No Imputation without Representation

Oliver Urs Lenz [1] **Daniel Peralta** [2] **Chris Cornelis** [1]

Abstract

By filling in missing values in datasets, imputation allows these datasets to be used with algorithms that cannot handle missing values by themselves. However, missing values may in principle contribute useful information that is lost through imputation. The missing-indicator approach can be used in combination with imputation to instead represent this information as a part of the dataset. There are several theoretical considerations why missing-indicators may or may not be beneficial, but there has not been any large-scale practical experiment on real-life datasets to test this question for machine learning predictions. We perform this experiment for three imputation strategies and a range of different classification algorithms, on the basis of twenty real-life datasets. In a follow-up experiment, we determine attribute-specific missingness thresholds for each classifier above which missing-indicators are more likely than not to increase classification performance. And in a second follow-up experiment, we evaluate numerical imputation of one-hot encoded categorical attributes. We reach the following conclusions. Firstly, missing-indicators generally increase classification performance. Secondly, with missing-indicators, nearest neighbour and iterative imputation do not lead to better performance than simple mean/mode imputation. Thirdly, for decision trees, pruning is necessary to prevent overfitting. Fourthly, the thresholds above which missing-indicators are more likely than not to improve performance are lower for categorical attributes than for numerical attributes. Lastly, mean imputation of numerical attributes preserves some of the information from missing values. Consequently, when not using missing-indicators it can be advantageous to apply mean imputation to one-hot encoded categorical attributes instead of mode imputation.

1. Introduction

Missing values are a frequent issue in real-life datasets, and the subject of a large body of ongoing research. Some implementations of machine learning algorithms can handle missing values natively, requiring no further action by practitioners. But whenever this is not the case, a common general strategy is to replace the missing value with an estimated value: imputation. An advantage of imputation is that we obtain a complete dataset, to which we can apply any and all algorithms that make no special provision for missing values. However, missing values may be informative, and a disadvantage of imputation is that it removes this information.

The missing-indicator approach (Cohen, 1968) is an old proposal to represent and thereby preserve the information encoded by missing values. For every original attribute, it adds a new binary 'indicator' or 'dummy' attribute that takes a value of 1 if the value for the original attribute is missing, and 0 if not.[1] The missing-indicator approach is often presented as an alternative to imputation, but since it does not resolve the missing values in the original attributes, it can only be used in addition to, not instead of imputation.

It is an open question whether missing-indicators should be used for predictive tasks in machine learning (Sperrin et al., 2020). Both imputation and the missing-indicator approach originate in the statistical literature. While imputation strategies have been the subject of a rich body of research, the missing-indicator approach has not received a large amount of attention, and is often dismissed or disregarded in overviews of approaches towards missing values.

In the context of machine learning, the effect of missing-indicators can be framed as follows. On the one hand, the addition of missing-indicators results in a more complete, higher-dimensional representation of the data. On the other hand, their omission corresponds to a form of dimensionality reduction, which may increase the efficiency and effective-

---

[1] Department of Applied Mathematics, Computer Science and Statistics, Ghent University [2] Department of Information Technology, Ghent University. Correspondence to: Oliver Urs Lenz <oliver.lenz@ugent.be>.

Presented at the Data-centric Machine Learning Research (DMLR) Workshop at ICML 2023, Honolulu, 29 July 2023.

[1] Some authors use the opposite convention, letting the indicator express non-missingness.

ness of a dataset by eliminating redundancy.

To determine whether this trade-off is useful, a key question is to which extent missing values in a given dataset are informative. If they are not, the phrase “missing at random” (MAR) (Rubin, 1976) is used to indicate that the distribution of missing values is dependent on the known values, while the stricter phrase “missing completely at random” (MCAR) denotes values that are distributed truly randomly. In contrast, informative missing values are often denoted as “missing not at random” (MNAR).

For real-life datasets, unless we have specific knowledge about the process responsible for the missing values, we have to assume some degree of informativeness in principle.[2] However, it has been argued that in practice, the attributes of a dataset can be sufficiently redundant that one can get away with assuming its missing values are MAR (Schafer, 1997). But even if this is so, imputation may not always perform optimally, in which case missing-indicators may still prove useful.

A more subtle point is that even when missing values are informative, the information they encode need not be lost completely through imputation. This is particularly evident in the case of numerically encoded binary attributes, where imputation can represent missing values as a third, intermediary value. More generally, Le Morvan et al. (2021) have recently observed that almost all deterministic imputation functions map records with missing values to distinct manifolds in the attribute space that can in principle be identified by sufficiently powerful algorithms. Nevertheless, missing-indicators can potentially make this learning task easier.

In light of these conflicting theoretical arguments, the usefulness of missing-indicators for real-life machine learning problems is an interesting empirical question. However, previous experiments in this direction have been limited in scope and number. These limitations include the use of only one or a handful of datasets, the use of datasets from which values have been removed artificially, and not comparing the same imputation strategies with and without missing-indicators.

The purpose of the present paper is straightforward. On the basis of twenty real-life classification problems with naturally occurring missing values, we measure the performance of a range of popular classification algorithms, using three common types of imputation, with and without missing-indicators. This allows us to evaluate the effect of using missing-indicators, as well as the choice of imputation strategy.

Moreover, we conduct two follow-up experiments to gain a better understanding of when and why missing-indicators can be useful. In the first, we determine whether this is influenced by the type (categorical or numerical) and the amount of missing values of a given attribute. In the second follow-up experiment, we test the hypothesis that numerical imputation partially preserves the information from missing values.

In Section 2, we provide a brief overview of the existing literature on missing-indicators, including previous experimental evaluations. In Section 3, we describe our experimental setup. We report our results in Section 4 and conclude in Section 5.

2. Background

We start with a brief discussion of the origins and reception of the missing-indicator approach, as well as previous experimental evaluations of the use of missing-indicators in prediction tasks.

