About the Effects of Data Imputation Techniques on ML Uncertainty

Cinzia Cappiello[1], Federico Cerutti[2], Camilla Sancricca[1] and Riccardo Zanelli[1]

[1]Politecnico di Milano, Milan, Italy
[2]University of Brescia, Brescia, Italy

Abstract

The data-driven culture is based on the importance of data analysis in supporting decision-making. In particular, machine learning technologies and tools are evolving quickly and becoming increasingly popular as an effective means to gain insights from raw data. However, it should be considered that Machine Learning (ML) models often generate uncertain results due mainly to their imperfect and statistical nature. In this paper, we focus on the fact that data preparation techniques can introduce additional uncertainty. Errors, missing values, and inconsistencies are frequently addressed using techniques that correct data using estimates and thus add further uncertainty. Focusing on the specific problem of incomplete data, this paper *(i)* investigates the effect of imputation techniques on the results' uncertainty, and *(ii)* identifies the techniques that minimize such an issue.

Keywords

Data Quality, Uncertainty, Data Imputation

1. Introduction

In the modern era of the data-driven culture, data analysis is critical in providing useful information to support companies' decisions. In particular, Machine Learning (ML) models help users effectively gain insights from raw data. However, dealing with ML requires managing uncertainty.

There are many sources of uncertainty in an ML-based analysis. Still, the main one concerns the imperfect nature of any models developed considering high variance data or samples. More formally, we need to distinguish between (at least) two different sources of uncertainty: *aleatoric*, and *epistemic* uncertainty [1, 2, 3]. Aleatoric uncertainty refers to the variability in the outcome of an experiment, which is due to inherently random effects (*e.g.*, flipping a fair coin): no additional source of information but Laplace's daemon — *i.e.*, "An intelligence that, at a given instant, could comprehend all the forces by which nature is animated and the respective situation of the beings that make it up" [4, p.2] — can reduce such variability. Data Quality plays a crucial role in managing this uncertainty: reliable and consistent data helps identify and quantify the inherent variability and randomness in the data itself. This allows for more accurate modeling and prediction of outcomes, helping decision-makers assess and manage the associated risks.

Epistemic uncertainty refers to the agent's epistemic state using the model, hence its lack of knowledge that — in principle — can be reduced based on additional data samples. One example of the impact of epistemic uncertainty can be seen in climate change research. In its assessment reports, the Intergovernmental Panel on Climate Change (IPCC) [5] explicitly acknowledges and quantifies uncertainties associated with climate projections. These uncertainties arise from factors such as limited historical data, incomplete understanding of climate processes, and modelling assumptions. By considering epistemic uncertainty, researchers and policymakers gain a more realistic understanding of potential outcomes, enabling them to make informed decisions and develop appropriate mitigation strategies. In this paper, we focus on this type of uncertainty.

We must also consider that an ML model's performance strictly depends on the quality of input data. Data Quality (DQ) is often defined as "fitness for use," i.e., the ability of a data collection to meet user requirements [6]. It might be affected by several aspects, such as syntactic or semantic errors, inconsistencies, or missing values. These issues can be addressed and/or mitigated by using data preparation techniques. Such techniques, on the one hand, improve the data quality level but, on the other hand, can contribute to increasing epistemic uncertainty. To validate such a statement, we started considering the case in which a data set contains missing values. One way to address such an issue is to fill in the missing data in some form, and once the data are complete, feed them to the model. The available imputation methods are vari-

Joint Workshops at 49th International Conference on Very Large Data Bases (VLDBW'23) — the 12th International Workshop on Quality in Databases (QDB'23), August 28 - September 1, 2023, Vancouver, Canada

cinzia.cappiello@polimi.it (C. Cappiello); federico.cerutti@unibs.it (F. Cerutti); camilla.sancricca@polimi.it (C. Sancricca); riccardo.zanelli@mail.polimi.it (R. Zanelli)
0000-0001-6062-5174 (C. Cappiello); 0000-0003-0755-0358 (F. Cerutti); 0000-0002-3820-7870 (C. Sancricca)

