Impact of Imputation Methods on the Performance of Classification Algorithms

Bojan Radišić
Faculty of Tourism and Rural Development in Požega, Josip Juraj Strossmayer University of Osijek
Vukovarska 17, 34000 Požega, Croatia
bradisic@ftrr.hr

Ivan Dunđer, Sanja Seljan
Faculty of Humanities and Social Sciences, University of Zagreb, Department of Information and Communication Sciences
Ivana Lučića 3, 10000 Zagreb, Croatia
{idundjer, sseljan}@ffzg.unizg.hr

Abstract*. This paper analyses the impact of different imputation methods on the performance of classification algorithms. This was done to predict students' academic success. The original dataset consisted of student data from the Faculty of Tourism and Rural Development. Within this dataset, 46 missing values (around 10%) related to study duration were identified. Imputation methods, such as Random Forest, Gradient Boosted Trees, K-Nearest Neighbors, Multiple Imputation by Chained Equations and arithmetic mean imputation, were applied. Using these methods resulted in five distinct datasets that were used to train and evaluate the classification algorithms using Monte Carlo validation to ensure model assessment stability. For each iteration, classification performance metrics were calculated. A comparative analysis of all models provides insight into how imputation methods affect the performance of classification algorithms.*

Keywords. Machine learning, Data imputation, Missing data, Monte Carlo validation, Accuracy

1 Introduction

Educational Data Mining (EDM) has become an increasingly prominent area of research since its appearance in the literature in 2007 (Romero & Ventura, 2007). In Croatia, higher education institutions systematically collect data on students through the ISVU system (Cro. Informacijski sustav visokih učilišta, Eng. Information System of Higher Education Institutions), which serves as the central database for academic records.

The ISVU system supports the management of various types of academic information, including enrolment history, ECTS credit points, exam results and student progress indicators. When such data is analysed using machine learning techniques, it can yield valuable insights into patterns of academic achievement and student behaviour.

These insights are often used to improve educational strategies, identify students at risk of dropping out, and monitor trends in student retention and academic success.

In this research, data was obtained from ISVU for the period from 2010 to 2018. The dataset included instances of missing information on study duration. Such omissions are usually attributed to factors such as manual data entry errors, technical malfunctions, student non-response, cancellation of study programs, or merging of heterogeneous datasets (Emmanuel et al., 2021). The primary objective of this research is to evaluate how different imputation methods affect the predictive performance of machine learning classifiers in the context of student academic success.

Unlike earlier work (Radišić et al., 2023), which was limited to simple statistical imputations such as mean, median and geometric mean, this study extends the analysis to machine learning-based methods (Random Forest, Gradient Boosted Trees, K-Nearest Neighbors) and a hybrid approach (MICE). In addition, the models were evaluated using Monte Carlo cross-validation with multiple classification algorithms (Random Forest, Support Vector Machine, Naïve Bayes, K-Nearest Neighbors). This broader methodological scope provides new insights into how different imputation techniques interact with classifiers, demonstrating that the Random Forest algorithm remains the most robust and reliable across diverse imputation strategies.

The paper is organized as follows: after the Introduction, the second section provides a literature review. The third section, Methodology, contains a description of the dataset, an analysis of implemented machine learning algorithms, three main imputation strategies, and the testing process. In the fourth section, the results are presented, comparing the models of each individual algorithm and comparing all models among themselves, using accuracy and F1-score. The last section provides a conclusion and suggestions for future research.

2 Literature review

Research on predicting student success using machine learning has grown rapidly in the past decade, reflecting the increasing availability of educational data and the demand for evidence-based decision-making in higher education. Numerous studies have examined how student performance can be modelled, which algorithms are most effective, and what factors contribute to academic outcomes. A total of 438 articles were initially identified, and after a rigorous selection and quality assessment process, 70 studies published between 2018 and 2023 were included in the final review (Pelima et al., 2024). Their analysis showed that most studies focus on student performance prediction in higher education using machine learning techniques, but do not explicitly address how missing data affects predictive performance. This limitation highlights the importance of this particular study, which directly examines the role of imputation techniques in educational data mining.

