Impact of missing values on the performance of machine learning algorithms

Radišić, Bojan; Seljan, Sanja; Dunđer, Ivan

Impact of missing values on the performance of machine learning algorithms

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

[1] *Josip Juraj Strossmayer University of Osijek, Faculty of Tourism and Rural Development in Požega, 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*

Abstract

Machine learning (ML) can be used to analyze and predict student success outcome in order to avoid various problems and to plan future actions for helping students overcome difficulties during their study. This paper analyzes data from a digital system of 309 students who were enrolled in the Specialist Study in Trade Business at the Faculty of Tourism and Rural Development from 2010 to 2018. The paper explores the impact of four different data sets on the performance of ML algorithms. The first data set is with partially missing data on the length of study (around 7%), the second one uses arithmetic means in place of missing data, the third is based on median values, whereas the fourth uses the geometric mean instead. Four popular ML algorithms were considered: k-Nearest Neighbors (KNN), Naïve Bayes (NB), Random Forest (RF) and Probabilistic Neural Network (PNN). All of them are used for predicting student success based on achieved ECTS credit points. The aim of this paper is to compare and analyze the impact of missing values on the results of individual ML algorithms.

Keywords

Machine learning, Neural Network, Missing data, Confusion matrix, Accuracy

1. Introduction

Educational data mining (EDM) has been used in literature since 2007 [1]. Higher education institutions in Croatia collect students' data in a student information system called ISVU – *Informacijski sustav visokih učilišta* (Information System of Higher Education Institutions).

The system enables management of student-related information, such as enrollment metadata, ECTS credits, grades, and student progress. This data can be useful for predicting student performance when paired with machine learning.

The results are often used to improve the education system and student learning outcomes, to detect dropout students, dropout rates etc.

The data set in this research is collected from ISVU and contained some missing data for the length (duration) of study in the period from 2010 to 2018. Missing values are usually attributed to human error when processing data, or to machine errors due to malfunctioning of equipment, to respondents' refusal to answer certain questions, to cancellation of study programs, and to merging of unrelated data [2].

Missing data in the field "Length of study" is replaced by applying traditional methods for processing incomplete data. Here, there authors chose the *medium substitution* (MS) method that allows solving the problem of incompleteness of data by replacing each missing entry with an average value [3]:

- arithmetic average (arithmetic mean),
- median, or
- geometric mean.

Proceedings of RTA-CSIT 2023, April 26–27, 2023 Tirana, Albania
EMAIL: bradisic@ftrr.hr (Bojan Radišić); sseljan@ffzg.unizg.hr (Sanja Seljan); idundjer@ffzg.unizg.hr (Ivan Dunđer)

2. Literature review

More than 300 relevant papers were written in English related to the prediction of study success, between 2009 and 2021 [4]. Most of the research is focused on various predictions of success during studies (grade point average, enrollments in higher years etc.).

In most of the research, two approaches are used: one uses classification of educational data [5], while the other uses methods to predict student success [6].

The papers compare different algorithms to determine a more accurate method for a selected data set. The Random Forest (RF) algorithm had one of the highest classification accuracies (90%) [7], while k-Nearest Neighbors (KNN) had the lowest classification accuracy, contributing to the early prediction of students with a high risk of failure [8].

For predicting students' academic achievement at the end of a four-year bachelor's degree study program called *Information Technology* at a public sector engineering university in Pakistan, the best accuracy (83,65%) was achieved by using Naïve Bayes (NB).

Probabilistic Neural Network (PNN) achieved the best accuracy results for MOOC's student results classification when compared to other classification algorithms [9]. It was tested with feature selection using neighborhood component analysis for classification, and here feature selection is done by using statistical measures, measures from information theory and interclass distance. The best accuracy score was 93,4%.

Machine learning (ML) algorithms can usually make correct predictions unless the data used to train the algorithms is wrong. In many cases, data is either missing or entered incorrectly by humans, resulting in incorrect predictions. One of the main problems with data quality are missing values. Missing values in the data set can significantly increase the computational cost, distort the outcome, frustrate and mislead researchers [10].

The most widely used methods that tackle this problem fall into three main categories:

1. Deletion Methods (listwise deletion, i.e. complete-case analysis, pairwise deletion, i.e. available-case analysis)
2. Single Imputation Methods (mean/mode substitution, linear interpolation, Hot deck/cold deck)
3. Model-Based Methods (regression, multiple imputation, k-Nearest Neighbors) [11].