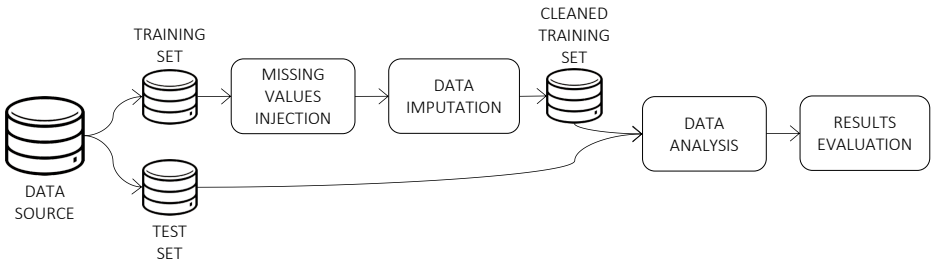

Figure 1: Pipeline of the experiments

ous: they go from traditional techniques in which null values are substituted with statistical information (*e.g.*, mean, median, mode) to more complex processes based on ML (*e.g.*, clustering and distance-based algorithms [7]). All these methods are imputing estimates, and therefore they add epistemic uncertainty.

This work investigates *(i)* the effect of this portion of uncertainty introduced by the data preparation process on the data analysis results and *(ii)* if the goal of mitigating uncertainty can be exploited to find the best preparation action within a specific context (*i.e.*, data and ML model characteristics).

The paper is organized as follows: Section 2 explores similar literature contributions and highlights the novel aspects of the presented paper. Section 3 describes the method that we used to investigate the impact of data preparation on the uncertainty of ML results; Section 4 presents the conducted experiments and discuss the obtained results, while Section 5 concludes the paper and presents future work.

2. Related Work

The problem of missing data has been increasingly spread in a variety of domains. For this reason, a lot of research contributions aim to define methods for efficiently performing data imputation and replacing the missing data with values that are as accurate as possible [7, 8].

Several papers propose implementing accurate and efficient data imputation methods by exploiting ML techniques. For example, the method presented in [9] proposes a novel k-Nearest-Neighbors (kNN) imputation method that iteratively imputes missing data selecting the kNN via calculating the Gray distance, *i.e.*, a technique used in the Gray system theory, rather than traditional distance metrics. Such a distance metric can deal with both numerical and categorical attributes.

Other methods [10, 11] make use of neural networks. The work presented in [10] builds a deep latent variable model to impute missing-at-random data. This model is based on autoencoders and has been proven to provide accurate single imputations, being competitive with other state-of-the-art methods. Finally, the method shown in [11] adapts the Generative Adversarial Networks (GAN) framework to impute the missing data. This method has been tested on various datasets and outperforms some state-of-the-art methods.

Some of the data imputation methods described above were also considered in the experiments of this paper.

Moreover, some studies have tried to put together several imputation methods. For example, the work in [12] proposes an adaptive iterative imputation framework that automatically finds, for each dataset column, the best data imputation model and configures it with the appropriate hyperparameters. The best single-column imputation method is computed by trying several methods until an imputation-stopping criterion, based on the incremental change in imputation quality, is met.

Within the same domain, several contributions have conducted comparisons between the different imputation methods present nowadays in the literature. For example, [13] depicts a comprehensive benchmark on six different methods involving standard, classical ML, and novel deep learning approaches to perform data imputation. The experiments were done on a huge set of real-world datasets, including three missingness patterns, *i.e.*, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In [14], the authors show a comparison between multiple existing data imputation techniques that are based on deep learning; moreover, they propose a set of improvements for each analyzed method.

In these cases, the data imputation methods are evaluated on their imputation quality, without considering the uncertainty that they can introduce in the performance of a ML model that will be executed on them.

A recent approach [15] focuses on studying the impact of data preparation on the ML model performance. This study investigates the impact of data cleaning actions on ML classification models. The authors consider different data cleaning methods for correcting outliers, duplicates, inconsistencies, mislabels, and missing val-

ues. The goal was to assign, for a specific setting (error type, data cleaning action, and ML application), a P (positive), N (negative), or S (insignificant) flag indicating the impact of the data cleaning on the ML performance.

Also in this case, the impact of data cleaning methods is evaluated on the basis of the final ML model performance, without considering the ML uncertainty.

Some contributions focused on creating their data imputation methods for particular contexts and then tried to validate them from the point of view of the introduced uncertainty [16, 17]. In particular, the work proposed in [16] aims to provide a tool to predict hospital readmission among Heart Failure patients and develops a new methodological framework to address the missing data using a Gaussian process latent variable model. In contrast, the method shown in [17] focuses on *well logs*, commonly used in geoscience, and proposes an approach to customize the hyper-parameters of a random forest model to predict the missing values.

However, none of the cited works considered using uncertainty to select the best data imputation method to apply in a given analysis context. Our work aims to explore this open issue.

