of missing values for study duration, standardization of variables, application of three clustering algorithms and comparison of the obtained solutions. In all three obtained data sets, the same student records were used, and the same range of possible number of clusters was considered. Therefore, the differences in the results can be attributed to the joint action of the imputation method and the clustering algorithm, and not to differences in the analyzed sample.

3.1. Dataset

The experimental analysis is based on a relatively small set of actual data on student academic achievement. Limited sample sizes and incomplete records are common limitations in educational data mining. The data were extracted from ISVU, and included records of 461 students enrolled in the Professional Study of Trade at the Faculty of Tourism and Rural Development in Požega, Josip Juraj Strossmayer University of Osijek, Croatia.

The dataset included students enrolled in the study in the period from 2010 to 2018. The dataset covered 195 male (42%) and 266 female (58%) students, and 264 full-time (57%) and 197 part-time (43%) students. The original dataset contained 46 missing values in the data on study duration. The original dataset also included a set of thirteen input variables covering the academic status of the students, demographic data and educational achievements.

In this study, the authors analyzed three imputed datasets in parallel, in which missing values were replaced by values using: mean substitution, Multiple Imputation by Chained Equations (MICE) and Random Forest-based imputation [4–6]. In short, a total of 461 students were analyzed, with 46 missing values for study duration (9.98%). Study duration was the only variable that varied depending on the imputation method in the three analyzed data sets, while the known values and all other data on students remained unchanged. This enabled a realistic comparison and analysis of the effects of the three imputation methods.

However, this analysis cannot determine which imputed value is closest to the actual values for study duration, but rather examines whether different imputed values lead to different clustering results. The analyzed variables describe the progress of the students during their studies, the circumstances of their enrolment, demographic characteristics and prior education. They include the total number of ECTS credits earned, the mode of study, enrolment and completion status, study duration, age at enrolment, weighted grade point average, gender, indicators of place and county of residence, area of prior education and the total number of credits earned at enrolment.

3.2. Missing-Data Imputation Strategies

Mean imputation replaces each missing value of study duration with an equal value calculated from the known data. However, assigning the same value to all incomplete records

© 2026 ITMA. All rights reserved.

reduces the differences among these students, and may reduce the variability of study duration [4].

Random Forest imputation estimates missing values using a set of decision trees built on the relationships between the available variables. This method can capture nonlinear relationships and interactions without assuming a single predetermined mathematical relationship between the variables [5]. Unlike mean imputation, each student with missing data can be assigned a different estimated study duration.

MICE estimates missing values by iterating through a series of conditional models. At each step, the values of the incomplete variable are estimated using the other available variables, after which the newly estimated values are used in subsequent steps [6]. Thus, the estimated study duration may depend on other characteristics of the individual student.

The three methods thus produce different datasets. Mean replacement assigns a single value to all incomplete records, while Random Forest and MICE may assign different values to different students. By comparing the clustering results obtained from these augmented data sets, one can examine whether the clustering algorithms are sensitive to the chosen method of imputation of study duration.

3.3. Clustering Algorithms

Three clustering algorithms were investigated in this paper: K-means, Ward's hierarchical clustering and Gaussian Mixture Model (GMM).

Before clustering, the analytical variables were standardized because variables such as total number of ECTS credits, age at enrolment, study duration and weighted average grade are expressed in different numerical scales. The same standardization procedure was applied separately to each augmented data set. Each clustering algorithm was evaluated on possible solutions containing between two and ten clusters. Applying the same range of possible clusters and the same evaluation criteria to all data sets enabled a direct comparison of the results.

All three algorithms are based on different approaches to clustering data: K-means forms groups around the cluster center, Ward's method creates a hierarchical structure by successively merging clusters, while GMM represents the data as a combination of probability distributions. Therefore, changes in data variability caused by imputation may affect each algorithm differently.

3.3.1. K-means Clustering Algorithm

Compared to other clustering algorithms, K-Means features a straightforward implementation, and produces clusters with clear similarity relationships among data points. K-means iteratively assigns each observation to the nearest centroid, and updates centroid positions until the solution stabilizes. Its objective function minimizes the within-cluster sum of squared errors (SSE), meaning that the algorithm performs best when the underlying clusters are compact and reasonably well separated. However, the algorithm may struggle to

© 2026 ITMA. All rights reserved.