Single Imputation Methods are used widely for their simplicity, especially when compared with more complex imputation methods [12].

One of the most important tools for precision analysis is the confusion matrix that consists of four values: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).

Accuracy and F1 score are commonly used for comparisons in order to determine the accuracy of ML algorithms using confusion matrix results [13]. Accuracy evaluates the ratio of the number of correct predictions and the total number of samples. F1 score is the harmonic mean of precision and recall [14].

3. Methodology

The data was collected from students of the Specialist Study in Trade Business at the Faculty of Tourism and Rural Development in Požega, Juraj Strossmayer University of Osijek, Croatia. Four data sets were used:

1. without missing data (i.e. missing values were dropped),
2. with missing data that is averaged using arithmetic means,
3. with missing data that uses median values instead,
4. with missing data that uses geometric means instead.

The following machine learning algorithms were employed in order to show the impact of the data sets on the algorithms: k-Nearest Neighbors (KNN), Naïve Bayes (NB), Random Forest (RF) and Probabilistic Neural Network (PNN). The test results are analyzed using confusion matrices, accuracy and F1 score.

3.1. Data set

The data used in this research was collected from 309 students of the Specialist Study in Trade Business. The study program has 4 semesters during two academic years and has a total of 120 ECTS credit points. The data set includes all students enrolled from the academic years 2010/2011 to 2018/2019. Overall, there are 203

female students (65,7%) and 106 male students (34,3%). Also, there are 44 full-time students (14,2%) and 265 part-time students (85,8%).

Ten input variables were chosen as listed and described in Table 1.

The output variable focuses on the collected, i.e. achieved *ECTS credits*. This is divided into three categories:

1. Full ECTS credits (120 ECTS) – FE,
2. Achieved ECTS credits (1-119 ECTS) – AE,
3. No ECTS credits (0 ECTS credits) – NE.

Table 1
Data features

Features	Description
Enrolled	Part-time students and full-time students
Gender	Male or Female
Country of birth	
Place of birth	Abroad or one of 21 counties in Croatia
Place of residence	Abroad or one of 21 counties in Croatia
Age	The age of the student at the time of enrollment
High school	Type of finished high school
Length of study	Time of enrollment until graduation/dropout
Graduated	Yes or no
Grade point average	

The first category are students who have successfully completed their studies with 120 ECTS points – in total 279 students. The second category are active students who passed some exams but did not manage to finish their studies (1-119 ECTS) – 11 students in total. The third category refers to passive students, i.e. those who didn't collect any ECTS credits (0 ECTS) – 19 students totally.

During pre-processing of the data set, the authors noticed that there are 22 missing data points for the variable (field) "Length of study", which is about 7% of the total number of entries. Therefore, four data sets were created:

1. Set one – data set is the initial one with the missing data in the field "Length of study",
2. Set two – missing data is replaced by using the arithmetic mean of values in the filed "Length of study" that were available for all other students in the data set,
3. Set three – missing data is replaced by using median values instead,
4. Set four – missing data is replaced by using geometric mean values instead.

3.2. Machine Learning

Machine learning (ML) algorithms are often used to predict student success. Student data such as grades, prior academic performance, demographics, and class attendance directly affect final student success. There are many different ML algorithms that can be used for predicting student success, including Decision Trees, Random Forest, k-Nearest Neighbors, Naïve Bayes, and Neural Networks.

3.2.1. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is an algorithm used for classification and regression. KNN is mostly used as a classifier. It classifies data based on the closest or neighboring training examples in each region of data points [15].

Advantages of KNN are as follows: the algorithm is simple and easy to implement; there is no need to build a model, tune several parameters, or make additional assumptions; and the algorithm is versatile [16]. It can also be used for proximity searching.

3.2.2. Naïve Bayes (NB)

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

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

The Naïve Bayes algorithm is easy to implement and fast to perform, which makes it popular for classifying large data sets. Sometimes, the feature independence assumption may be wrong in some cases, which may lead to worse classification results. It is a probabilistic classifier, which means it predicts based on the probability of an object [17].