Finally, a paper that implements a similar approach *w.r.t.* our method is [18]. However, the authors focus on a totally different purpose: they systematically inject errors, *e.g.*, missing values and encoding errors, into the input data to estimate the prediction quality of a ML model. Their goal was to estimate the output quality of ML models on unseen, unlabeled serving data, in order to automate the validation of black boxes.

3. Measuring the Impact of Data Preparation on the Decision Uncertainty

This section presents the pipeline — illustrated in Figure 1 — implemented to investigate the impact of data preparation, whose application introduces approximate data, on the uncertainty of ML outcomes.

In this work, we focus on the Completeness DQ dimension *i.e.*, the degree to which a given data collection includes the data describing the corresponding set of real-world objects [6]. It is affected by missing values and can be improved by applying data imputation techniques. Note that the considered input dataset is free of DQ problems. For this reason, we have to inject missing values to perform the data imputation techniques.

The Experiment Pipeline As Figure 1 depicts, the input of the pipeline is a *Data Source*, which is split into two datasets: the *Training Set* and the *Test Set*. Each dataset is the input of the *Data Analysis* phase. This splitting is done before injecting the DQ errors: in this way, dirty instances of the *Training Set* are created, an ML model is trained on them, and it is finally evaluated on the same original instance of the *Test Set*.

The *Missing Values Injection* phase generates five instances of the *Training Set* at different levels of quality by injecting a different percentage of missing values (from 50% to 10%, with a decreasing step of 10%) uniformly. The *targeted class* is excluded from the injection and is not corrupted.

Following this procedure, the injected missing values are Missing Completely At Random (MCAR), *i.e.*, the probability of a data point being missed is independent of the observed and unobserved data. An injection above 50% of DQ errors has not been performed in our experiments since the variance of the model performance, trained with so many mistakes, was too high and was no longer considered reliable.

The obtained five dirty datasets are the input of the *Data Imputation* phase, in which a data imputation technique is applied to fill the missing values. In this phase, several imputation methods have been compared.

The five cleaned datasets obtained as the output of the *Data Imputation* are fed to the *Data Analysis* phase, where an ML model is trained on them. The resulting five ML models are finally evaluated on the same *Test Set*, computing their prediction performance and related epistemic uncertainty. Two sets of scores are the output of this phase: five scores (each one related to the ML model executed on one of the five cleaned datasets) related to the model performance and another set of scores for the uncertainty. The method is repeated for all the selected data source/ML algorithm/data imputation method combinations.

The Pipeline with Feature Selection The same pipeline is also performed with an additional step of *feature selection*. In this case, the input dataset is first analyzed through a *feature selection* method. The output is a subset of the original dataset that keeps only the four most relevant features. The resulting dataset is the input *Data Source* of this set of experiments.

4. Experiments & Results

This section describes the setup used to run the experiments and the results obtained following the method proposed in Section 3.

4.1. Experimental Setup

Different data sources have been selected to run the experiments: Boston,[1] Wine,[2] California,[3] House,[4] Concrete.[5] Table 1 lists their main characteristics. All these datasets have a numeric target label, and regression ML models were adopted to perform the *Data Analysis* phase (see Section 3). For this reason, the CatBoost algorithm from the *catboost* Python library [19] and the Gaussian Process regressor from the *scikit-learn* Python library [20] have been selected as ML analysis algorithms. In addition, the Boruta [21] method for *feature selection* has been adopted. It is an ML-based method that evaluates each feature's importance in a dataset and returns the most relevant ones.

In order to include a diversified set of data imputation techniques, we consider seven types of them, divided into four macro-categories. For each category, we select one or more representative methods, even though it is known that some are less effective than others. The considered methods are the following:

(1) **Single-column imputation with aggregated values** computes an aggregated value like the mean, the median, or the most frequent to substitute the missing ones.

ML-based imputation exploits ML algorithms, such as: *(2)* **k-Nearest Neighbours (KNN)** [9] estimates each sample's missing value with the mean value of its nearest neighbours; *(3)* **Generative Adversarial Imputation Nets (GAIN)** [11] uses generative adversarial networks (GANs) for estimating missing values by training a GAN, which consists of two neural networks: a generator network, which generates the missing data, and a discriminator one, to distinguish between the real data and the ones that were just generated; *(4)* **MIWAE** [10] uses an autoencoder, a neural network trained to encode the observed data into a lower-dimensional space. This allows the autoencoder to learn a compact representation of the data, which can be used to predict the missing values.