Several machine learning algorithms, including Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Logistic Regression (LR), have proven to be effective in predicting academic success among students (Yağcı, 2022). In the study, performance on midterms emerged as a key predictor of outcome, i.e. academic success.

Similarly, (Campanilla, 2024) used a Naïve Bayes classifier to predict study completion. The results showed that out of 272 enrolled students, 200 of them (73.6%) had a higher grade point average (GPA) and a higher probability of completing their studies, while the remaining 72 (26.4%) were considered at risk for completing their studies. Based on these insights, the study advocated for a revision of institutional strategies to better address and reduce student attrition.

One model has been developed for the early identification of at-risk students using supervised machine learning techniques, including Support Vector Machine (Martinez et al., 2024). Trained on engagement, demographic and academic performance data, the model demonstrated high accuracy, with the SVM achieving 94% precision, confirming its effectiveness in early detection of academic risk.

A systematic review of 33 peer-reviewed studies published between 2010 and 2023 focused on the methods, datasets and frameworks used to handle missing values (Setiawan et al., 2023). The authors concluded that although no single imputation method is universally superior, statistical methods, K-Nearest Neighbors, Random Forest and hybrid frameworks are among the most commonly used, and their effectiveness is greatly influenced by the nature of the dataset and the objectives of the analysis.

The impact of four imputation techniques – Multivariate Imputation by Chained Equations (MICE), HMISC, Amelia and MissForest – on the predictive performance of eight supervised machine learning algorithms was investigated in a study focusing on the F1-score as the primary evaluation metric (Maale et al., 2025). The authors found that all imputation methods were sensitive to both proportion and mechanism of missingness (Missing Completely at Random – MCAR, Missing at Random – MAR, Missing Not at Random – MNAR), with MissForest achieving the highest average F1-score, while MICE performed consistently well, especially in MAR conditions.

Recent developments show that state-of-the-art (SOTA) imputation methods, especially those based on deep learning and adversarial networks, are being applied in data mining. A thorough review of GAN-based algorithms for imputation of missing data, has shown that they outperform traditional statistical and machine learning methods in accuracy (Shahbazian & Greco, 2023). Through experiments and analysis, the authors have demonstrated that GANs represent a state-of-the-art approach for handling incomplete datasets across various domains.

This particular study therefore contributes by systematically comparing classical, machine learning-based and hybrid imputations, providing a benchmark that can guide future integration of SOTA techniques in educational contexts.

3 Methodology

The data was collected from students of the Professional Study of Commerce at the Faculty of Tourism and Rural Development in Požega, Juraj Strossmayer University of Osijek, Croatia, from 2010 to 2018. Missing values were imputed using five methods: Random Forest (RF), Gradient Boosted Trees (GBT), N-Nearest Neighbors (KNN), MICE and arithmetic mean (Average).

After imputation, classification was performed using four machine learning algorithms: Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). Each algorithm with different imputation combination was evaluated using the Monte Carlo cross-validation method with 30 random and 70/30 train-test splits. All obtained models were compared based on accuracy and F1-score.

3.1. Dataset

The original dataset used in this research was collected from 461 students of the Professional Study of Commerce at the Faculty of Tourism and Rural Development in Požega, Juraj Strossmayer University of Osijek, Croatia.

The study program lasts 6 semesters over three academic years, and has a total of 180 ECTS credit points. The dataset includes all students enrolled from the academic years 2010/2011 to 2018/2019. Out of the total 461 students, there were 195 male students (42%) and 266 female students (58%). There were 264 full-

time students (57%) and 197 part-time students (43%). Eleven input variables, so-called data features, were selected for this study, as listed and described in Table 1.