3.2.3. Random Forest (RF)

Random Forest (RF) is an algorithm used for classification and regression. Random Forest is easy to use and a stable classifier with many

interesting properties. This algorithm contains numerous decision trees on different subsets of the data set, and takes the average to increase the predictive accuracy of that data set. Advantages of the Random Forest algorithm are in the automatization of lost values in data and its efficiency in handling large data sets. On the other hand, disadvantages are in the context of more computing and more resources that are needed for efficient results [18].

3.2.4. Probabilistic Neural Network (PNN)

Probabilistic Neural Networks (PNNs) are a group of artificial neural networks built using the Parzen's approach to devise a family of probability density function estimators that would asymptotically approach Bayes optimal by minimizing the "expected risk", which is known as "Bayes strategies". PNNs have shown great potential for tackling complex scientific and engineering problems [19].

3.3. Testing

The data set was split into two sets – training and test set, in a ratio of 80:20, respectively. There were 247 entries in the training set, and 62 entries in the test set. The software KNIME was used for training and testing all of the selected machine learning algorithms. The authors also used a seed value during splitting data in the data set to avoid data inconsistencies. It means that always the same rows from the data set were taken for training and testing.

At first, the initial data set was tested and analyzed by ignoring missing data. After that, three different data sets were tested and analyzed, by using arithmetic mean, median and geometric mean instead of missing data from the initial data set.

All four data sets were trained and tested by the same four algorithms: KNN, NB, RF and PNN.

4. Results

The next three subsections discuss the research results that were obtained by applying confusion matrices, accuracy statistics and the F1 score. All results are separately presented in a tabular form or graphically with regard to each ML algorithm (KNN, NB, RF and PNN).

4.1. Confusion matrix

The purpose of the confusion matrix in machine learning is to evaluate the classification performance of used ML algorithms. A confusion matrix is a two-dimensional table showing the overall results of true positive, true negative, false positive, and false negative predictions made by the tested algorithms.

In this paper, 62 samples were selected for testing. At first, KNN is used for testing of all four data sets. The confusion matrix for all four data sets is shown in Table 2.

In the first confusion matrix, the initial data set failed to classify all the data so that 5 remained uncategorized, and a total of 57 were categorized. This is because in the used software only KNN tries to include the missing data points, whereas other algorithms ignore them.

Overall, only in the first confusion matrix (with the initial data) one NE was correctly detected, in contrast to the other confusion matrices where it was not recognized. In all data sets there was an improvement in the recognition of FE when compared to the initial data set. The averaged data set and the median data set correctly identified AE. In the geometric mean data set, only FC is identified, whereas AE and NE were missed.

Table 2
KNN confusion matrix

	Predicted data → / Actual data ↓	Achieved ECTS	Full ECTS	No ECTS
Initial data	Achieved ECTS (AE)	0	1	0
	Full ECTS (FE)	0	51	0
	No ECTS (NE)	1	3	1
Average	Achieved ECTS (AE)	1	3	0
	Full ECTS (FE)	1	53	0
	No ECTS (NE)	1	3	0
Median	Achieved ECTS (AE)	1	3	0
	Full ECTS (FE)	0	54	0
	No ECTS (NE)	1	3	0
Geometric mean	Achieved ECTS (AE)	0	2	0
	Full ECTS (FE)	1	54	0
	No ECTS (NE)	1	4	0

In the second case, NB is used for testing all of the four data sets (Table 3).

Table 3
NB confusion matrix

	Predicted data / Actual data	Achieved ECTS	Full ECTS	No ECTS
Initial data	Achieved ECTS (AE)	0	2	0
	Full ECTS (FE)	1	54	0
	No ECTS (NE)	1	1	3
Average	Achieved ECTS (AE)	2	0	2
	Full ECTS (FE)	0	54	0
	No ECTS (NE)	2	0	2
Median	Achieved ECTS (AE)	2	0	2
	Full ECTS (FE)	0	54	0
	No ECTS (NE)	2	0	2
Geometric mean	Achieved ECTS (AE)	2	0	0
	Full ECTS (FE)	0	55	0
	No ECTS (NE)	2	0	3

In the first confusion matrix with the initial data set, NB correctly predicted 54 FE and 3 NE, but AE was not recognized. On the other hand, in all other confusion matrices NB recognized 2 instances of AE. Only in the first set, three cases of FE were not recognized. In other data sets FT was identified correctly, with no false recognitions. Note that in the last confusion matrix with the geometric mean data set, only NE was wrongly recognized as AE, while all other cases were correctly recognized.