Multiple imputation creates copies of the original data and estimates the missing values through an iterative process. We consider the *(5)* **Multiple Imputation by Chained Equations (MICE)** [22] technique: *(i)* random imputation is applied to each missing column; *(ii)* the missing values are set back one feature at a time; *(iii)* an ML model is fitted to impute the values using the rest of data as training set; *(iv)* the training set is updated with the predicted column. For the experiments, the selected ML model is *KNNRegressor* from *scikit-learn* Python library [20].

Statistics-based imputation considers **Matrix Factorization (MF)** techniques. We select two of them: *(6)* **basic MF** and *(7)* **Singular Value Decomposition (SVD)** [8]. These processes assume that input data are noisy observations produced by a linear combination of a small set of principal components. They estimate the missing data by splitting them into two or more low-dimensional matrices and reconstructing the original one based on a linear combination.

To evaluate the accuracy and the uncertainty of the results, we used the following evaluation metrics. The *Root Mean Squared Error* (RMSE), *i.e.*, a measure of the average difference between the predicted and actual values, has been used to evaluate the prediction accuracy.

Moreover, one common approach for estimating the (epistemic) uncertainty of ML models is to use the *standard deviation* of the algorithm prediction, *i.e.*, a measure of the variation of the predicted values with respect to their average. A high standard deviation indicates that the predicted values are more variable and, therefore, less reliable.

The standard deviation of the results was estimated directly by the CatBoost and Gaussian Process algorithms. For example, for CatBoost, the value of the uncertainty was extracted from the model evaluation function, which, in this case, was set to *RMSEWithUncertainty* — an evaluation metric provided by the *catboost* Python library [19].

The method presented in Section 3 has been executed 16 times with different random seeds for each combination of data source/Ml algorithm/data imputation method.

[1] https://www.kaggle.com/datasets/avish5787/boston-data-set (on 29th May 2023).
[2] https://www.kaggle.com/datasets/shelvigarg/wine-quality-dataset (on 29th May 2023).
[3] https://www.kaggle.com/datasets/dhirajnirne/california-housing-data (on 29th May 2023).
[4] https://www.kaggle.com/datasets/lespin/house-prices-dataset (on 29th May 2023).
[5] https://www.kaggle.com/datasets/rithikkotha/concrete-dataset (on 29th May 2023).

	Tuples	*Features*	*Numeric f.*	*Categoric f.*	*Target class*
boston	506	13	13	0	numeric
california	1,000	9	8	1	numeric
house	1,460	80	34	46	numeric
str	1,030	8	8	0	numeric
wine	1599	12	11	1	numeric