The output variable focuses on the accumulated, i.e. achieved *ECTS credit points*, and is divided into three categories:

1. Full ECTS credit points (180 ECTS),
2. Achieved ECTS credit points (1-179 ECTS),
3. No ECTS credit points (0 ECTS).

Table 1. Data features

Features	Description and possible values
Special status	Admission based on special status
Enrolled	Part-time or Full-time students
Gender	Male or Female
County of residence	One of the 21 counties in Croatia
Residence in the city of Požega	Residence in Požega or not
Residence in Požega-Slavonia County	Residence in Požega-Slavonia County or not
Age	Student's age at the time of enrolment
High school	Type of completed high school
Duration of study	Time from enrolment to completion (graduation) or dropout
Points upon enrolment	Points when enrolling in a study program
Grade point average	Grade point average during study

The dataset includes three student categories: the first is comprised of 172 students who have earned all 180 ECTS credit points and completed their studies; the second group consists of 199 students who are still active, have passed certain courses, but have not accumulated all 180 ECTS credit points; and the third category includes 90 students classified as passive, as they have not earned any ECTS credit points.

The dataset was missing 46 values related to the *duration of study*. Therefore, they were imputed using five different methods for handling missing data: arithmetic mean, K-Nearest Neighbors (KNN), Multiple Imputation by Chained Equations (MICE), Random Forest (RF), and Gradient Boosted Trees (GBT).

To assess the consistency of the different techniques for imputing missing values, a correlation analysis was performed between four advanced imputation methods: KNN, MICE, RF and GBT. Pearson correlation was calculated for rows that originally contained missing values in the variable *duration of study*.

According to Table 2, it is evident that the highest correlation was observed between Random Forest and Gradient Boosted Trees (r = 0.969), which is expected given their similar architectures based on ensemble trees. The MICE method showed moderate correlation with both GBT (r = 0.554) and Random Forest (r = 0.515).

Table 2. Correlation matrix

		RF	GBT	MICE	KNN
RF	Pearson's r	—			
	df	—			
	p-value	—			
GBT	Pearson's r	0.969***	—		
	df	44	—		
	p-value	<.001	—		
MICE	Pearson's r	0.515***	0.554***	—	
	df	44	44	—	
	p-value	<.001	<.001	—	
KNN	Pearson's r	0.332*	0.364*	0.594***	—
	df	44	44	44	—
	p-value	0.024	0.013	<.001	—

Note: * p < .05, ** p < .01, *** p < .001

KNN showed a slightly higher correlation with MICE (r = 0.594), but relatively low correlations with Random Forest (r = 0.332) and GBT (r = 0.364), indicating that it relies on local proximity between samples, unlike tree-based methods.

The correlation analysis highlights strong agreement between tree-based machine learning models, particularly Random Forest and GBT. MICE and KNN, although algorithmically different, still exhibit moderate consistency with ensemble methods.

3.2. Machine learning

Machine learning techniques are often applied to predict academic success (achievements) among students. Key factors such as exam scores, prior educational performance, demographic details and class attendance play a significant role in determining study outcomes. Various machine learning models can be used for this predictive task, such as Decision Trees, Random Forest, K-Nearest Neighbors, Naïve Bayes and Neural Networks.

3.2.1. Random Forest (RF)

The Random Forest (RF) algorithm is widely applied in both classification and regression tasks due to its robustness and predictive power. It operates by generating multiple decision trees from various random subsets of the original dataset, and aggregates

their outputs, typically through averaging or majority voting, to improve overall prediction accuracy. Among its key strengths are the ability to automatically handle missing data, and its efficiency when applied to large and complex datasets.

Nevertheless, its use can be computationally demanding, requiring significant processing power and memory resources, especially when the number of trees or the amount of data increases (Kovač et al., 2022).

3.2.2. Naïve Bayes (NB)

Naïve Bayes (NB) is a type of classification algorithm based on Bayes' theorem, as stated below:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Naïve Bayes is a probabilistic classifier that assumes independence between features. Despite its simplicity, it is often competitive with more complex algorithms, especially in high-dimensional settings. In the educational context, Naïve Bayes has been effectively applied to predict student performance and academic success (Nakhipova et al., 2024).