Table 4
RF confusion matrix

	Predicted data / Actual data	Achieved ECTS	Full ECTS	No ECTS
Initial data	Achieved ECTS (AE)	0	1	1
	Full ECTS (FE)	0	55	0
	No ECTS (NE)	1	1	3
Average	Achieved ECTS (AE)	3	0	1
	Full ECTS (FE)	0	54	0
	No ECTS (NE)	4	0	0
Median	Achieved ECTS (AE)	2	0	2
	Full ECTS (FE)	0	54	0
	No ECTS (NE)	3	0	1
Geometric mean	Achieved ECTS (AE)	2	0	0
	Full ECTS (FE)	0	55	0
	No ECTS (NE)	3	0	2

In the third case, RF is used for testing all of the data sets (Table 4). There are some similarities when compared to the NB case.

Like in the previous case with the NB confusion matrix, in the first confusion matrix with the initial data set, RF correctly predicted 55 FE and 3 NE instances, but AE was not recognized. Only in the first set, two cases of FE were not recognized. In other data sets FT was correctly identified, with no false recognitions, which is similar to the NB case.

Note the last confusion matrix with the geometric mean data set, where only NE was wrongly recognized as AE, while all other cases were correctly recognized (same as in the case of NB).

In the fourth case, PNN is used for testing all of the four data sets (Table 5). In all data sets, PNN did not recognize any of the AE. Only in the initial data set NE cases were recognized, while in the other data sets NE cases were not recognized.

In the averaged data set and the median data set, there was an equal number of correct and incorrect FE recognitions (eight wrong recognitions for FE).

Table 5
PNN confusion matrix

	Predicted data / Actual data	Achieved ECTS	Full ECTS	No ECTS
Initial data	Achieved ECTS (AE)	0	2	0
	Full ECTS (FE)	0	55	0
	No ECTS (NE)	0	4	1
Average	Achieved ECTS (AE)	0	4	0
	Full ECTS (FE)	0	54	0
	No ECTS (NE)	0	4	0
Median	Achieved ECTS (AE)	0	4	0
	Full ECTS (FE)	0	54	0
	No ECTS (NE)	0	4	0
Geometric mean	Achieved ECTS (AE)	0	2	0
	Full ECTS (FE)	0	55	0
	No ECTS (NE)	0	5	0

4.2. Accuracy statistics

Accuracy is a commonly used performance metric in machine learning to evaluate the quality of predictions, by using confusion matrix values: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). The accuracy score is calculated as a quotient of the

number of correct predictions and the total number of predictions.

The formula for accuracy calculation is given below [20]:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Overall, RF exhibited the highest precision (0,935), while on the other hand, PNN had the lowest accuracy of 0,903.

The results are presented below individually for each algorithm.

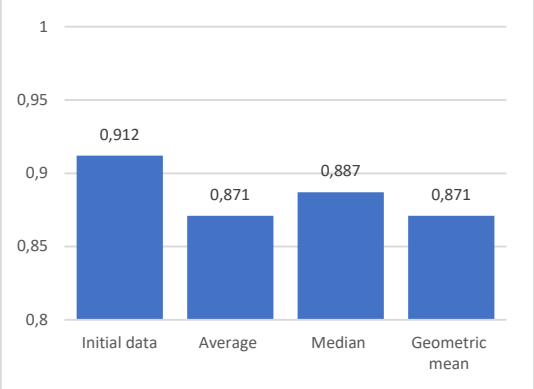

Figure 1: Accuracy KNN

Figure 1 shows the accuracy results for KNN. The initial data set generated the highest accurate of 0,912.

Data sets that use averaged and geometric mean values obtained equally the lowest accuracy of 0,871 (difference about 4,5%). The data set with median values obtained an accuracy of 0,887.