2.1. Origins and Reception

The missing-indicator approach originates in the literature on linear regression. It dates back to at least Cohen (1968), who pointed out that values in real-life datasets are typically not missing completely at random, and that the distribution of missing values may in particular depend on the values of the attribute that is to be predicted. He proposed that each attribute could be said to have two ‘aspects’, its value, and whether that value is present to begin with, which should be encoded with a pair of variables. For missing attribute values, the first of these variables was to be filled in with the mean of the known values, although other applications might call for different values. Cohen’s proposal was subsequently expanded by Cohen & Cohen (1975), but received only limited recognition in the following years (Kim & Curry, 1977; Stumpf, 1978; Chow, 1979; Hutcheson & Prather, 1981; Anderson et al., 1983; Orme & Reis, 1991).

Cohen’s proposal was subjected to a formal analysis by Jones (1996), who showed that, if one assumes that missing values are MAR, and the true linear regression model does not contain any terms related to missingness, it produces biased estimates of the regression coefficients (unless the sample covariance between independent variables is zero). However, these assumptions run directly counter to the position set out by Cohen & Cohen (1975) that a priori, the missingness of each attribute is a possible explanatory factor, that it is safer not to assume that missing values are distributed randomly, and that the usefulness of missing-indicators is ultimately an empirical question.

---

[2] This is acknowledged by authors working under the assumption of MAR, e.g. “When data are missing for reasons beyond the investigator’s control, one can never be certain whether MAR holds. The MAR hypothesis in such datasets cannot be formally tested unless the missing values, or at least a sample of them, are available from an external source.” (Schafer, 1997)

Allison (2001), motivated by Jones (1996) and working under the general assumption of MAR, dismissed missing-indicators as "clearly unacceptable", before conceding that they in fact produce optimal estimates when the missing value is not just missing, but cannot exist, such as the marital quality of an unmarried couple. However, this semantic distinction may not always be clear-cut in practice, and the more pertinent question may be whether missing values are informative. Allison (2010) later acknowledged that missing-indicators may lead to better predictions and their use for that purpose was acceptable. Missing-indicators have also been dismissed by Pigott (2001); Schafer & Graham (2002); Graham (2009); Aste et al. (2015), and are frequently omitted in overviews of missing data strategies (Schafer, 1997; Enders, 2010; Eirola, 2014; García et al., 2015; Das et al., 2018).

2.2. Previous Experiments

Only a handful of experimental comparisons of missing data approaches have included the missing-indicator approach, and these have been limited in scope. Vamplew & Adams (1992) and Ng & Yusoff (2011) only use a single dataset with randomly removed values, and base their evaluation on the performance of a single algorithm (respectively a neural network and linear regression). Pereira Barata et al. (2019) use three classification algorithms and 22 datasets, but again with randomly removed values, explicitly assuming an MCAR context. They conclude that imputation outperforms missing-indicators, but the comparison is not like-for-like, since it involves several forms of imputation but only combines indicator attributes with zero imputation. Van der Heijden et al. (2006) compare missing-indicators with zero imputation against several other forms of imputation without missing-indicators on one real dataset, for logistic regression. However, they do not evaluate predictive performance.

Ding & Simonoff (2010) conduct a more extensive investigation, using insights from a series of Monte Carlo simulations to systematically remove values from 36 datasets to simulate different forms of missingness. They use these datasets to compare zero imputation[3] with indicator attributes against mean/mode imputation without, as well as a number of other missing data approaches, for logistic regression. In addition, the authors evaluate a related representation of missing values[4] on the same set of 36 datasets, and on one real-life dataset with missing values, for decision trees. They find that there is strong evidence that representing missing values is the best approach when they are informative; when this is not the case their results show no strong difference.

The comparison by Grzymala-Busse & Hu (2000) is based on 10 datasets with naturally occurring missing values. However, the setting is purely categorical — all attributes are transformed into categorical attributes — the only form of imputation is mode imputation, and the missing value approaches are evaluated on the basis of the LERS classifier (Learning from Examples based on Rough Sets (Grzymala-Busse, 1988)).

Marlin (2008) compares zero imputation with missing-indicators (*augmentation with response indicators*) against several forms of imputation without, for logistic regression and neural networks, on the basis of an extensive series of simulations, one dataset with artificially removed values, and three real datasets. For the real datasets, there is no strong difference in performance between the different approaches.

Most recently, building on earlier experiments with simulated regression datasets (Josse et al., 2020; Le Morvan et al., 2021), Perez-Lebel et al. (2022) compare four different imputation techniques with and without missing-indicators (*missingness mask*) on seven prediction tasks derived from four real medical datasets, and conclude that missing-indicators consistently improve performance for gradient boosted trees, ridge regression and logistic regression.

We point out that the Missingness in Attribute (MIA) proposal (Twala et al., 2008) for decision trees and decision tree ensembles can be understood as an implicit combination of missing-indicators with automatic imputation, and has also been shown to outperform imputation without missing-indicators in small-scale experimental studies (Josse et al., 2020; Perez-Lebel et al., 2022).

Finally, even experimental comparisons of missing data that do not feature the missing-indicator approach generally do not involve more than a handful of real-life datasets with naturally occurring missing values. We have only found Luengo et al. (2012a;b), who use 21 datasets from the UCI repository, but 12 of these are problematic.[5]

---

[3] Presumably, Ding & Simonoff (2010) use one-hot encoding for categorical attributes, in which case zero imputation is equivalent to treating missing values as a separate category, but they do not state this explicitly.

[4] For categorical values, encoding missing values as a separate category, for numerical values, encoding missing values as an extremely large value that can always be split from the other values.

[5] The target column of the *echocardiogram* dataset ('alive-at-1') is supposed to denote whether a patient survived for at least one year, but it doesn't appear to agree with the columns from which it is derived, that denote how long a patient (has) survived and whether they were alive at the end of that period. The *audiology* dataset has a large number of small classes with complex labels and should perhaps be analysed with multi-label classification. In addition, it has ordinal attributes where the order of the values is not entirely clear, and three different values that potentially denote missingness ('?', 'unmeasured' and 'absent'), and it is not completely clear how they relate to each other. The *house-votes-84* dataset contains '?' values, but its documentation explicitly

Table 1: Classification Algorithms.