Table 1
Data sources profile

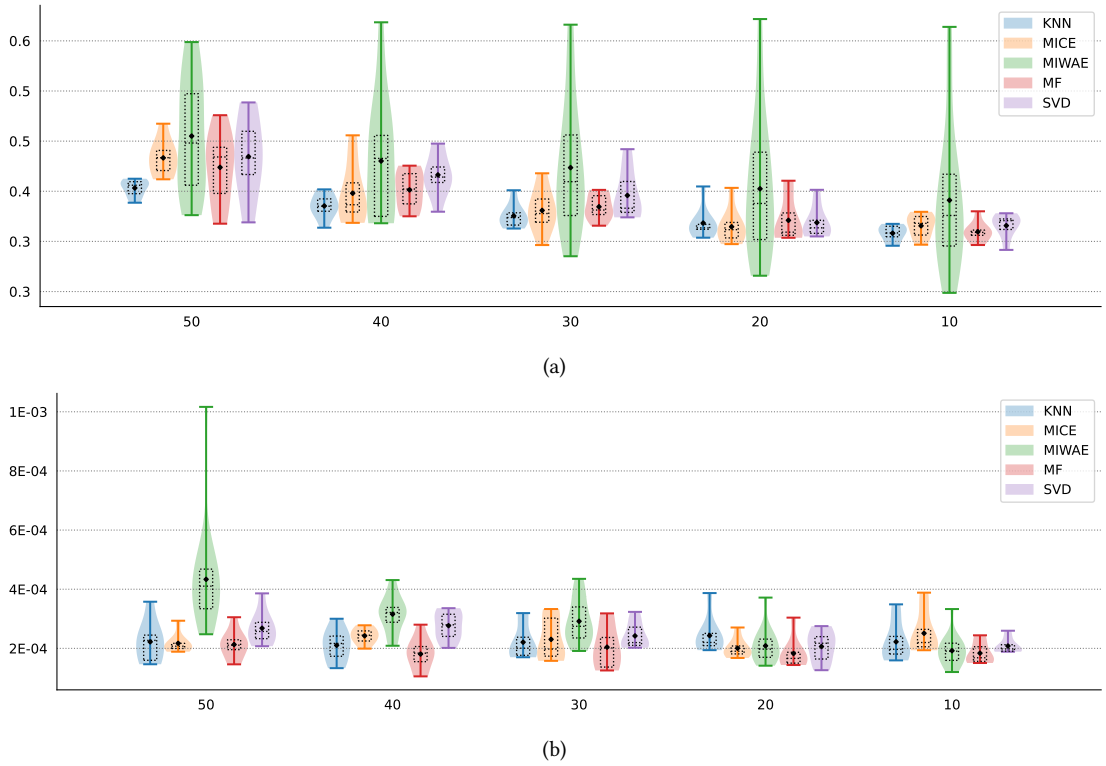

Figure 2: Distributions of RMSE (a) and Standard Deviation (b) of using KNN, MIWAE, MF, MICE, SVD on the *house* dataset when analyzing it with CatBoost, varying the completeness.

4.2. Results Evaluation

This section shows the preliminary results we obtained applying the method described in Section 3. Experiments have been conducted for the data sources, ML algorithms, and data imputation techniques listed in Section 4.1.

From the experiments' results, the role of uncertainty introduced by data preparation arises: it can be used as a support in identifying the best data preparation method to apply in a specific *analysis context*, *i.e.*, a combination of the data source and the ML algorithm selected for its analysis. When applying two data imputation methods leads to equivalent analysis results (in terms of performance), the best one can be identified by evaluating their uncertainty.

Figure 2 depicts an example of the aggregated results obtained for the combination CatBoost/*house* dataset. In particular, Figure 2a plots the model performance (RMSE) and Figure 2b the uncertainty distribution for the five imputation methods that give the best analysis results, varying the completeness. The y-axes represent values assumed by the RMSE and uncertainty, respectively, while the x-axis pictures the Completeness level.

From visual inspection of Figure 2, it emerges that applying k-Nearest Neighbours (KNN), Multiple imputation (MICE), Matrix Factorization (MF), and Singular Value Decomposition (SVD) yields ML model performance (RMSEs) that are very similar to each other, and it becomes difficult to determine which one is better. However, by analyzing the uncertainty, one can argue that MF outperforms the others since — on average — it leads to lower values.

Moreover, we rank the data imputation methods based on the analysis performance and the uncertainty they introduced. For each *analysis context*, we compute the ML model performance and related uncertainty for the selected ML algorithms using the original (cleaned) dataset, and we use this value as a baseline. Then, for each combination of dataset/ML algorithm/data imputation technique, we *(i)* run our method (see Section 3) several times, *(ii)* aggregate the results by the median, and *(iii)* compute the median distance between the five extracted scores (both for RMSE and uncertainty) and the baseline.

Data imputation methods were sorted in ascending order of their median distance from the baseline to extract the rankings. The closer the score is to the original values, the more reliable the data imputation method is. Table