converge when applied to non-convex datasets. In summary, the K-Means algorithm iteratively identifies cluster centers. The notation $(C_1,\ C_2,\ \ldots,\ C_k)$ denotes the partitioning of clusters. The algorithm aims to minimize the sum of squared errors [19], as defined by the following formulas (1, 2):

$$M = \sum_{i=1}^{k} \sum_{x \in C_i} \left|\left| x - \mu_i \right|\right|^2 \tag{1}$$

where μ_i is the mean vector of C_i:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{2}$$

3.3.2. Ward's Hierarchical Clustering Algorithm

Ward's algorithm is an agglomerative hierarchical algorithm that starts from singletons, and at each step merges the pair of clusters that yields the smallest increase in within-cluster variance, computed with Euclidean distance on z-standardized features. A partition with K clusters is obtained by cutting the resulting dendrogram at the level producing K branches [20]. Compared with K-means, Ward's algorithm provides a multi-scale, deterministic solution (no random seeding) rather than a single-seeded partition for a fixed K. In addition, it offers structural diagnostics via the dendrogram to justify the probable K, and optimizes the incremental increase of SSE during merges, whereas K-means alternates assignment/centroid updates to minimize total SSE for a pre-specified K. Given the datasets used in this research, by uncovering nested relationships among observations, Ward's algorithm complements the flat partitioning produced by K-means, and tends to yield compact, interpretable clusters.

3.3.3. Gaussian Mixture Model (GMM)

A Gaussian Mixture Model is a common mixture model, where the probability density is given by a mixture of Gaussian distributions (3):

$$p(x) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(x \mid \mu_k, \Sigma_k) \tag{3}$$

where:

- x is a d-dimensional vector,
- μ_k is the mean vector of the k-th Gaussian component,
- π_k is the mixing coefficient (mixture weight) of the k-th Gaussian component,
- Σ_k is the covariance matrix of the k-th Gaussian component, and
- $\mathcal{N}(x \mid \mu_k, \Sigma_k)$ is the multivariate normal density function for the k-th component (4):

$$\mathcal{N}(x \mid \mu_k, \Sigma_k) = \frac{1}{\sqrt{2\pi^d |\Sigma_k|}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right) \tag{4}$$

Unlike K-means, which exhibits equal spherical covariances and equal priors, GMM accommodates unequal shapes, sizes and densities, and provides soft membership

© 2026 ITMA. All rights reserved.

probabilities. Compared to Ward's hierarchical clustering, GMM yields a flat, probabilistic partition rather than a merge tree, trading hierarchy for distributional flexibility [21].

3.4. Evaluation of Cluster Structure Using the Silhouette Coefficient

This section examines the internal consistency of cluster solutions obtained from three datasets produced through different imputation techniques: Random Forest (RF), mean substitution and MICE. The silhouette coefficient was selected because it provides an interpretable measure of within-cluster cohesion and between-cluster separation [8]. It was used to determine the optimal number of clusters. By using three clustering algorithms (K-means, Ward's hierarchical clustering and GMM) on each of the three datasets (imputed with MICE, Random Forest and mean substitution), 9 new datasets were created from which the level of impact of different imputation algorithms on the optimal number of clusters for each individual algorithm could be analyzed. The silhouette coefficient values for k values (2–10) represent the main criterion for determining the number of clusters and their comparison.

The silhouette coefficient compares the average distance from other records in the assigned cluster with the average distance from the records in the nearest other cluster [8]. Values near 1 indicate that the record is well distributed and clearly separated from neighboring clusters. Values closer to 0 indicate that the record is close to the boundary between clusters, while negative values indicate that it would fit better in another cluster. The average silhouette coefficient was calculated for each considered value of the number of clusters, and the solution with the highest value was selected.

The robustness of the clustering results was assessed by examining whether the same optimal number of clusters was selected in all three imputed data sets, and by comparing the silhouette coefficient values for solutions with two to ten clusters. The number of iterations required for convergence was also compared for K-means and GMM. In addition, cluster profiles were analyzed to determine whether groups of students with similar characteristics appeared in the different data sets. This is important because an equal number of clusters does not necessarily mean that they represent the same types of students.

Based on the theoretical considerations above, the following section presents an empirical evaluation of the clustering algorithms applied to student academic achievement data under different implementations of missing-data imputation. The aim is to assess their robustness, convergence behavior and validation outcomes.

4. Results

4.1. Optimal Number of Clusters

Cluster validation was performed using the silhouette coefficient due to its interpretability and applicability to all selected algorithms. This ensured the comparability of the results. Using the silhouette coefficient, the optimal number of clusters in all combinations of datasets was determined.

© 2026 ITMA. All rights reserved.