Name	Description
NN-1	Nearest neighbours (Fix & Hodges, 1951) with (Boscovich) 1-distance
NN-2	Nearest neighbours with (Euclidean) 2-distance
NN-1-D	Nearest neighbours with 1-distance, distance-weighted (Dudani, 1976)
NN-2-D	Nearest neighbours with 2-distance, distance-weighted
SVM-L	Soft-margin Support Vector Machine (Cortes & Vapnik, 1995) with linear kernel
SVM-G	Soft-margin Support Vector Machine with Gaussian kernel
LR	Multinomial logistic regression (Cox, 1966)
MLP	Multilayer perceptron (Rosenblatt, 1961) with ReLu activation (Fukushima, 1969), Glorot initialisation (Glorot & Bengio, 2010) and Adam optimisation (Kingma & Ba, 2015)
CART	Classification and Regression Tree (Breiman et al., 1984)
RF	Random Forest (Breiman, 2001)
ERT	Extremely Randomised Trees (Geurts et al., 2006)
ABT	Ada-boosted trees (Freund & Schapire, 1995) with SAMME (stagewise additive modeling using a multi-class exponential loss function) (Zhu et al., 2009)
GBM	Gradient Boosting Machine (Friedman, 2001)

3. Experimental Setup

To evaluate the effect of the missing-indicator approach on classification performance, we conduct a series of experiments, using the Python machine learning library *scikit-learn* (Pedregosa et al., 2011).

3.1. Questions

The aim of our experiments is to answer the following questions:

- Do missing-indicators increase performance, and does it matter which imputation strategy they are paired with?
- When do missing-indicators start to become useful in terms of missingness?
- Does using mean imputation instead of mode imputation allow for more information to be learned from missing categorical values?

3.2. Evaluation

We preprocess datasets by standardising numerical attributes and one-hot encoding categorical attributes (as required by the implementations in scikit-learn).

We measure classification performance by performing stratified five-fold cross-validation, repeating this for five different random states (which determine both the dataset splits and the initialisation of algorithms with a random component), and calculating the mean area under the receiver operator curve (AUROC). For multi-class datasets, we use the extension of AUROC defined by Hand & Till (2001).

To compare two alternatives A and B, we consider the p-value of a one-sided Wilcoxon signed-rank test (Wilcoxon, 1945) on the mean AUROC scores for our selection of datasets. When we compare A vs B, a score below 0.5 means that A increased performance on our selection of datasets; the lower the scores, the more confident we can be that this generalises to other similar datasets. Conversely, a score higher than 0.5 means that A decreased performance on our selection of datasets.

3.3. Imputation Strategies

We consider the following three imputation strategies:

- *Mean/mode imputation* replaces missing values of numerical and categorical attributes by, respectively, the mean and the mode of the non-missing values.
- *Nearest neighbour imputation* (Troyanskaya et al., 2001) replaces missing values of numerical and categorical attributes by, respectively, the mean and the mode of the 5 nearest non-missing values, with distance determined by the corresponding non-missing values for the other attributes.
- *Iterative imputation*, as implemented in scikit-learn, based on Van Buuren & Groothuis-Oudshoorn (2011),

---

states that these values are not unknown, but indicate different forms of abstention. The *ozone* dataset is a time-series problem, while the task associated with the *sponge* and *water-treatment* datasets is clustering, with no obvious target for classification among their respective attributes. Finally, the *breast-cancer* (9), *cleveland* (7), *dermatology* (8), *lung-cancer* (5), *post-operative* (3) and *wisconsin* (16) datasets contain only very few missing values, and any performance difference between missing value approaches on these datasets may to a large extent be coincidental.

Table 2: Real-life Classification Datasets with Missing Values from the UCI Repository for Machine Learning.

Dataset	Records	Classes	Attributes			Missing value rate		
			Num	Cat	Total	Num	Cat	Total
adult	48842	2	5	8	13	0.0	0.017	0.010
agaricus-lepiota	8124	2	1	21	22	0.0	0.015	0.014
aps-failure	76000	2	170	0	170	0.083		0.083
arrhythmia	443	10	279	0	279	0.0032		0.0032
bands	540	2	19	15	34	0.054	0.054	0.054
ckd	400	2	14	10	24	0.14	0.059	0.11
crx	690	2	6	9	15	0.0060	0.0068	0.0065
dress-sales	500	2	3	9	12	0.20	0.19	0.19
exasens	399	4	7	0	7	0.43		0.43
hcc	165	2	49	0	49	0.10		0.10
heart-disease	1611	2	13	1	14	0.18	0.0	0.17
hepatitis	155	2	19	0	19	0.057		0.057
horse-colic	368	2	19	1	20	0.25	0.39	0.26
mammographic-masses	961	2	2	2	4	0.042	0.041	0.042
mi	1700	8	111	0	111	0.085		0.085
nomao	34465	2	89	29	118	0.38	0.37	0.38
primary-tumor	330	15	16	1	17	0.029	0.20	0.039
secom	1567	2	590	0	590	0.045		0.045
soybean	683	19	22	13	35	0.099	0.096	0.098
thyroid0387	9172	18	7	16	23	0.22	0.0021	0.069

predicts missing values of one attribute on the basis of the other attribute values using a round-robin approach. For numerical attributes, this uses Bayesian ridge regression (Tipping, 2001), initialised with mean imputation, while for categorical attributes, we use logistic regression, initialised with mode imputation.

The scikit-learn implementations of nearest neighbour and iterative imputation can currently only impute numerical features, so we had to adapt them for categorical imputation. In all other aspects, we follow the default settings of scikit-learn.[6]

3.4. Classification Algorithms

We consider the classification algorithms listed in Table 1, as implemented in scikit-learn. Hyperparameters take their default values, except for SVM-L, LR and MLP, where we increase the maximum number of iterations to 10 000 to increase the probability of convergence.

For a number of these algorithms, specific ways have been proposed to handle missing values: e.g. NN-2-D (Dixon, 1979), SVM-G (Śmieja et al., 2019), MLP (Tresp et al., 1994; Śmieja et al., 2018; Ipsen et al., 2020) and CART (Quinlan, 1989; Twala et al., 2008). The purpose of the present experiment is to evaluate the general approach of using imputation with missing-indicators when these solutions have not been implemented, as is the case in scikit-learn.