3.2.3. K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm is a widely used nonparametric method primarily designed for classification, although it can also be applied to regression problems. It assigns a class to a new data point based on the majority votes of its k nearest (closest) neighbours within the training data, where proximity is typically measured by Euclidean distance or other distance metrics.

KNN is particularly valued for its intuitive logic and minimal model training requirements, making it simple to implement and adapt to different domains. One of the main advantages of KNN is that it does not require the construction of an explicit model or strong assumptions about the data distribution. This characteristic allows for flexible application, especially in scenarios involving proximity-based decision-making or pattern recognition (García et al., 2016).

3.2.4. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a widely used supervised machine learning algorithm designed for binary classification tasks. It works by constructing an optimal hyperplane that maximizes the margin between data points of different classes, which improves its generalization performance. One of its key strengths lies in its ability to handle both linearly and non-linearly separable data using kernel functions, which map the data into higher-dimensional spaces.

SVM is built on the principle of structural risk minimization, which contributes to its robust performance in various applications such as text classification, biomedical data analysis, financial modelling and others. However, the limitations of the method limitations include the high computational cost associated with solving large-scale quadratic programming problems, and sensitivity to kernel choice and parameter tuning (Tanveer et al., 2024).

3.3. Imputation of missing values

Addressing the problem of missing data is a key aspect of data preparation, especially in studies that rely on predictive modelling or statistical inference. Setiawan et al. (2023) provide a structured overview of imputation strategies, categorizing them into three core groups: statistical methods, machine learning-based methods and hybrid or ensemble approaches.

Statistical imputation methods are the simplest and most widely adopted, mainly due to their ease of use and minimal computational demands. These include methods such as replacing missing entries (values) with the mean, median or mode of the observed data.

Machine learning-based methods offer greater flexibility, especially when dealing with nonlinear interactions or multidimensional data. Algorithms such as KNN impute missing values based on the similarity between observations, while Random Forest-based imputation leverages ensemble decision trees to generate plausible (probabilistic) estimates. Similarly, GBT have demonstrated the ability to internally handle missing data during training, making them particularly useful in automated machine learning pipelines. These methods typically yield more accurate imputations, especially when the data structure is complex and not supported well by simple statistical heuristics.

Hybrid and ensemble approaches combine the strengths of different methods to increase robustness. One of the most well-known among them is MICE, which iteratively models each variable with missing values based on the others, thus capturing more nuanced dependencies. These approaches are particularly promising in large datasets where traditional imputation methods fall short.

3.4. Testing

From the original, i.e. initial dataset that contained missing values, five new datasets were created. In each dataset, missing values were imputed using different imputation strategies:

- statistical: arithmetic mean,
- based on machine learning: KNN, RF and GBT,
- hybrid and ensemble: MICE.

Four machine learning algorithms were applied to predict student academic success: RF, NB, SVM and KNN using all five datasets. Each algorithm was evaluated through the Monte Carlo cross-validation method with 30 random and 70/30 train-test splits.

There were 322 entries in the training set and 132 entries in the test set per iteration. A total of 20 different models were created.

4 Research results and discussion

All models were analysed both horizontally, by comparing models within each individual algorithm, and vertically, by comparing all models with each other. The comparisons focused on model accuracy and F1-score.

4.1. Accuracy

This section presents a comparative evaluation of the accuracy of four machine learning algorithms: KNN, NB, RF and SVM, based on their performance in different missing value imputation methods.

Table 3. KNN algorithm – accuracy metric

Imputation Method	Mean	Median	Standard Deviation	Minimum	Maximum
RF	0.491	0.489	0.0344	0.424	0.568
Average	0.493	0.493	0.0389	0.403	0.561
GBT	0.501	0.496	0.037	0.417	0.576
MICE	0.475	0.478	0.0318	0.396	0.532
KNN	0.489	0.489	0.0319	0.439	0.554

Table 3 shows that the highest average accuracy for the KNN algorithm was achieved using GBT imputation (Mean = 0.501), followed by Average (0.493) and RF (0.491). This suggests that GBT provides a small performance advantage for the KNN classifier. The lowest average accuracy was observed with MICE imputation (0.475), indicating that this method may not be as effective in preserving class-relevant information for the KNN algorithm.