4.1.1. K-means Algorithm

Across the three datasets, K-means produced remarkably high silhouette trajectories, with the optimal solution consistently emerging at k = 8, as shown in Table 1. All three imputations selected $k = 8$ as the optimal solution, with silhouette scores differing only marginally: 0.1965 for MICE, 0.2014 for Random Forest and 0.2001 for mean substitution.

Table 1. Silhouette scores for K-Means

Number of clusters (k)	MICE	RF	Mean
2	0.187166816	0.186506	0.184494
3	0.183074993	0.18316	0.180997
4	0.181939152	0.180555	0.180802
5	0.185993738	0.185683	0.186184
6	0.158871751	0.159043	0.157965
7	0.172075513	0.175989	0.16801
8	0.196457316	0.201392	0.200137
9	0.191144898	0.186508	0.19237
10	0.181889198	0.189998	0.187211

K-means proved to be robust and resilient in choosing the imputation algorithm in this dataset. Despite the conceptual imputation differences between MICE, RF and mean substitution, the values of all clusters were almost identical, both in terms of silhouette values and the optimal number of clusters. These results imply that different imputation approaches do not significantly alter the resulting centroid-based cluster structure.

The difference between the best and second-best K-means solutions was 0.0053 for MICE, 0.0114 for Random Forest and 0.0078 for mean replacement. Although these differences are relatively small, in all three cases the highest value of the silhouette coefficient was achieved with the same number of clusters. This further supports the eight-cluster solution, as the same result was independently obtained in all three augmented data sets.

4.1.2. Gaussian Mixture Model (GMM)

GMM shows the largest variability across imputations, although the optimal number of clusters remains consistently k = 2, as shown in Table 2. For all three imputations, the optimal solution occurred at the following silhouette scores: 0.1760 (MICE), 0.1755 (Random Forest) and 0.1749 (mean substitution). As with the K-means algorithm, GMM always selects the optimal number of clusters to be two.

Table 2. Silhouette scores for GMM

Number of clusters (k)	MICE	RF	Mean
2	0.175959119	0.175467101	0.174902859
3	0.169106643	0.168920605	0.167094112

© 2026 ITMA. All rights reserved.

4	0.096446515	0.098578995	0.091341137
5	0.118028715	0.091073922	0.078863105
6	0.084468073	0.099635466	0.083843409
7	0.073147017	0.131275781	0.142458214
8	0.142311613	0.16852431	0.162545368
9	0.10967622	0.11380898	0.1629654
10	0.116177889	0.125453508	0.124061508

This indicates that the dataset consists of only two large, distinct clusters. Accordingly, GMM is less useful for detailed cluster analysis across all analyzed datasets, and should be interpreted primarily for identifying general patterns. The difference between the best and second best GMM solutions was 0.0069 for MICE, 0.0065 for Random Forest and 0.0078 for mean replacement. The chosen two-component solution was therefore consistent across all three datasets. GMM recognizes the general difference between students with lower and higher academic achievement, rather than the more detailed division provided by K-means.

4.1.3. Ward's hierarchical clustering algorithm

Ward's algorithm displays the greatest variation in optimal cluster numbers across datasets, as shown in Table 3. For Ward's algorithm, the optimal number of clusters varied substantially across imputations: $k = 10$ for MICE, $k = 6$ for Random Forest and $k = 4$ for mean substitution.

Table 3. Silhouette scores for Ward's algorithm

Number of clusters (k)	MICE	RF	Mean
2	0.162177191	0.178542088	0.141984938
3	0.167613458	0.172349335	0.177524323
4	0.181598588	0.18024678	0.194438855
5	0.153725944	0.186885398	0.186971859
6	0.166540391	0.194903458	0.173961404
7	0.167403635	0.191928324	0.191623527
8	0.171710753	0.182383155	0.173446328
9	0.177430728	0.187525854	0.168616764
10	0.184428137	0.191701679	0.17614808

Ward's algorithm showed a strong dependence on imputation approaches. For the RF dataset, the cluster structure is clearly, logically and consistently organized through multiple levels of grouping, without illogical mergers or sudden jumps in similarity (k=6). The mean substitution dataset gives more compact segmentations (k=4), whereas the MICE algorithm introduces greater variability that pushes the optimal solution towards higher values of k (k=10). These differences indicate that Ward's hierarchy is very sensitive to different dataset imputations. The difference between the best and second-best Ward solutions was only 0.0028 for MICE, 0.0030 for Random Forest and 0.0028 for arithmetic mean replacement.

© 2026 ITMA. All rights reserved.