3.5. Datasets

We use twenty real-life datasets with naturally occurring missing values from the UCI repository for machine learning (Dua & Graff, 2019) (Table 2). These datasets are quite varied — they cover a number of different domains and contain between 155 and 76 000 records, between 4 and 590 attributes, between 2 and 21 decision classes and missing value rates between 0.0032 and 0.43.

We have preprocessed these datasets in the following manner. We have removed attributes that were non-informative according to the accompanying documentation, as well as identifiers and alternative target values. When it was clear from the description that an attribute was categorical, we have treated it as such, even if it was originally represented with numerals. Conversely, where the possible values of an attribute admitted a semantic order, we have encoded them numerically. We have left binary attributes in their original encoding (categorical or numerical). To enable 5-fold cross-validation, we have removed classes with fewer than 5 records.

The individual datasets are described in Appendix A.

---

[6] For the *nomao* dataset, iterative imputation diverged, so we had to restrict imputation to the interval $[-100, 100]$.

Table 3: One-sided p-values, Imputation With Missing-Indicators versus Without.

Classifier	Imputation strategy Mean/mode	Neighbours	Iterative
NN-1	0.0088	0.0015	0.0017
NN-2	0.015	0.0024	0.00048
NN-1-D	0.0045	0.0019	0.0011
NN-2-D	0.0019	0.0031	0.00027
SVM-L	0.13	0.27	0.099
SVM-G	0.0032	0.0027	0.0021
LR	0.079	0.063	0.068
MLP	0.0027	0.0063	0.0056
CART	0.44	0.39	0.40
RF	0.038	0.051	0.17
ERT	0.28	0.0099	0.026
ABT	0.089	0.078	0.47
GBM	0.17	0.012	0.36

Table 4: One-sided p-values, Missing-Indicators with Iterative and Nearest Neighbour versus Mean/Mode Imputation.

Classifier	Imputation strategy Neighbours	Iterative
NN-1	0.94	0.15
NN-2	0.78	0.19
NN-1-D	0.97	0.55
NN-2-D	0.84	0.23
SVM-L	0.53	0.61
SVM-G	0.47	0.94
LR	0.40	0.83
MLP	0.30	0.55
CART	0.69	0.79
RF	0.61	0.86
ERT	0.61	0.64
ABT	0.33	0.78
GBM	0.93	0.85

4. Results and Discussion

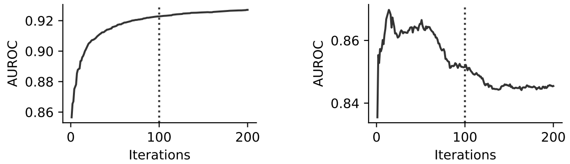
Using the experimental setup detailed in the previous section, we now try to answer the questions listed in Subsection 3.1. Full AUROC scores are provided in Appendix B.

4.1. Do Missing-Indicators Increase Performance, and Does It Matter Which Imputation Strategy They Are Paired With?

The p-values obtained by comparing imputation with and without missing-indicators are displayed in Table 3. Missing-indicators generally lead to increased performance — with the notable exception of CART, to which we return below. The more complicated imputation strategies do not result in much better results than mean/mode imputation when we pair imputation with missing-indicators (Table 4). At best, nearest neighbour and iterative imputation only lead to a modest improvement, and for many classifiers, they actually decrease performance. Therefore, we focus on mean/mode imputation for the remainder of this section.

A possible reason for the failure of missing-indicators to increase performance with CART, is that by default, the scikit-learn implementation of this classifier does not perform pruning, making it prone to overfitting. To test this hypothesis, we repeat our experiment for CART and mean imputation, but this time we apply cost complexity pruning ($\alpha = 0.01$). This clearly improves performance ($p = 0.0069$ without missing-indicators, $p = 0.015$ with missing-indicators), and now missing-indicators have a slight advantage ($p = 0.23$).

We have also taken a closer look at ERT and GBM, for which the performance increase from missing-indicators is not very significant. For ERT, this may be due to underfitting. If we increase the number of trees from the default 100 to 1000, this improves performance ($p = 0.0011$ without missing-indicators, $p = 0.0032$ with missing-indicators), and makes the advantage of missing-indicators somewhat clearer ($p = 0.092$).



(a) adult (b) mammographic-masses

Figure 1: GBM test AUROC for two illustrative datasets, using mean/mode imputation without missing-indicators, for one random state and one cross-validation fold. The default hyperparameter value of 100 iterations leads to under- (a) and overfitting (b).

For GBM, the default choice of 100 iterations of gradient descent can lead to both under- or overfitting, depending on the dataset (Fig. 1). We believe that it is generally preferable to continue training until an early-stopping criterion is met. However, applying the same criterion as with MLP[7] does not improve performance over the default of 100 ($p = 0.81$ without missing-indicators, $p = 0.85$ with missing-indicators) and does not change the relative advantage due to missing-indicators ($p = 0.20$).

---

[7] Setting aside 10% of the data for validation, stopping when validation loss has not decreased by at least 0.0001 for ten iterations, with a maximum of 10 000 iterations.

Table 5: Thresholds above Which Missing-Indicators Are More Likely Than Not to Increase AUROC, in Terms of the Absolute Number of Missing Values or the Missing Rate.

Classifier	Missing values		Missing rate	
	Cat	Num	Cat	Num
NN-1	1	302		
NN-2	2	130		
NN-1-D	1	291		
NN-2-D	1	73		
SVM-L			0.0	0.0
SVM-G			0.0	0.40
LR			0.0	0.0
CART			0.0	0.12
ERT			0.0	1.0
ABT	1	23200		
GBM			0.0	0.0

4.2. When Do Missing-Indicators Start to Become Useful in Terms of Missingness?

The theoretical motivation for representing missing values through missing-indicators is that this allows classifiers to learn the information encoded in their distribution. In principle, this should be easier when there are more examples to learn from. We can use this to obtain a better understanding of when missing-indicators might be useful on a per-attribute level.

We test this with the following additional experiment. For each attribute with missing values in each dataset, we reduce the original dataset by removing all other attributes with missing values. We thus obtain 1148 derived datasets, on which we again apply each of our classifiers (with pruning for CART, 1000 trees for ERT and early-stopping for GBM) and consider whether missing-indicators increase or decrease AUROC (we dismiss ties). Finally, for each classifier we fit a logistic regression model with cluster robust covariance (clustered by the originating dataset), with the following potential parameters: categoricalness (whether the attribute is categorical) and either the number of missing values (log-transformed) or the missing rate. We use the Akaike information criterion (Akaike, 1971) to decide whether to select these parameters.