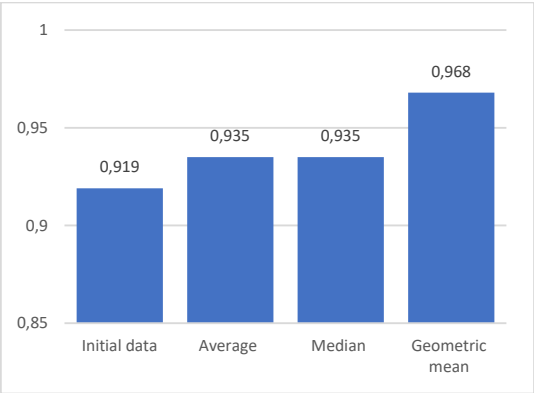

Figure 2: Accuracy NB

When speaking of NB, it was the only case where the initial data set generated the lowest accuracy (0,919) when compared to the other three data sets, as shown in Figure 2. Data sets with averaged and median values generated better accuracy for NB (0,935). In this case, the data set with geometric mean values obtained the best accuracy of 0,968, and that is 5,33% better than the initial data set using NB.

Figure 3 shows accuracy results for RF. The initial data set scored the highest accuracy of 0,935.

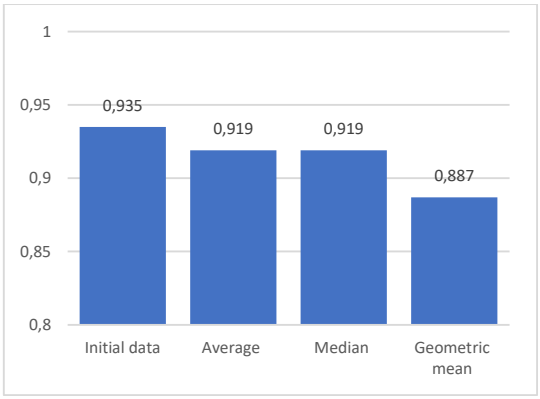

Figure 3: Accuracy RF

In the case of RF, data sets with averaged and median values generated equally the lowest accuracy of 0,919. The data set with geometric mean values generated the lowest accuracy of 0,887, which is 5,13% worse than the initial data set for RF.

PNN obtained the highest accuracy of 0,903 for the initial data set, as shown in Figure 4. The data sets with averaged and median values generated the worst accuracy of 0,871, which is 3,54% worse than the initial data set for PNN. The data set with geometric mean values obtained a score of 0,887.

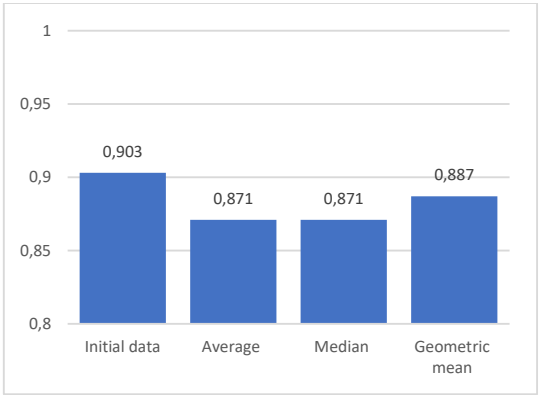

Figure 4: Accuracy PNN

As already mentioned, only in the NB case there is an improvement with regard to the initial

data set. KNN, RF and PNN got worse, but no correlation was observed between them.

4.3. F1 score

The F1 score is a popular performance metric in machine learning used to evaluate the quality of predictions in order to complement the accuracy measure. It is most often applied in unbalanced data sets. The F1 score is defined as the harmonic mean of precision and recall, and it is given by the following formulas [20]:

$$F1score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

where TP, FP and FN are values from the confusion matrix.

Figure 5 shows the F1 score results for all data sets using the KNN algorithm. Results for all data sets show high accuracy for FE, i.e. over 0,93 in all data sets using KNN. Other values are lower than 0,35.

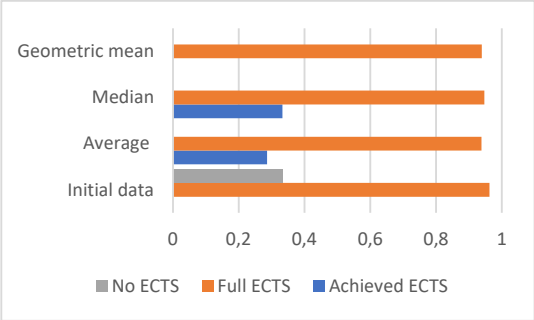

Figure 5: F1 score for KNN

In the median data set, KNN scores 0,333 for AE. In the average data set, KNN scores 0,286 for AE. In the initial data set, KNN scores 0,333 for NE.