		Rankings without feature selection				*Rankings with feature selection*			
CatBoost		RMSE-distance		Uncertainty-distance		RMSE-distance		Uncertainty-distance	
boston	*1*	MF	0.11884	KNN	5e-05	KNN	0.13111	KNN	7e-05
	2	SVD	0.13992	MICE	6e-05	MICE	0.13252	MF	0.00011
	3	KNN	0.15831	MF	6e-05	MEDIAN	0.16652	MIRACLE	0.00011
	4	MEDIAN	0.16375	SVD	7e-05	GAIN	0.19912	MICE	0.00014
	5	MICE	0.16489	MIRACLE	9e-05	MEAN	0.20864	SVD	0.00018
california	*1*	MF	0.18043	GAIN	2e-05	MF	0.24315	KNN	2e-05
	2	SVD	0.21464	MF	2e-05	GAIN	0.24458	MICE	2e-05
	3	MICE	0.22342	SVD	2e-05	MIRACLE	0.25869	MF	3e-05
	4	KNN	0.23966	MIRACLE	3e-05	MIWAE	0.26749	SVD	5e-05
	5	MEAN	0.24033	KNN	5e-05	KNN	0.27179	GAIN	8e-05
house	*1*	KNN	0.139	MF	1e-05	MICE	0.17602	KNN	1e-05
	2	MICE	0.14301	MIRACLE	2e-05	KNN	0.17849	MIRACLE	1e-05
	3	MF	0.14648	KNN	4e-05	SVD	0.18807	MICE	3e-05
	4	MIWAE	0.14697	MICE	4e-05	MF	0.20437	MF	3e-05
	5	SVD	0.1483	SVD	5e-05	MEDIAN	0.20977	SVD	4e-05
wine	*1*	GAIN	0.09704	MEAN	0.00011	MICE	0.17407	MEDIAN	0.00019
	2	MF	0.16057	MEDIAN	0.00013	MEAN	0.1839	GAIN	0.00023
	3	MIWAE	0.16117	MIWAE	0.00013	MEDIAN *(*)*	0.18537	MIRACLE	0.00023
	4	SVD	0.16877	MF	0.00015	MF	0.18547	MEAN	0.00024
	5	MICE	0.17426	SVD	0.00015	SVD	0.20672	MIWAE	0.00024
concrete	*1*	MF	0.15369	KNN	3e-05	MEAN	0.20749	SVD	7e-05
	2	MICE	0.15676	MF	4e-05	MF	0.21162	MICE	8e-05
	3	MIWAE *(†)*	0.15934	SVD	9e-05	KNN	0.21179	KNN	9e-05
	4	SVD	0.17455	MICE	0.00012	MICE	0.21313	MEDIAN	9e-05
	5	KNN	0.17532	GAIN	0.00016	MIWAE	0.2134	MIWAE	9e-05
Gaussian Process									
boston	*1*	SVD	0.17893	KNN	0.45759	KNN	0.15131	KNN	0.39494
	2	GAIN	0.19831	GAIN	0.47994	MICE	0.17001	MICE	0.46441
	3	MF	0.20524	MF	0.4841	SVD *(‡)*	0.18452	MF	0.48834
	4	MIWAE	0.23681	MICE	0.50455	GAIN	0.1909	GAIN	0.52218
	5	KNN	0.24607	SVD	0.57109	MEDIAN	0.19812	SVD	0.58669
california	*1*	SVD	0.19445	KNN	0.52332	MF	0.20394	MICE	0.36205
	2	MF	0.2031	MICE	0.5583	MIWAE	0.26859	KNN	0.37373
	3	MEAN	0.23877	GAIN	0.56757	MEAN	0.26934	MF	0.45056
	4	MICE	0.24154	MF	0.57137	MICE	0.27192	SVD	0.46501
	5	KNN	0.24468	SVD	0.67285	GAIN	0.29764	GAIN	0.49432
house	*1*	SVD	0.16725	MIRACLE	0.20324	KNN	0.18913	KNN	0.24427
	2	MICE	0.17225	MF	0.25711	MICE	0.20438	MF	0.25389
	3	KNN	0.17235	KNN	0.30113	MF *(#)*	0.20585	MICE	0.30185
	4	MF	0.17635	SVD	0.32236	MEAN	0.20631	SVD	0.44806
	5	GAIN	0.19969	MICE	0.32672	SVD	0.20789	GAIN	0.68732
wine	*1*	SVD	0.0742	MIRACLE	0.30437	KNN	0.17269	MICE	0.83604
	2	MICE	0.10617	MF	0.67813	SVD	0.1749	KNN	0.89423
	3	KNN	0.10943	MICE	0.80923	MEAN	0.18762	MF	0.98518
	4	MF	0.11109	KNN	0.80974	MICE	0.19375	SVD	1.04923
	5	GAIN	0.12256	SVD	0.87162	MF	0.20993	GAIN	1.20151
concrete	*1*	MF	0.19415	MF	0.67996	MICE	0.23003	MICE	0.47709
	2	SVD	0.20085	KNN	0.68979	KNN	0.25437	MEDIAN	0.48248
	3	MICE	0.20471	MICE	0.74398	MEDIAN	0.25936	KNN	0.55295
	4	KNN	0.20947	SVD	0.8873	MF	0.2618	MF	0.55554
	5	MEAN	0.22111	GAIN	1.04528	SVD	0.27333	SVD	0.61117

Table 2
Data imputation methods rankings based on the RMSE/uncertainty median distance from the baseline (*i.e.*, the RMSE/uncertainty computed with the original dataset).