In terms of variability, the lowest standard deviations were recorded for MICE (0.0318) and KNN imputation (0.0319), suggesting that these imputations yielded more consistent performance results, albeit with lower mean accuracy. For KNN algorithm, GBT imputation seems to achieve the best balance between performance and variability.

Table 4 shows that the RF algorithm demonstrated consistently high performance across all imputation methods. The highest average accuracy was recorded for KNN imputation (Mean = 0.810), followed by Average (0.806) and GBT (0.804). These results indicate that RF maintains robust predictive power regardless of the imputation method. In terms of consistency, GBT imputation yielded the lowest standard deviation (0.0208), indicating the most stable performance across iterations. The minimum and maximum values further highlight the effectiveness of RF. All methods achieved high minimum accuracy values ranged from 0.734 (Average) to 0.77 (GBT and KNN), and maximum values ranged from 0.842 (GBT) to 0.885 (Average), indicating the potential for near-optimal classification across different imputations.

Table 4. RF algorithm – accuracy metric

Imputation Method	Mean	Median	Standard Deviation	Minimum	Maximum
RF	0.8	0.799	0.0278	0.741	0.863
Average	0.806	0.809	0.035	0.734	0.885
GBT	0.804	0.802	0.0208	0.77	0.842
MICE	0.799	0.806	0.0242	0.755	0.871
KNN	0.81	0.809	0.025	0.77	0.863

Overall, RF combined with KNN imputation provided the best trade-off between high accuracy and acceptable variability. GBT offered the most consistent results, while Average achieved the highest peak performance.

The highest mean accuracy was observed with the Average imputation method (0.675), followed closely by GBT (0.674) and RF (0.671) for the NB algorithm, as shown in Table 5.

Table 5. NB algorithm – accuracy metric

Imputation Method	Mean	Median	Standard Deviation	Minimum	Maximum
RF	0.671	0.673	0.0422	0.583	0.748
Average	0.675	0.68	0.0476	0.59	0.763
GBT	0.674	0.68	0.053	0.496	0.763
MICE	0.659	0.669	0.0454	0.532	0.741
KNN	0.663	0.662	0.0541	0.561	0.763

This suggests that these three methods offer comparable overall predictive accuracy when used with the NB classifier. The highest performance variability was noted for GBT (SD = 0.0530) and KNN (SD = 0.0541), despite their relatively strong mean scores, implying that their effectiveness and performance may be more context-dependent or sensitive to data variation. Among the evaluated methods, Average imputation emerges as the most balanced option for the NB algorithm.

Table 6 shows that the highest average accuracy of the SVM algorithm was achieved with the RF imputation method (Mean = 0.713), indicating a slightly superior overall performance compared to other imputation methods.

Table 6. SVM algorithm – accuracy metric

Imputation Method	Mean	Median	Standard Deviation	Minimum	Maximum
RF	0.713	0.712	0.0347	0.633	0.777
Average	0.71	0.719	0.0355	0.626	0.755
GBT	0.706	0.709	0.029	0.655	0.77
MICE	0.695	0.701	0.0387	0.604	0.763
KNN	0.706	0.712	0.0441	0.583	0.791

This was followed closely by Average (0.710) and GBT/KNN (both 0.706), suggesting that all methods yielded relatively similar average results. Average imputation recorded the highest median value (0.719), although its mean was slightly lower than RF. This may indicate a more consistent central performance, with fewer lower outliers compared to other imputation methods. In terms of variability, GBT imputation had the lowest standard deviation (0.0290), indicating the most stable performance. The widest range of results was observed with KNN imputation, where the minimum accuracy dropped to 0.583, the lowest among all methods, but also achieved the highest maximum (0.791). This reinforces the interpretation that although KNN imputations may occasionally exhibit excellent performance, they also pose a risk of instability. For the most part, the SVM algorithm performs well across all imputation methods, with RF and Average imputations providing the best balance between accuracy and consistency.