We find that for most classifiers, either the absolute or the relative number of missing values is an informative parameter with positive coefficient. For MLP, neither parameter is informative, while for RF, the number of missing values is an informative parameter with negative coefficient, for which we have no explanation at present. For every classifier, categoricalness is an informative parameter with positive coefficient, meaning that missing-indicators are more beneficial for categorical than for numerical attributes.

Table 6: One-sided p-values, Mean Imputation after One-hot Encoding versus Mode Imputation of Missing Categorical Values.

Classifier	Without —	With missing-indicators
NN-1	0.020	0.077
NN-2	0.14	0.031
NN-1-D	0.016	0.12
NN-2-D	0.16	0.031
SVM-L	0.43	0.57
SVM-G	0.17	0.56
LR	0.81	0.057
MLP	0.16	0.60
CART	0.44	0.30
RF	0.046	0.57
ERT	0.030	0.95
ABT	0.48	0.62
GBM	0.077	0.54

The fitted logistic regression models allow us to calculate attribute-specific thresholds above which missing-indicators are more likely than not to increase AUROC, for all classifiers except MLP and RF (Table 5). In many cases, these thresholds are 1 or 0.0, indicating that missing-indicators are always likely to increase AUROC.

4.3. Does Using Mean Imputation Instead of Mode Imputation Allow for More Information to Be Learned from Missing Categorical Values?

As indicated above, missing-indicators are generally more likely to increase performance for categorical than for numerical attributes. A potential explanation for this is the fact that the mode of a categorical attribute is one of the non-missing values, whereas the mean of a numerical attribute is generally not equal to one of the non-missing values. Therefore, categorical imputation renders missing values truly indistinguishable from non-missing values, whereas numerical imputation does not — the information expressed by missing values may be partially recoverable, as argued by Le Morvan et al. (2021) and discussed in the Introduction.

We can achieve a similar partial representation of missing categorical values by changing the order in which we perform imputation and one-hot encoding, i.e. by performing numerical imputation on one-hot encoded categorical attributes with missing values. For imputation without missing-indicators, this indeed leads to better performance for some classifiers, while in combination with missing-indicators, it does not make much of a difference (Table 6)[8].

---

[8] LR is an exception here, we have no explanation for this.

5. Conclusion

We have presented the first large-scale experimental evaluation of the effect of the missing-indicator approach on classification performance, conducted on real datasets with naturally occurring missing values, paired with three different imputation techniques. The central question was whether, on balance, more benefit can be derived from the additional information encoded in a representation of missing values, or from the lower-dimensional projection of the data obtained by omitting missing-indicators.

On the whole, missing-indicators increase performance for the classification algorithms that we considered. An exception was CART, which suffers from overfitting in its default scikit-learn configuration. When pruning is applied, missing-indicators do increase performance. For ERT, the advantage of missing-indicators becomes more significant when underfitting is controlled.

We also found that, in the presence of missing-indicators, nearest neighbour and iterative imputation do not significantly increase performance over simple mean/mode imputation. This is a useful finding, because implementations of more sophisticated imputation strategies may not always be available to practitioners working in different frameworks, or easy to apply.

In a follow-up experiment, we determined attribute-specific missingness thresholds above which missing-indicators are more likely than not to increase performance. For categorical attributes, this threshold is generally very low, while for numerical attributes, there is more variation among classifiers, in particular as to whether this threshold is absolute or relative to the total number of records.

The greater usefulness of missing-indicators for categorical than for numerical attributes can be explained by the fact that the mean of a numerical attribute is not generally identical to any of the non-missing values, and that mean imputation therefore preserves some of the information of missing values. This is supported by the results of a further experiment, which showed that, in the absence of missing-indicators, applying mean imputation to one-hot encoded categorical attributes results in somewhat better performance than mode imputation.

We conclude that the combination of mean/mode imputation with missing-indicators is a safe default approach towards missing values in classification tasks. While over- or underfitting is a concern for certain classifiers, it is a concern for these classifiers with or without missing-indicators. However, practitioners may want to omit missing-indicators when the classification algorithm to be used has a special provision for missing values, when the missingness thresholds that we determined are not met, or on the basis of specific information about the distribution of missing values in the dataset. The use of missing-indicators can also be combined with dimensionality reduction algorithms to increase the information density of the resulting dataset.

While we have considered the use of missing-indicators with imputation, they can in principle also be used to supplement other, learner-specific solutions for missing-values. Whether this makes sense and increases performance will differ from case to case, and we leave this as an open question. In any case, we believe that going forward, any experimental evaluation of such learner-specific proposals should take missing-indicators into account.

The problem of missing data has been the subject of a rich body of theoretical literature. We hope to have contributed with this paper to the practical evaluation of some of that theory. In particular, we are happy to have identified twenty real-life datasets with missing values, and hope that in the future, more such datasets will be collected.

Acknowledgements

The research reported in this paper was conducted with the financial support of the Odysseus programme of the Research Foundation – Flanders (FWO). We would like to express our thanks to Geert van der Heijden for answering a question about Van der Heijden et al. (2006).

References

Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, pp. 267–281. Akadémiai Kiadó, 1971.

Allison, P. D. *Missing Data*. Sage Publications, Thousand Oaks, California, 2001.

Allison, P. D. Missing data. In Marsden, P. V. and Wright, J. D. (eds.), *Handbook of Survey Research*, chapter 20, pp. 631–657. Emerald Group Publishing, Bingley, England, second edition, 2010.

Anderson, A. B., Basilevsky, A., and Hum, D. P. J. Missing data: A review of the literature. In Rossi, P. H., Wright, J. D., and Anderson, A. B. (eds.), *Handbook of Survey Research*, Quantitive Studies in Social Relations, chapter 12, pp. 415–494. Academic Press, New York, 1983.

Aste, M., Boninsegna, M., Freno, A., and Trentin, E. Techniques for dealing with incomplete data: a tutorial and survey. *Pattern Analysis and Applications*, 18(1):1–29, 2015.