These small differences show that several considered solutions had similar values, and the selected number of clusters should not be considered the only correct representation of the student population. The results show that Ward's hierarchical method, which tends to reduce variability within clusters, responds to relatively small changes in the position of individual records in the data set. Substitution with the arithmetic mean extracted a less detailed solution with four clusters, the Random Forest solution with six clusters, and the MICE solution with ten clusters, which was the upper limit of the considered range. Such a pattern is consistent with the greater variability retained by more complex imputation methods. However, this analysis cannot prove a direct interrelationship between the variability of the imputed values and the selected number of clusters.

In relation to RQ1, the results show that the choice of imputation approach has an algorithm-specific impact on the determination of the optimal number of clusters. K-means consistently selects the number of clusters across all imputations, GMM selects them uniformly, while Ward's algorithm displays greater variability. These results imply that the imputation of the input datasets affects the optimal number of clusters only in the case of hierarchical algorithms. On the other hand, algorithms based on centroids and distribution remain largely insensitive to the choice of dataset imputation technique.

4.2. Convergence and Algorithmic Stability

Convergence was evaluated for iterative clustering algorithms using the number of iterations required until stabilization. Across the evaluated values of k (2–10), K-means converged in an average of 10.67, 9.89 and 13.78 iterations for the mean-substitution, Random Forest, and MICE-imputed datasets, respectively (min–max ranges: 7–18, 8–15 and 9–25). Corresponding averages for the GMM were 6.56, 7.56 and 7.78 iterations (min–max ranges: 3–13, 3–18 and 6–13). For the optimal number of clusters (k = 8 for K-means and k = 2 for GMM), K-means converged in 8–9 iterations, while GMM converged in 13 iterations across all datasets. Ward's algorithm is non-iterative.

The number of iterations varied depending on the value of the cluster number and the data set because small changes in distance can affect the initial assignment of records to clusters, and subsequently shift their centers. Nevertheless, all solutions obtained by K-means show stability and optimization of K-means in all observed data sets. Ward's method does not update the cluster centers by repeating the procedure, and therefore small changes caused by imputation can affect some of the initial merging decisions and then the overall subsequent structure. This may explain why Ward's method is more sensitive to imputation than K-means and GMM.

4.3. Cluster Analysis and Interpretation

Applying three clustering techniques (K-means, Ward's algorithm and GMM) revealed a consistent and interpretable structure in students' academic behavior and progression. Although the number of clusters varied depending on the approach, all three approaches point

© 2026 ITMA. All rights reserved.

to the presence of distinct student subgroups characterized by different levels of achievement, study duration and demographic attributes.

4.3.1. K-means Algorithm

K-means produced eight clusters in all three datasets, with nearly identical centroids across imputations, which provides strong evidence of structural robustness, as shown in Table 4.

Table 4. K-means clusters for all datasets

k	ECTS			Study duration			Age at enrolment			Weighted GPA		
	Me	RF	MI	Me	RF	MI	Me	RF	MI	Me	RF	MI
0	122.1	122.1	155.2	4.9	4.9	5.9	19.4	19.4	19.3	3.1	3.1	3.1
1	129.7	129.7	143.6	8.2	8.3	7.7	20.9	20.9	20.7	2.8	2.8	2.9
2	131.5	130.9	128.8	8.9	9.2	9.3	25.9	25.8	26.3	3.1	3.0	3.1
3	4.5	4.4	6.4	2.6	2.6	2.9	20.6	20.6	20.3	0.4	0.4	0.8
4	81.9	81.5	106.0	5.0	4.9	7.9	19.6	19.6	33.4	2.1	2.1	2.7
5	152	152	143.3	7.9	7.9	6.4	22.7	22.7	19.6	2.9	2.9	2.9
6	139.1	137.6	152	5.9	5.8	8.1	19.4	19.5	22.7	3.0	3.0	2.9
7	112.9	112.9	9.3	7.5	7.7	2.8	33.8	33.8	20.3	2.8	2.8	1.3

*Note: Me=Mean, RF = Random Forest, MI = MICE

For a comparative analysis of clusters obtained using the K-means algorithm across three differently imputed datasets (Mean, RF and MICE), a qualitative assessment of profile similarity can be carried out as follows:

1. **High-achieving students** with approximately 152 ECTS and on a positive study completion track are present in all three tables:

   - Mean: Cluster 5
   - RF: Cluster 5
   - MICE: Cluster 6

2. **Older long-term studying students** (age > 33, study duration ≈ 8 years) appear consistently:

   - Mean: Cluster 7 (Age 33.8, Study Duration 7.48)
   - RF: Cluster 7 (Age 33.8, Study Duration 7.67)
   - MICE: Cluster 4 (Age 33.4, Study Duration 7.99)

3. **Very low-achieving early school leavers** with minimal ECTS (≈ 4–9) and very low weighted GPAs (0.4–1.3):

   - Mean: Cluster 3 (ECTS 4.48, Weighted GPA 0.45)
   - RF: Cluster 3 (ECTS 4.40, Weighted GPA 0.41)
   - MICE: Cluster 3 (ECTS 6.40, Weighted GPA 0.77) & Cluster 7 (ECTS 9.33, Weighted GPA 1.30).