Overall, the comparative analysis of the four classification algorithms (RF, SVM, NB and KNN) reveals clear differences in their classification accuracy and stability when applied to datasets with imputed values. RF emerged as the most accurate and stable algorithm, achieving average accuracy above 0.80 across all imputation methods. Support Vector Machine ranked second, with average accuracy scores ranging from 0.695 to 0.713, and showed strong consistency. Naïve Bayes showed moderate performance, with average accuracy results between 0.659 and 0.675, depending on the imputation method. Finally, K-Nearest Neighbors consistently yielded the lowest accuracy, with average scores below 0.50. Despite occasional improvements (e.g. with GBT), the algorithm exhibited high sensitivity to data variations and instability, making it the least suitable option among the evaluated models.

4.2. F1-score

This section presents a comparative evaluation of four machine learning algorithms (KNN, NB, RF and SVM) for the F1-score, based on their performance across different missing value imputation methods.

Table 7. KNN algorithm – F1-score metric

Imputation Method	Mean	Median	Standard Deviation	Minimum	Maximum
RF	0.444	0.500	0.134	0.111	0.621
Average	0.448	0.494	0.127	0.111	0.632
GBT	0.454	0.495	0.133	0.093	0.623
MICE	0.431	0.475	0.123	0.167	0.602
KNN	0.448	0.492	0.126	0.093	0.641

Table 7 shows that the highest mean F1-score (0.454) was obtained using GBT imputation, suggesting a modest performance benefit of this approach when applied with the KNN classification algorithm. In contrast, MICE imputation resulted in the lowest average F1-score (0.431), indicating a relatively less favourable outcome. The greatest variability in performance was observed with KNN imputation (Min = 0.093, Max = 0.641), which may reflect increased sensitivity of the model to specific data characteristics or imputation methods. In general, GBT imputation appears to provide the most favourable trade-off between predictive accuracy and consistency for the KNN classifier.

Table 8 demonstrates that the highest average F1-score was achieved using KNN imputation (0.839), which may indicate a slight advantage in the performance of the RF algorithm when combined with this method. Notably, all imputation methods reached a maximum F1-score of 1.000, suggesting that, under certain conditions, perfect classification was attainable regardless of the method applied. Furthermore, KNN

imputation produced the lowest standard deviation (0.0999), indicating the most consistent model performance.

Across all imputation methods, standard deviations remained relatively low (≤ 0.107), supporting the conclusion that the RF algorithm maintained a high level of classification stability regardless of the chosen imputation method. These results highlight the robustness and reliability of the RF classifier across different data preprocessing scenarios.

Table 8. RF algorithm – F1-score metric

Imputation Method	Mean	Median	Standard Deviation	Minimum	Maximum
RF	0.828	0.786	0.107	0.638	1.000
Average	0.832	0.792	0.105	0.621	1.000
GBT	0.831	0.789	0.106	0.667	1.000
MICE	0.828	0.780	0.107	0.674	1.000
KNN	0.839	0.790	0.0999	0.695	1.000

The analysis results presented in Table 9 show that the highest average F1-score for the NB algorithm was achieved using the KNN imputation method (0.700), followed closely by the GBT imputation method, which recorded a score of 0.697. This indicates strong classification performance associated with these two imputation methods.

Table 9. NB algorithm – F1-score metric

Imputation Method	Mean	Median	Standard Deviation	Minimum	Maximum
RF	0.689	0.689	0.114	0.457	0.984
Average	0.674	0.687	0.145	0.0377	0.984
GBT	0.697	0.686	0.109	0.453	0.958
MICE	0.680	0.675	0.153	0.0465	1.000
KNN	0.700	0.680	0.124	0.424	0.984

In contrast, the NB_Average method yielded the lowest minimum F1-score (0.0377), indicating potential instability or poor performance in certain iterations. The MICE imputation method reached a perfect maximum F1-score of 1.000, demonstrating its potential for optimal classification under certain conditions. However, this method also exhibited the highest standard deviation (0.153), indicating substantial variability and less consistent results. GBT imputation again demonstrated a favourable balance, achieving high performance with a low standard deviation (0.109), reinforcing its consistency across multiple iterations.