Breiman, L. Random forests. *Machine Learning*, 45(1): 5–32, 2001.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and Regression Trees*. The Wadsworth statistics/probability series. Wadsworth, Monterey, California, 1984.

van Buuren, S. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011.

Candillier, L. and Lemaire, V. Design and analysis of the Nomao Challenge: Active learning in the real-world. In *ECML-PKDD 2012: Active Learning in Real-world Applications Workshop*, 2012.

Cestnik, B., Kononenko, I., and Bratko, I. ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. In *EWSL 87: Proceedings of the 2nd European Working Session on Learning*, pp. 31–45. Sigma Press, 1987.

Chow, W. K. A look at various estimators in logistic models in the presence of missing values. Technical Report N-1324-HEW, Rand Corporation, Santa Monica, California, 1979.

Cohen, J. Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70(6):426–443, 1968.

Cohen, J. and Cohen, P. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.*, chapter 7. Missing Data, pp. 265–290. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1975.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Cox, D. R. Some procedures connected with the logistic qualitative response curve. in (fn david, ed.) research papers in statistics: Essays in honour of j. neyman's 70th birthday. In David, F. N. (ed.), *Research Papers in Statistics: Festschrift for J. Neyman*, pp. 55–71. John Wiley & Sons, London, 1966.

Das, S., Datta, S., and Chaudhuri, B. B. Handling data irregularities in classification: Foundations, trends, and future challenges. *Pattern Recognition*, 81:674–693, 2018.

Detrano, R., Yiannikas, J., Salcedo, E. E., Rincon, G., Go, R. T., Williams, G., and Leatherman, J. Bayesian probability analysis: a prospective demonstration of its clinical utility in diagnosing coronary disease. *Circulation*, 69(3): 541–547, 1984.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., and Froelicher, V. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5): 304–310, 1989.

Ding, Y. and Simonoff, J. S. An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research*, 11(1):131–170, 2010.

Dixon, J. K. Pattern recognition with partly missing data. *IEEE Transactions on Systems, Man, and Cybernetics*, 9 (10):617–621, 1979.

Dua, D. and Graff, C. UCI machine learning repository, 2019. URL `http://archive.ics.uci.edu/ml`.

Dudani, S. A. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, 6(4):325–327, 1976.

Efron, B. and Gong, G. Statistical theory and the computer. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pp. 3–7. Springer, 1981.

Eirola, E. *Machine learning methods for incomplete data and variable selection*. PhD thesis, Aalto University, Espoo, 2014.

Elter, M., Schulz-Wendtland, R., and Wittenberg, T. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Medical physics*, 34(11):4164–4172, 2007.

Enders, C. K. *Applied Missing Data Analysis*. Methodology in the Social Sciences. The Guilford Press, New York, 2010.

Evans, B. and Fisher, D. Overcoming process delays with decision tree induction. *IEEE Expert*, 9(1):60–66, 1994.

Ferreira Costa, C. and Nascimento, M. A. IDA 2016 industrial challenge: Using machine learning for predicting failures. In *IDA 2016: Proceedings of the 15th International Symposium on Intelligent Data Analysis*, volume 9897 of *Lecture Notes in Computer Science*, pp. 381–386. Springer, 2016.

Fix, E. and Hodges, Jr, J. Discriminatory analysis — nonparametric discrimination: Consistency properties. Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

Freund, Y. and Schapire, R. E. A desicion-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT '95: Proceedings of the Second European Conference on Computational Learning Theory*, volume 904 of *Lecture Notes in Computer Science*, pp. 23–37. Springer, 1995.

Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

Fukushima, K. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333, 1969.

García, S., Luengo, J., and Herrera, F. *Data Preprocessing in Data Mining*, volume 72 of *Intelligent Systems Reference Library*, chapter 4. Dealing with Missing Values. Springer, Cham, Zug, 2015.

Geurts, P., Ernst, D., and Wehenkel, L. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS 2010: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Golovenkin, S. E., Bac, J., Chervov, A., Mirkes, E. M., Orlova, Y. V., Barillot, E., Gorban, A. N., and Zinovyev, A. Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data. *GigaScience*, 9(11):giaa128, 2020.

Graham, J. W. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60:549–576, 2009.

Grzymala-Busse, J. W. Knowledge acquisition under uncertainty—a rough set approach. *Journal of Intelligent and Robotic Systems*, 1(1):3–16, 1988.

Grzymala-Busse, J. W. and Hu, M. A comparison of several approaches to missing attribute values in data mining. In *RSCTC 2000: Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing*, volume 2005 of *Lecture Notes in Artificial Intelligence*, pp. 378–385. Springer, 2000.

Güvenir, H. A., Acar, B., Demiröz, G., and Çekin, A. A supervised machine learning algorithm for arrhythmia analysis. In *Proceedings of the 24th Annual Meeting of Computers in Cardiology*, volume 24 of *Computers in Cardiology*, pp. 433–436. IEEE, 1997.

Hand, D. J. and Till, R. J. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning*, 45(2):171–186, 2001.

van der Heijden, G. J. M. G., Donders, A. R. T., Stijnen, T., and Moons, K. G. M. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *Journal of Clinical Epidemiology*, 59 (10):1102–1109, 2006.

Hutcheson, Jr, J. D. and Prather, J. E. Interpreting the effects of missing data in survey research. *Southeastern Political Review*, 9(2):129–143, 1981.

Ipsen, N., Mattei, P.-A., and Frellsen, J. How to deal with missing data in supervised deep learning? In *Artemiss 2020: First ICML Workshop on the Art of Learning with Missing Values*, 2020.

Jones, M. P. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, 91(433): 222–230, 1996.

Josse, J., Prost, N., Scornet, E., and Varoquaux, G. On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*, 2020. doi: 10.48550/ARXIV.1902.06931.

Kim, J.-O. and Curry, J. The treatment of missing data in multivariate analysis. *Sociological Methods & Research*, 6(2):215–240, 1977.

Kingma, D. P. and Ba, J. L. Adam: A method for stochastic optimization. In *ICLR 2015: 3rd International Conference on Learning Representations*, 2015.

Kohavi, R. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *KDD-96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207. AAAI Press, 1996.