2 lists the extracted rankings and their related distances. We performed a Kruskal-Wallis [23] nonparametric test to determine if there are statistically significant differences between the methods in each ranking. White cells in Table 2 are statistically significant ($p < 0.01$) results according to the Kruskal-Wallis test. Among the non-statistically significant ones—in grey—the following couples are statistically significant ($p < 0.01$) according to a pairwise analysis performed using the Mann-Whitney test [24]: *(*)* MICE ≠ SVD; *(†)* MICE ≠ {MIWAE, KNN}; *(‡)* KNN ≠ {SVD, MEDIAN}; *(#)* KNN ≠ {MEAN, SVD}.

From Table 2, we can appreciate, again, that uncertainty can be used to discriminate between different imputation methods with absolute values of distance

from the baseline very close considering the ML model performance achieved. From these tables, it is evident that whenever two imputation methods have median distances from the baseline that are very close, the uncertainty they introduce is always different and can be sorted accordingly.

For example, for the combination of CatBoost algorithm/*concrete* dataset, the first two imputation methods are very close to each other, *i.e.*, MF and MICE; however, the uncertainty introduced by MF is much smaller than the other one. We can conclude that the first method is better than the second one. The above statement applies to all the tested *analysis contexts*.

We also aggregate the rankings results in the following manner: *(i)* aggregating all results together; *(ii)* aggregating, for each dataset, results obtained applying the two algorithms with and without feature selection; *(iii)* aggregating, for the 4 combinations of CatBoost-Gaussian Process/with-without feature selection all dataset-related results. For each aggregation, we sum the median distances reported in Table 2 and sort the imputation methods in ascending order of that sum, creating aggregated rankings.

From the aggregated results, we can state that:

(i) Considering all the results together, the best-4 methods turned out to be MICE, MF, SVD, and KNN. Moreover, their aggregated distance values are very close to each other both for RMSE and uncertainty. SVD imputation has slightly higher uncertainty than the others.

(ii) The best-4 methods found in *(i)*, in general, appear in the first four positions of the rankings obtained for each dataset aggregation. There may be variations in the third and fourth positions of the aggregated rankings, where other imputation methods can appear. However, the uncertainty of the latter methods is always higher than the best-4 methods.

(iii) The best-4 methods are coherent for all algorithms/with-without feature selection aggregations. However, the position of these methods changes based on the considered combination. We can notice that CatBoost and Gaussian Process algorithms have very similar RMSE-related rankings *(1)* with feature selection, in which the first 3 positions are the same, and *(2)* without feature selection.

It is possible to draw some conclusions from the conducted experiments. First of all, there is no absolute "best" imputation method that fits all situations: identifying the imputation method to prefer depends on the *analysis context*. However, it is possible to observe that the imputation methods that outperform the others are KNN, MICE, and MF.

From a more general perspective, it is also possible to state that neural network-based imputation techniques, in some cases, are the best ones. However, they have very high uncertainty and are less reliable. This is especially the case of data sources with low dimensionality, *i.e.*, the number of tuples and features, as neural networks need much more data to build a reliable ML model.

As regards the single-column imputation with aggregated values techniques, it is possible to highlight that the uncertainty introduced by these methods is higher for lower completeness values. This happens since substituting an aggregated value introduces a higher approximation concerning the other methods.

5. Conclusions and Future Work

The paper presents a set of experiments to evaluate the effects of data imputation techniques on ML-based analysis uncertainty. The obtained results highlight that besides performance, uncertainty can be an additional metric to consider for defining the data preparation method to prefer. Future work will focus on extending the experiments considering the other DQ dimensions. Our vision is to exploit these experimental results and experience already gained in similar contexts to design a self-service environment that supports data scientists in finding and recommending data preparation techniques to maximize the results' accuracy while minimizing uncertainty.

Acknowledgments

This research was supported by EU Horizon Framework grant agreement 101069543 (CS-AWARE-NEXT).

References

[1] S. C. Hora, Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management, Reliability Engineering & System Safety 54 (1996) 217–223. Citation Key: HORA1996217.

[2] E. Hüllermeier, W. Waegeman, Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods, Machine Learning 110 (2021) 457–506. doi:10.1007/s10994-021-05946-3. arXiv:1910.09457.

[3] F. Cerutti, L. M. Kaplan, M. Sensoy, Evidential reasoning and learning: a survey, in: L. D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, ijcai.org, 2022, pp.

5418–5425. URL: https://doi.org/10.24963/ijcai.2022/760. doi:10.24963/ijcai.2022/760.

[4] P. S. Laplace, A Philosophical Essay on Probabilities, Springer, 1825.