© 2026 ITMA. All rights reserved.

4.3.2. Gaussian Mixture Model

To compare clusters obtained using the GMM algorithm across three differently imputed datasets (Mean, RF and MICE), a qualitative analysis of profile stability can be conducted based on the component parameters (centroids and probabilities) for each of the two clusters (Table 5).

Table 5. Gaussian Mixture Model clusters for all datasets

k	ECTS			Study duration			Age at enrolment			Weighted GPA		
	Me	RF	MI	Me	RF	MI	Me	RF	MI	Me	RF	MI
0	82.9	83.2	83.2	4.7	4.7	4.6	19.9	19.9	19.9	2.1	2.1	2.1
1	124.1	123.8	123.8	7.7	7.9	7.9	26.5	26.6	26.6	2.8	2.8	2.8

*Note: Me=Mean, RF = Random Forest, MI = MICE

GMM converges to exactly two clusters in all three cases (Mean, RF and MICE), confirming its high robustness to the choice of imputation approach. Both clusters exhibit very similar, nearly identical profiles across all variables among the three datasets, as shown below:

Cluster 0: Lower-Achieving Students with Shorter Study Duration

- ECTS: ≈ 82.9–83.2
- Study Duration: ≈ 4.6–4.7 years
- Age at Enrolment: ≈ 19.9 years
- Weighted GPA: ≈ 2.1

Cluster 1: Higher-Achieving Students with Longer Study Duration

- ECTS: ≈ 123.8–124.1
- Study Duration: ≈ 7.7–7.9 years
- Age at Enrolment: ≈ 26.5–26.6 years
- Weighted GPA: ≈ 2.8

4.3.3. Ward's Algorithm

To compare clusters obtained using the Ward's hierarchical clustering across three differently imputed datasets (Mean, RF and MICE), a qualitative analysis of profile stability can be performed based on the centroid values for each cluster (Table 6).

Table 6. Ward's algorithm clusters for all datasets

k	ECTS			Study duration			Age at enrolment			Weighted GPA		
	Me	RF	MI	Me	RF	MI	Me	RF	MI	Me	RF	MI
0	139.6	146.6	143.1	7.1	6.5	8.5	23.6	20.1	22.1	3.0	3.0	2.9
1	14.1	137.8	1.1	2.9	8.0	3.4	20.9	23.0	21.3	1.0	3.0	0.0
2	100.6	126.6	118.6	6.5	7.9	4.9	22.0	28.1	19.4	2.5	2.9	3.0

© 2026 ITMA. All rights reserved.

3	152.0	152.0	110.7	7.9	7.9	5.4	22.7	22.7	19.5	2.9	2.9	2.9
4		8.7	146.4		3.0	6.2		20.9	19.4		0.7	3.0
5		12.6	132.7		2.5	9.0		20.1	25.8		1.5	3.0
6			118.5			7.6			35.1			2.9
7			152.0			8.1			22.7			2.9
8			0.0			3.0			20.4			0.0
9			40.2			2.2			20.4			2.7

*Note: Me=Mean, RF = Random Forest, MI = MICE

Despite the differing number of clusters, several recurring student profiles can be identified across all imputations:

1. **High-Achieving Students** – this group remains clearly separated by high academic performance metrics:

   - Mean: Cluster 0 (ECTS 139.6, Weighted 3.0)
   - RF: Cluster 0 (ECTS 146.6, Weighted GPA 3.0) & Cluster 1 (ECTS 137.8, Weighted GPA 3.0)
   - MICE: Clusters 0, 4, 5 (ECTS 132–146, Weighted GPA ≈ 3.0)

2. **Early Dropout Students** – extremely low ECTS (≈ 0–15) and very low weighted GPA (≈ 0.0–1.5):