F1 score results for all data sets for NB are shown in Figure 6. All scores are above 0,5. The averaged data set and the median data set scored a perfect 1. The geometric mean data set scored highest overall.

Furthermore, the geometric mean data set with NB achieved the highest accuracy when compared to all other given data sets, and it is the most accurate algorithm in all observed cases.

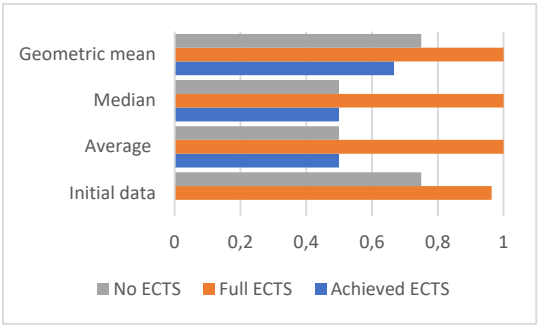

Figure 6: F1 score for NB

Figure 7 shows the F1 score results for all data sets using the RF algorithm. Results for all data sets show high accuracy for FE, whereas the averaged data set and the median data set scored a perfect 1.

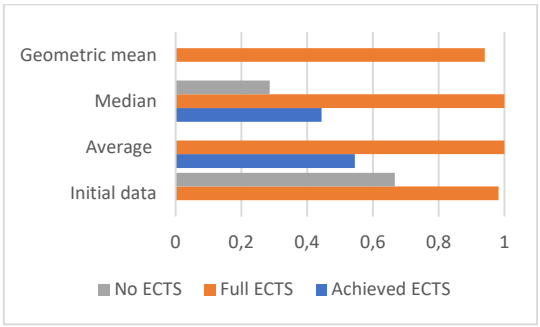

Figure 7: F1 score for RF

Other results range from 0,286 for the median data set score for NE to 0,667 for NE in the initial data set.

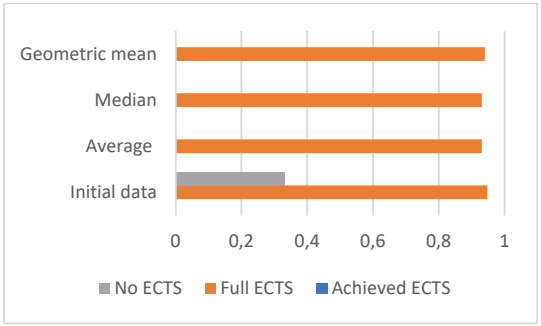

Figure 8: F1 score for PNN

F1 score results for all data sets for PNN are shown in Figure 8. Results for all data sets show high accuracy for FE, with a score over 0,93 in all sets using PNN. The initial data set scored 0,333 for NE.

5. Conclusion

The aim of this paper was to explore the impact of four different data sets containing missing values on the performance of selected ML algorithms. The four data sets were different with regard to the method used for filling the missing values – initial data set with missing values; missing values replaced by arithmetic mean values; missing values replaced by median values; and missing values replaced by geometric mean values.

Four ML algorithms were used in this research to test the accuracy of predicted ECTS scores (full ECTS points achieved, partial ECTS points achieved, no ECTS credits achieved): KNN, NB, RF and PNN.

All ML algorithms showed high accuracy ranging from 0,903 to 0,935 for the initial data set. The initial data set was lacking some data, therefore different types of averaged values were used in place of missing entries.

When comparing accuracy results of the initial data set with the remaining data sets for KNN, RF and PNN, there is a drop in accuracy of up to 5,13%, as it was in the case for the geometric mean data set used by RF. The NB showed an improvement in accuracy in all three data sets when compared to the initial data set – up to 5,33% in the case of the geometric mean data set.

Also, NB showed high values of F1 score, which further confirms that NB was positively impacted by all of the three data sets that replaced the missing data. It is certainly worth mentioning that the geometric mean data set achieved the highest accuracy for NB, and this is generally the highest accuracy for all observed cases in this work (0,968).

In conclusion, even though the data set was relatively small, the impact of the various data sets on the results achieved by the analyzed ML algorithms was noticeable.

For future work, the authors plan to experiment with other methods for replacing missing data, and to compare the effectiveness of various ML algorithms.