Based on these findings, it can be concluded that the NB algorithm generally exhibits stable performance when combined with different imputation methods. Among the tested methods, KNN and GBT imputations produced the most favourable outcomes in terms of average F1-score, offering a desirable combination of predictive accuracy and reliability. In contrast, Average and MICE imputations introduced greater variability, which may pose a risk of reduced performance consistency.

Table 10. SVM algorithm – F1-score metric

Imputation Method	Mean	Median	Standard Deviation	Minimum	Maximum
RF	0.754	0.678	0.154	0.520	1.000
Average	0.753	0.682	0.157	0.529	1.000
GBT	0.749	0.667	0.149	0.536	1.000
MICE	0.740	0.669	0.162	0.509	1.000
KNN	0.748	0.681	0.162	0.420	1.000

Table 10 shows that the highest average F1-score for the SVM algorithm was observed with RF imputation (0.754), indicating slightly better overall performance compared to other imputation methods. Notably, all methods reached a maximum F1-score of 1.000, suggesting that perfect classification was achieved in at least some iterations across all imputations. However, KNN imputation produced the lowest minimum value (0.420), pointing to potential sensitivity of this configuration to input data variability or specific data patterns. Among all evaluated methods, GBT imputation exhibited the lowest standard deviation (0.149), signifying the most stable performance across repeated testing. Overall, the choice of imputation method should align with the primary analytical objective, whether it is maximizing classification accuracy or ensuring robust and stable results across diverse data conditions.

A comparison of the four classification algorithms shows that Random Forest consistently produced the best results, with average F1-scores exceeding 0.82 across all imputation methods, and a peak value of

0.839 using KNN imputation. SVM ranked second, with mean F1-scores around 0.75 and good consistency, particularly when paired with GBT imputation, which provided the most stable results for this algorithm.

Naïve Bayes demonstrated moderate but more variable performance with KNN and GBT imputations, achieving relatively high mean scores around 0.70. The lowest performance was observed with the KNN algorithm, with average F1-scores generally below 0.46. Although GBT imputation slightly improved its results in the case of the KNN algorithm, KNN remained the least accurate and most sensitive to data variations.

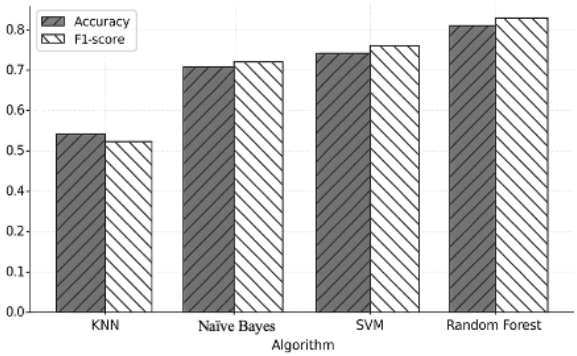

Figure 1. Highest mean values for accuracy and F1-score for each algoritm

Finally, Fig. 1 shows that Random Forest (RF) demonstrated the highest accuracy (up to 0.810) and F1-score (up to 0.839), making it the most robust and reliable classifier among the imputation methods. Support Vector Machine (SVM) showed slightly lower performance, but high stability, especially with GBT imputation. Naïve Bayes (NB) achieved moderate results, and was more sensitive to imputation variability, while K-Nearest Neighbors (KNN) yielded the weakest predictive performance among all evaluated classifiers. RF is recommended as the primary algorithm, with SVM as a strong alternative, while KNN should be avoided due to low accuracy and instability.