   - Mean: Cluster 1 (ECTS 14.1, Weighted GPA 1.0)
   - RF: Clusters 4 & 5 (ECTS 8.7 & 12.6, Weighted 0.7 & 1.5)
   - MICE: Clusters 1, 8, 9 (ECTS 0.0–40.2, Weighted GPA 0.0–2.7)

3. **Students with Moderate Results** – medium ECTS (≈ 100–120) and moderate Weighted GPA (≈ 2.5–2.9):

   - Mean: Cluster 2 (ECTS 100.6, Weighted GPA 2.5)
   - RF: Not clearly separated as a single cluster (blended with others)
   - MICE: Clusters 2 & 3 (ECTS 110.7 & 118.6, Weighted GPA 2.9 & 3.0)

The difference between the best and second-best K-means solutions was 0.0053 for MICE, 0.0114 for Random Forest and 0.0078 for arithmetic mean replacement. Although these differences are relatively small, in all three cases the highest value of the silhouette coefficient was achieved with an equal number of clusters. This further supports the eight-cluster solution because the same result was independently obtained in all three supplemented data sets. However, choosing an equal number of clusters does not necessarily mean that individual clusters represent equal groups of students. Since cluster labels are assigned arbitrarily, corresponding groups in the three analyses may have different labels. Stability therefore needs to be assessed by comparing the academic and demographic characteristics of students within each cluster.

© 2026 ITMA. All rights reserved.

4.4. Distributional Effects of Imputation

Table 7 compares the distribution of all study duration values with the values that were imputed from the 46 originally incomplete records. The imputed values are not considered to be known or correct values. They are compared to illustrate how each imputation method changes the variability and distribution of study duration.

Table 7. Distributional diagnostics for study duration

Group	Unique	Mean	SD	Minimum	Maximum
Observed	298	5.904	3.595	0.170	16.230
Mean	1	5.962	0.000	5.962	5.962
RF	46	6.363	2.019	1.420	9.732
MICE	46	6.369	2.276	−0.281	10.258

Looking at the entire imputed data set, the average study duration remains similar: 5.910 years after mean replacement and 5.950 years after imputation using Random Forest and the MICE method. However, the standard deviation increases from 3.411 years with mean replacement to 3.472 years with the Random Forest method and 3.487 years with the MICE method. Similar overall average values may therefore mask important differences in the distribution of imputed values.

5. Discussion

In this section, the implications of the findings are interpreted in relation to the research questions and the existing literature. The findings are considered at both the cluster-selection and student-profile levels. This distinction is important because an imputation method may change the number of clusters selected by an algorithm while the main types of students identified in those clusters remain recognizable.

5.1. Impacts of Imputation on Clustering (RQ1)

The results clearly show that the choice of imputation algorithm substantially affects hierarchical clustering (Ward's algorithm), leading to substantial variation in the optimal number of clusters. On the other hand, K-means and GMM remain largely stable regardless of the data imputation technique.

The analysis of imputed values helps to explain these differences. Mean substitution assigns the same value to all incomplete records, thereby reducing the differences among these students. Random Forest and MICE use data from the remaining variables, and the imputed values are distributed over a wider range. Since Ward's method successively merges clusters in an attempt to reduce the variability within them, changes in distances caused by imputation can affect the cluster structure. K-means also uses squared distances, but

© 2026 ITMA. All rights reserved.

successive updates of cluster centers in this data set have been shown to be more robust to such changes.

The answer to RQ1 therefore depends on the clustering algorithm. The imputation method has little effect on the selected number of clusters in K-means and GMM, but it clearly affects Ward's method. In addition, the small differences between the best and second-best silhouette coefficient values for Ward's method indicate that several possible solutions had similar internal structure.

5.2. Robustness of K-means and GMM (RQ2)

The consistency of K-means across imputations underscores its reliability in identifying stable, fine-grained student subgroups. Also, the two-cluster GMM approach remained virtually unchanged, indicating that this algorithm also consistently separates the data into the same basic, broad groups, even when the input data or processing procedures (e.g. imputation) change, but is not necessarily sensitive to small differences within those groups. These results confirm that certain clustering algorithms are inherently more robust to preprocessing variations.

In relation to RQ2, the strongest evidence of robustness exists when the numerical results and the interpretation of the clusters coincide, and it is K-means that retains eight clusters with similar centers, while GMM retains two components with almost identical general profiles. Ward's method selects different numbers of clusters, but identifies comparable groups of successful students, early dropouts and students with moderate academic performance.