6. References

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

[2] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, O. Tabona, A survey on missing data in machine learning, Journal of Big Data, 8, article number: 140 (2021). doi:10.1186/s40537-021-00516-9.

[3] C. Wang, N. Shakhovska, A. Sachenko, M. Komar, A New Approach for Missing Data Imputation in Big Data Interface, Information Technology and Control, 49(4), (2020) 541–555. doi:10.5755/j01.itc.49.4.27386.

[4] B. Albreiki, N. Zaki, H. Alashwal, A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques, Education Sciences, 11(9), (2021). doi:10.3390/educsci11090552.

[5] B. Chakraborty, K. Chakma, A. Mukherjee, A density-based clustering algorithm and experiments on student dataset with noises using Rough set theory, in: Proceedings of the 2nd IEEE international conference on engineering and technology (ICETECH), 2016, pp. 431–436. doi:10.1109/ICETECH.2016.7569290 .

[6] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, G. Van Erven, Educational data mining : Predictive analysis of academic performance of public school students in the capital of Brazil, Journal of Business Research, 94, (2018) 335–343. doi:10.1016/j.jbusres.2018.02.012.

[7] R. Asif, A. Merceron, S. A. Ali, N. G. Haider, Analyzing undergraduate students' performance using educational data mining, Computers and Education, 113, (2017) 177–194. doi:10.1016/j.compedu.2017.05.007.

[8] M. Yağcı, Educational data mining: prediction of students' academic performance using machine learning algorithms, Smart Learning Environments, 9(11), (2022). doi:10.1186/s40561-022-00192-z.

[9] A. M. Nazif, A. A. Hesham Sedky, O. M. Badawy, MOOC's Student Results Classification by Comparing PNN and other Classifiers with Features Selection, in: Proceedings of the 21st International Arab Conference on Information Technology (ACIT), 2020, pp. 1–9, doi:10.1109/ACIT50332.2020.9300123.

[10] S. Khan, A. Latiful Haque, SICE: an improved missing data imputation technique,

Journal of Big Data, 7, article number: 37 (2020). doi:10.1186/s40537-020-00313-w.
[11] C. M. Salgado, C. Azevedo, H. Proença, S. M. Vieira, Missing Data, in: MIT Critical Data (Eds.), Secondary Analysis of Electronic Health Records, Springer, Cham, 2016. doi: 10.1007/978-3-319-43742-2_13.
[12] M. B. Mohammed, H. S. Zulkafli, M. B. Adam, N. Ali, I. A. Baba, Comparison of five imputation methods in handling missing data in a continuous frequency table, in: Proceedings of the AIP Conference Proceedings, 2021, 2355(1). doi:10.1063/5.0053286.
[13] K. M. Ting, Confusion Matrix, in: C. Sammut, G. I. Webb (Eds.), Encyclopedia of Machine Learning, Springer, doi:10.1007/978-0-387-30164-8_157.
[14] M. S. Naser, A. H. Alavi, Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences, Architecture, Structures and Construction (2021). doi:10.1007/s44150-021-00015-8.
[15] K. Taunk, S. De, S. Verma, A. Swetapadma, A Brief Review of Nearest Neighbor Algorithm for Learning and Classification, in: Proceedings of the International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1225–1260. doi:10.1109/ICCS45141.2019.9065747.
[16] O. Harrison, Machine Learning Basics with the K-Nearest Neighbors Algorithm, 2018. URL: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.
[17] D. Berrar, Bayes' theorem and naive Bayes classifier, Encyclopedia of Bioinformatics and Computational Biology, 1, (2019). doi: 10.1016/B978-0-12-809633-8.20473-1.
[18] A. Kovač, I. Dunđer, S. Seljan, An overview of machine learning algorithms for detecting phishing attacks on electronic messaging services, in: Proceedings of the 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), 2022, pp. 954–961, doi: 10.23919/MIPRO55190.2022.9803517.
[19] B. Mohebali, A. Tahmassebi, A. Meyer-Baese, A. H. Gandomi, Probabilistic neural networks, in: P. Samui, S. Chakraborty, D. T. Bui, R. C. Deo (Eds.), Handbook of Probabilistic Models, Elsevier, Amsterdam, 2020, pp. 347–367.
[20] S. Ghoneim, Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on?, 2019. URL: https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124.