5 Conclusion and future research

The comparative analysis of imputation methods and machine learning algorithms clearly demonstrates the significant impact that handling missing values has on classification results in educational data mining. Among the evaluated models, Random Forest consistently achieved the highest accuracy and F1-score, demonstrating both robustness and reliability across all imputation methods. Support Vector Machine followed closely, offering stable and competitive results, particularly with ensemble imputations like GBT. Naïve Bayes, although less accurate, remained a computationally efficient alternative, especially when paired with GBT or KNN imputations. Conversely, K-Nearest Neighbors exhibited the weakest performance and the highest sensitivity to data variations, limiting its applicability in this context. These findings reinforce the importance of carefully selecting both the imputation method and classification algorithm in predictive analytics, especially in domains with limited and incomplete educational data.

A limitation of this study is that the evaluation was conducted on a single institutional dataset with approximately 10% missing values. Future research should replicate the experiments on more varied datasets from different universities and educational contexts to enhance generalizability. In addition, testing the robustness of imputation methods under higher proportions of missingness (e.g. 20-30%) would provide deeper insights into their stability. Finally, while this paper focused on statistical, machine learning-based and hybrid imputations, future work should extend the analysis to state-of-the-art approaches such as deep learning-based and GAN-based imputations, which have shown promising results in other domains.

Acknowledgments

This research was supported by the Faculty of Humanities and Social Sciences, University of Zagreb, Croatia [Institutional research project, 2025].

References

Campanilla, B. S. (2024). Forecasting Degree Completion: A Naïve Bayes Predictive Model for Students' Success. *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 1–4. https://doi.org/10.1109/ICAECT60202.2024.10469398

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1), 140. https://doi.org/10.1186/s40537-021-00516-9

García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, *98*, 1–29. https://doi.org/10.1016/j.knosys.2015.12.006

Kovač, A., Dunđer, I., & Seljan, S. (2022). An overview of machine learning algorithms for detecting phishing attacks on electronic messaging services. *2022 45th Jubilee International*

Convention on Information, Communication and Electronic Technology (MIPRO), 954–961. https://doi.org/10.23919/mipro55190.2022.9803517

Maale, F. D., Okyere, G. A., & Awe, O. O. (2025). Effects of Imputation Techniques on Predictive Performance of Supervised Machine Learning Algorithms. In O. O. Awe & E. A. Vance (Eds.), *Practical Statistical Learning and Data Science Methods* (pp. 29–48). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72215-8_2

Martinez, A. L. J., Sood, K., & Mahto, R. (2024). *Early Detection of At-Risk Students Using Machine Learning* (arXiv:2412.09483). arXiv. https://doi.org/10.48550/arXiv.2412.09483

Nakhipova, V., Kerimbekov, Y., Umarova, Z., Suleimenova, L., Botayeva, S., Ibashova, A., & Zhumatayev, N. (2024). Use of the Naive Bayes Classifier Algorithm in Machine Learning for Student Performance Prediction. *International Journal of Information and Education Technology*, *14*(1), 92–98. https://doi.org/10.18178/ijiet.2024.14.1.2028

Pelima, L. R., Sukmana, Y., & Rosmansyah, Y. (2024). Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review. *IEEE Access*, *12*, 23451–23465. https://doi.org/10.1109/ACCESS.2024.3361479

Radišić, B., Seljan, S., & Dunđer, I. (2023). *Impact of missing values on the performance of machine learning algorithms*. 54–62. https://urn.nsk.hr/urn:nbn:hr:277:483201

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146. https://doi.org/10.1016/j.eswa.2006.04.005

Setiawan, I., Gernowo, R., & Warsito, B. (2023). A Systematic Literature Review On Missing Values: Research Trends, Datasets, Methods and Frameworks. *E3S Web of Conferences*, *448*, 02020. https://doi.org/10.1051/e3sconf/202344802020

Shahbazian, R., & Greco, S. (2023). Generative Adversarial Networks Assist Missing Data Imputation: A Comprehensive Survey and Evaluation. *IEEE Access*, *11*, 88908–88928. https://doi.org/10.1109/ACCESS.2023.3306721

Tanveer, M., Rajani, T., Rastogi, R., Shao, Y. H., & Ganaie, M. A. (2024). Comprehensive review on twin support vector machines. *Annals of Operations Research*, *339*(3), 1223–1268. https://doi.org/10.1007/s10479-022-04575-w

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z