The results therefore suggest two forms of robustness. The first concerns whether the same number of clusters and similar numerical results are obtained in differently imputed data sets. The second concerns whether the resulting clusters represent similar groups of students. This distinction is important because the algorithm can select the same number of clusters even though their characteristics have changed. On the other hand, it can select a different number of clusters, and still retain the same general student profiles.

This distinction is also important for the practical application of the clustering results. Higher education institutions can still identify broader groups, such as successful students and students who dropped out of their studies, even if the algorithm selects six clusters instead of four. However, the number of students belonging to a narrower, more precisely defined group may change. Before using the clustering results for planning student support, it is necessary to examine both the stability of the number of clusters and the similarity of the resulting student profiles.

5.3. Practical Implications and Limitations

The results suggest a practical procedure for clustering educational data containing missing values. The number and proportion of missing values should be reported as a starting point. After that, it is necessary to compare datasets of originally recorded and imputed values. The cluster evaluation should be repeated on datasets obtained by different imputation methods, and competing solutions with similar values of the silhouette coefficient should be

© 2026 ITMA. All rights reserved.

presented. Finally, the cluster profiles should be compared on the basis of multiple variables, especially those that were not imputed.

Several limitations reduce the generalizability of the obtained results. The research is based on 461 records from a single higher education institution, and considers only one variable with missing values. Different results may be obtained on larger datasets, data from multiple institutions or datasets in which values of several related variables are missing. The analysis considered solutions containing between two and ten clusters, and the evaluation is primarily based on the silhouette coefficient. The use of other internal validation measures, stability analysis on repeated samples or external outcomes could support different solutions.

Finally, clustering reveals associations and similarities among students, but does not establish causal explanations for their academic outcomes. The resulting profiles should not be interpreted as evidence that one student characteristic causes another, nor should they be used alone to automate decisions that could significantly affect students. Any institutional implementation should include expert judgment, privacy protection, periodic re-evaluation of the model and consideration of potential adverse effects on students whose records contain missing values.

6. Conclusion

In conclusion, this study demonstrates that the impact of data imputation on clustering outcomes is highly dependent on the choice of clustering algorithm. Ward's hierarchical clustering algorithm has been shown to be sensitive to different input datasets, while K-means and GMM have been shown to be robust in the analyzed dataset. K-means consistently selected eight clusters across all three completed datasets, while GMM consistently selected two components. In contrast, Ward's method selected ten clusters for MICE, six for Random Forest and four for mean substitution. These results demonstrate that changes introduced by imputation do not affect all clustering algorithms in the same way. By comparing multiple clustering algorithms and different imputation approaches, this research provides an understanding of how carefully selected preprocessing choices can shape and identify student profiles. The analysis of the imputed values helps to explain the observed differences. Mean substitution assigns the same study-duration value to all incomplete records, and therefore reduces the differences among those students, whereas Random Forest and MICE retain substantially more variation. The resulting changes in the positions and distances between observations may consequently influence the cluster structure, particularly when Ward's method is applied. These insights contribute to the methodological discussion in educational data mining, and offer practical guidance for researchers and practitioners in the field.

The obtained results are consistent with the findings from previous studies. Although these conclusions are predictable from a theoretical perspective, their empirical confirmation on educational data on student academic achievement provides practical applications and analysis for data mining in education. Despite the differences in the number of clusters selected by Ward's method, comparable profiles of high-achieving students, early school leavers and students with moderate academic results remained recognizable across the three completed datasets. Therefore, robustness should be considered both in terms of whether the

© 2026 ITMA. All rights reserved.

same number of clusters is obtained, and whether the resulting clusters represent similar groups of students.

Overall, the results for RQ1 and RQ2 show that hierarchical clustering is more sensitive to imputation choices, while centroid- and distribution-based algorithms provide more stable and robust results. Consequently, educational clustering studies should compare several plausible imputation methods, examine the distribution of the imputed values, and report closely competing clustering solutions instead of relying exclusively on the solution with the highest silhouette coefficient. Before clustering results are used to support institutional decisions, both the stability of the selected number of clusters and the consistency of the resulting student profiles should be evaluated.

Future research should examine larger and more diverse student populations, datasets containing missing values in several variables and additional missing-data mechanisms. The use of repeated algorithm initializations, resampling-based stability measures and external educational outcomes could provide further evidence regarding the reliability and practical relevance of the identified student profiles.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results. All authors declare that there are no competing interests.

References

[1] Onan, A.; Toçoğlu, M. A. Weighted word embeddings and clustering-based identification of question topics in MOOC discussion forum posts. *Comput Appl Eng Educ*, **2021**; vol. 29: 675–689. DOI:10.1002/cae.22252.