Le Morvan, M., Josse, J., Scornet, E., and Varoquaux, G. What's a good imputation to predict with missing values? In *NeurIPS 2021: Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems*, volume 34 of *Advances in neural information processing systems*, pp. 11530–11540. NIPS Foundation, 2021.

Lincoff, G. H. *The Audubon Society Field Guide to North American Mushrooms*. Alfred A Knopf, New York, 1981.

Luengo, J., García, S., and Herrera, F. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems*, 32(1):77–108, 2012a.

Luengo, J., Sáez, J. A., and Herrera, F. Missing data imputation for fuzzy rule-based classification systems. *Soft Computing*, 16(5):863–881, 2012b.

Marlin, B. M. *Missing Data Problems in Machine Learning*. PhD thesis, University of Toronto, 2008.

McCann, M., Li, Y., Maguire, L., and Johnston, A. Causality challenge: benchmarking relevant signal components for effective monitoring and process control. In *NIPS 2008:*

Proceedings of Workshop on Causality, volume 6 of *Proceedings of Machine Learning Research*, pp. 277–288. JMLR Workshop and Conference Proceedings, 2008.

McLeish, M. and Cecile, M. Enhancing medical expert systems with knowledge obtained from statistical data. *Annals of Mathematics and Artificial Intelligence*, 2(1–4): 261–276, 1990.

Michalski, R. S. and Chilausky, R. L. Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. *International Journal of Policy Analysis and Information Systems*, 4(2):125–161, 1980.

Ng, C. G. and Yusoff, M. S. B. Missing values in data analysis: Ignore or impute? *Education in Medicine Journal*, 3(1), 2011.

Orme, J. G. and Reis, J. Multiple regression with missing data. *Journal of Social Service Research*, 15(1–2):61–91, 1991.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Pereira Barata, A., Takes, F. W., van den Herik, H. J., and Veenman, C. J. Imputation methods outperform missing-indicator for data missing completely at random. In *ICDM 2019: Proceedings of the Workshops*, pp. 407–414. IEEE, 2019.

Perez-Lebel, A., Varoquaux, G., Le Morvan, M., Josse, J., and Poline, J.-B. Benchmarking missing-values approaches for predictive models on health databases. *GigaScience*, 11(1):giac013, 2022.

Pigott, T. D. A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383, 2001.

Quinlan, J. R. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3):221–234, 1987.

Quinlan, J. R. Unknown attribute values in induction. In *Proceedings of the Sixth International Workshop on Machine Learning*, pp. 164–168. Morgan Kaufmann, 1989.

Quinlan, J. R., Compton, P. J., Horn, K. A., and Lazarus, L. Inductive knowledge acquisition: a case study. In *Proceedings of the Second Australian Conference on Applications of Expert Systems*, pp. 157–173. Turing Institute Press, 1986.

Rosenblatt, F. Principles of neurodynamics — perceptrons and the theory of brain mechanisms. Technical Report VG-1196-G-8, Cornell Aeronautical Laboratory, Buffalo, New York, 1961.

Rossiev, D. A., Golovenkin, S. E., Shulman, V., and Matjushin, G. Neural networks for forecasting of myocardial infarction complications. In *Proceedings of the Second International Symposium on Neuroinformatics and Neurocomputers*, pp. 292–298. IEEE, 1995.

Rubin, D. B. Inference and missing data. *Biometrika*, 63(3): 581–592, 1976.

Rubini, L. J. and Eswaran, P. Generating comparative analysis of early stage prediction of chronic kidney disease. *International Journal of Modern Engineering Research*, 5(7):49–55, 2015.

Santos, M. S., Abreu, P. H., García-Laencina, P. J., Simão, A., and Carvalho, A. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. *Journal of biomedical informatics*, 58:49–59, 2015.

Schafer, J. L. *Analysis of Incomplete Multivariate Data*, volume 72 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1997.

Schafer, J. L. and Graham, J. W. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002.

Schlimmer, J. C. *Concept Acquisition Through Representational Adjustment*. PhD thesis, University of California, Irvine, 1987.

Śmieja, M., Struski, Ł., Tabor, J., Zieliński, B., and Spurek, P. Processing of missing data by neural networks. In *NeurIPS 2018: Proceedings of the Thirty-second Annual Conference on Neural Information Processing Systems*, volume 31 of *Advances in neural information processing systems*, pp. 689–696. NIPS Foundation, 2018.

Śmieja, M., Struski, Ł., Tabor, J., and Marzec, M. Generalized RBF kernel for incomplete data. *Knowledge-Based Systems*, 173:150–162, 2019.

Soltani Zarrin, P., Röckendorf, N., and Wenger, C. In-vitro classification of saliva samples of COPD patients and healthy controls using machine learning tools. *IEEE Access*, 8:168053–168060, 2020.

Sperrin, M., Martin, G. P., Sisk, R., and Peek, N. Missing data should be handled differently for prediction than for description or causal explanation. *Journal of Clinical Epidemiology*, 125:183–187, 2020.

Stumpf, S. A. A note on handling missing data. *Journal of Management*, 4(1):65–73, 1978.

Tipping, M. E. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

Tresp, V., Neuneier, R., and Ahmad, S. Efficient methods for dealing with missing data in supervised learning. In *NIPS-94: Proceedings of the Eighth Annual Conference on Neural Information Processing Systems*, volume 7 of *Advances in neural information processing systems*, pp. 689–696. MIT Press, 1994.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

Twala, B. E., Jones, M., and Hand, D. J. Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 29(7):950–956, 2008.

Vamplew, P. and Adams, A. Missing values in a backpropagation neural net. In *ACNN '92: Proceedings of the Third Australian Conference on Neural Networks*, pp. 64–66. Sydney University Electrical Engineering, 1992.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Zhu, J., Zou, H., Rosset, S., and Hastie, T. Multi-class AdaBoost. *Statistics and Its Interface*, 2(3):349–360, 2009.

A. Datasets

We have used the following twenty datasets in our experiments, all from the UCI repository for machine learning (Dua & Graff, 2019).

adult (Kohavi, 1996)

48 842 1994 census records of American adults. The task is to predict whether each person earns more than $50 000 per year (11 687) or not (37 155), based on 13 census questions.