[5] M. D. Mastrandrea, K. J. Mach, G.-K. Plattner, O. Edenhofer, T. F. Stocker, C. B. Field, K. L. Ebi, P. R. Matschoss, The IPCC AR5 guidance note on consistent treatment of uncertainties: A common approach across the working groups, Climatic Change 108 (2011) 675. doi:10.1007/s10584-011-0178-6.

[6] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, J. Manag. Inf. Syst. 12 (1996) 5–33. URL: https://doi.org/10.1080/07421222.1996.11518099. doi:10.1080/07421222.1996.11518099.

[7] T. Thomas, E. Rajabi, A systematic review of machine learning-based missing value imputation techniques, Data Technol. Appl. 55 (2021) 558–585. URL: https://doi.org/10.1108/DTA-12-2020-0298. doi:10.1108/DTA-12-2020-0298.

[8] Y. Luo, Evaluating the state of the art in missing data imputation for clinical data, Briefings Bioinform. 23 (2022). URL: https://doi.org/10.1093/bib/bbab489. doi:10.1093/bib/bbab489.

[9] S. Zhang, Nearest neighbor selection for iteratively knn imputation, J. Syst. Softw. 85 (2012) 2541–2552. URL: https://doi.org/10.1016/j.jss.2012.05.073. doi:10.1016/j.jss.2012.05.073.

[10] P. Mattei, J. Frellsen, MIWAE: deep generative modelling and imputation of incomplete data sets, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 4413–4423. URL: http://proceedings.mlr.press/v97/mattei19a.html.

[11] J. Yoon, J. Jordon, M. van der Schaar, GAIN: missing data imputation using generative adversarial nets, in: J. G. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 5675–5684. URL: http://proceedings.mlr.press/v80/yoon18a.html.

[12] D. Jarrett, B. Cebere, T. Liu, A. Curth, M. van der Schaar, Hyperimpute: Generalized iterative imputation with automatic model selection, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, S. Sabato (Eds.), International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 9916–9937. URL: https://proceedings.mlr.press/v162/jarrett22a.html.

[13] S. Jäger, A. Allhorn, F. Bießmann, A benchmark for data imputation methods, Frontiers in big Data 4 (2021) 693674.

[14] R. D. Camino, C. A. Hammerschmidt, R. State, Improving missing data imputation with deep generative models, arXiv preprint arXiv:1902.10666 (2019).

[15] P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, C. Zhang, Cleanml: A study for evaluating the impact of data cleaning on ml classification tasks, in: 2021 IEEE 37th International Conference on Data Engineering (ICDE), IEEE, 2021, pp. 13–24.

[16] Z. Hu, D. Du, A new analytical framework for missing data imputation and classification with uncertainty: Missing data imputation and heart failure readmission prediction, PLOS ONE 15 (2020) 1–15. URL: https://doi.org/10.1371/journal.pone.0237724. doi:10.1371/journal.pone.0237724.

[17] R. Feng, D. Grana, N. Balling, Imputation of missing well log data by random forest and its uncertainty analysis, Comput. Geosci. 152 (2021) 104763. URL: https://doi.org/10.1016/j.cageo.2021.104763. doi:10.1016/j.cageo.2021.104763.

[18] S. Schelter, T. Rukat, F. Bießmann, Learning to validate the predictions of black box classifiers on unseen data, in: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 2020, pp. 1289–1299.

[19] A. V. Dorogush, V. Ershov, A. Gulin, Catboost: gradient boosting with categorical features support, CoRR abs/1810.11363 (2018). URL: http://arxiv.org/abs/1810.11363. arXiv:1810.11363.

[20] F. P. et al., Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830. URL: https://dl.acm.org/doi/10.5555/1953048.2078195. doi:10.5555/1953048.2078195.

[21] M. B. Kursa, A. Jankowski, W. R. Rudnicki, Boruta - A system for feature selection, Fundam. Informaticae 101 (2010) 271–285. URL: https://doi.org/10.3233/FI-2010-288. doi:10.3233/FI-2010-288.

[22] I. R. White, P. Royston, A. M. Wood, Multiple imputation using chained equations: Issues and guidance for practice, Statistics in Medicine 30 (2011) 377–399. doi:https://doi.org/10.1002/sim.4067.

[23] W. H. Kruskal, W. A. Wallis, Use of ranks in one-criterion variance analysis, Journal of the American statistical Association 47 (1952) 583–621.

[24] H. B. Mann, D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, The annals of mathematical statistics (1947) 50–60.