[2] Mao, L. (2022). Resource Sharing Method of Basic Computer Education Based on Mixed Gaussian Model. *Mobile Information Systems*, 6325329, 9 pages, DOI:10.1155/2022/6325329.

[3] Kisić, E.; Raspopović Milić, M. R.; Jović, J.; Zdravković, N.; Selimović, F. (2024). Comparative Analysis of Student Performance Across Different Cohorts in Higher Education. In *eLearning* (pp. 24–31).

[4] Little, R. J. A.; Rubin, D. B. Statistical Analysis with Missing Data, 3rd ed.; Wiley: Hoboken, NJ, USA, 2019. https://doi.org/10.1002/9781119482260

[5] Stekhoven, D. J.; Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 2012, 28, 112–118. https://doi.org/10.1093/bioinformatics/btr597

[6] van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 2011, 45, 1–67. https://doi.org/10.18637/jss.v045.i03

[7] Jain, A. K. Data clustering: 50 years beyond K-means. Pattern Recognition Letters 2010, 31, 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

[8] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65, https://doi.org/10.1016/0377-0427(87)90125-7

© 2026 ITMA. All rights reserved.

[9] Radišić, B.; Seljan, S.; Dunđer, I. Impacts of missing value imputation on clustering algorithms in academic achievement data. In *Recent Trends and Challenges in Information Systems and Technologies: WorldCIST 2026, Volume 1*; Rocha, A.; Adeli, H.; Moreira, F., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2027; Lecture Notes in Networks and Systems, Volume 2091, pp. 23–35. https://doi.org/10.1007/978-3-032-32029-2_3

[10] Ahmed, A.; Zualkernan, I.; Elghazaly, H. (2021, July). Unsupervised clustering of skills for an online learning platform. In 2021 International Conference on Advanced Learning Technologies (ICALT) (pp. 200–202). DOI:10.1109/ICALT52272.2021.00066.

[11] Antolijao, T. D. et al. (2024). Performance Analysis of Multi-Feature Data-to-Image Encoding Across Different Clustering Methods for Labeling Anomalous Machine Data Under Predictive Maintenance, 2024 IEEE 16th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, Baguio, Philippines, 2024, pp. 1–6, DOI: 10.1109/HNICEM64917.2024.11258645.

[12] Iatrellis, O.; Savvas, I. K.; Fitsilis, P.; Gerogiannis, V. C. (2021). A two-phase machine learning approach for predicting student outcomes. Education and Information Technologies, 26(1), 69–88, https://doi.org/10.1007/s10639-020-10260-x

[13] Jin, J. (2025). Student behavior patterns in vocational education big data based on clustering algorithm. Discov Artif Intell 5, 197 (2025). DOI:10.1007/s44163-025-00433-3.

[14] Urs, A. L. A. D.; Sudharshan, A. (2025). A multi-factor machine learning framework for predicting and profiling student academic performance using behavioral, financial, and wearable data. MethodsX 2025, 15, 103673. https://doi.org/10.1016/j.mex.2025.103673

[15] Hendriks, R. A.; de Jong, P. G.; Admiraal, W. F.; Reinders, M. E. (2024). Motivation for learning in campus-integrated MOOCs: Self-determined students, grade hunters and teacher trusters. *Computers and Education Open*, *6*, 100158. DOI:10.1016/j.caeo.2023.100158.

[16] von Luxburg, U. Clustering stability: An overview. *Foundations and Trends in Machine Learning* **2010**, *2*, 235–274. https://doi.org/10.1561/2200000008

[17] Audigier, V.; Niang, N. Clustering with missing data: Which equivalent for Rubin's rules? *Advances in Data Analysis and Classification* **2023**, *17*, 623–657. https://doi.org/10.1007/s11634-022-00519-1

[18] Radišić, B.; Dunđer, I.; Seljan, S. (2025). Impact of Imputation Methods on the Performance of Classification Algorithms. In Proceedings of the 36th International Scientific Conference: Central European Conference on Information and Intelligent Systems: CECIIS 2025 (pp. 419–427). Varaždin: University of Zagreb Faculty of Organization and Informatics.

[19] Lloyd, S. (1982). Least squares quantization in PCM. IEEE transactions on information theory, 28(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

[20] Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236–244., https://doi.org/10.1080/01621459.1963.10500845

[21] McLachlan, G. J.; Peel, D. Finite Mixture Models; Wiley: New York, NY, USA, 2000. https://doi.org/10.1002/0471721182

© 2026 ITMA. All rights reserved.