We have removed the 'fnlwgt' attribute, the weight that needs to be applied to each record to obtain a representative socio-economic sample within each US state.

The version of the dataset used by Kohavi (1996) has 45 222 records — these are the records without missing values.

The data was extracted from the 1994 census database by Barry Becker.

agaricus-lepiota (Schlimmer, 1987)

8124 mushrooms from the *Agaricus* and *Lepiota* families, to be classified as edible (4208) or poisonous (3916) on the basis of 22 physical characteristics.

It is unclear whether the missing values, all in 'stalk-root', represent actually missing information, or a missing stalk-root.

This dataset was created on the basis of the information provided by Lincoff (1981). The version of the dataset used by Schlimmer (1987) contained only 3078 mushrooms (from 23 species). Although Schlimmer (1987) claims 23 attributes, he lists only 22, so this is most likely a mistake (possibly due to the number of mushroom species).

aps-failure (Ferreira Costa & Nascimento, 2016)

76 000 component failures in Scania trucks. The task is to predict whether a specific component in the air pressure system (APS) has failed (1375) or some other component (74 625), based on 170 measurements.

This dataset was provided by Tony Lindgren of the Department of Computer and Systems Sciences at Stockholm University and Jonas Biteus at Scania for the industrial challenge at the 15th International Symposium on Intelligent Data Analysis (IDA) in 2016.

arrhythmia (Güvenir et al., 1997)

452 patients in 13 classes, indicating the presence and type of arrhythmia. The 279 attributes consist of the age, sex, height and weight of the patients as well as a large number of characteristics of their ECG recordings.

This dataset is strongly imbalanced: 245 patients have no arrhythmia, while there are five classes with fewer than 10 records.

bands (Evans & Fisher, 1994)

540 rotogravure printing cylinders, displaying banding (228) or not (312), to be classified on the basis of 34 attributes describing the printing press and its use.

We have preprocessed this dataset by removing the 'timestamp', 'cylinder number', 'customer' and 'job number' attributes, which do not really form categories, as well as the 'ink color' attribute, which only has one value. None of these attributes are used by Evans & Fisher (1994).

There are additional differences with respect to the variant of this dataset used by Evans & Fisher (1994). That variant does not have the 'cylinder division', 'press type', 'paper mill location', 'callper' and 'roller durometer' attributes, but does have additional 'blade oscillation' and 'basis weight' attributes, for a total of 31. It also covers a shorter time period than is contained in the final version of the dataset.

ckd (Rubini & Eswaran, 2015)

400 people, 250 of which with chronic kidney disease (CKD), 150 without, to be classified on the basis of 24 measurements.

The origin of this dataset is not explained by Rubini & Eswaran (2015).

crx (Quinlan, 1987)

690 credit card applications, 307 of which were approved and 383 of which were not.

The data was provided by a large bank. The meaning of the 15 attributes is confidential.

dress-sales

500 dresses offered for sale by AliExpress between August and October 2013, recommended (210) or not (290) on the basis of 12 properties.

We have preprocessed this dataset by eliminating spelling variations and interpreting certain values as missing values.

This dataset was created by Muhammad Usman and Adeel Ahmed at the Air University in Islamabad, who do not appear to have used it in any publication. It is unclear what the meaning of the two classes is. The documentation suggests that there is a connection with the number of sales of each dress, which are also provided, but there doesn't appear to be any direct link.

exasens (Soltani Zarrin et al., 2020)

399 patients of the medical clinic in Borstel, near Sülfeld, Germany, and healthy controls, to be classified as healthy (160) or having chronic obstructive pulmonary disease (COPD, 79), asthma (80) or a respiratory infection (80) based on 7 attributes: age, gender, smoking status and four values expressing saliva permittivity.

The dataset used by Soltani Zarrin et al. (2020) only contains the healthy and COPD patients.

hcc (Santos et al., 2015)

165 hepatocellular carcinoma (HCC) patients of the Coimbra University Hospital. The task is to predict 1-year survival (102) or not (63) on the basis of 49 attributes expressing risk factors, comorbidies and a range of tests.

heart-disease (Detrano et al., 1989)

1611 patients from five hospitals, with (903) or without (708) heart disease, defined as more than 50% narrowing of any major bloodvessel. There are 14 attributes, including the hospital, the age and sex of the patient and the results of a number of tests.

We have preprocessed this dataset by reducing the id-attribute to only identify the source hospital.

This dataset has a complicated history. The data was collected by Dr Robert Detrano at the Cleveland Clinic and at the Veterans Administration Medical Center in Long Beach, Dr Andras Janosi at the Hungarian Institute of Cardiology in Budapest, Dr William Steinbrunn at the University Hospital in Zürich, and Dr Matthias Pfisterer at the University Hospital in Basel.

On the basis of the id-attribute, it is possible to identify several batches of records: one batch from Cleveland (303 records), two batches from Budapest (428 and 351 records), two batches from Long Beach (200 and 201 records), one batch from Zürich (58 records) and one batch from Basel (73 records). On the basis of these numbers, we can deduce that the dataset used by Detrano et al. (1989) does not contain the second batches from Budapest and Long Beach, nor three records at the end of the first batch from Budapest. It contains 85 records from Basel, which means that 12 records are missing. The pilot study by Detrano et al. (1984) used only 154 patients from Cleveland.

The first batch of Long Beach records appears to have three duplicate pairs of records, with the same or nearly the same name, social security number, age, sex and other attribute values. Given that four of the clinical attribute values are slightly different, it is unclear whether these are truly duplicate records or separate examinations of the same patients.

Nevertheless, we have decided to remove the second record of each pair during preprocessing.

hepatitis **(Efron & Gong, 1981)**

155 chronic hepatitis patients, 33 of which died and 122 of which lived. There are 19 attributes, consisting of patient characteristics, symptoms and test results.

The data was collected by Dr Peter Gregory.

horse-colic **(McLeish & Cecile, 1990)**

368 horses with colic presented to the Ontario Veterinary College hospital in Guelph. The task is to predict whether (in retrospect) the lesion was surgical (232) or not (136), based on 20 symptoms and measurements.

We have preprocessed this dataset by deleting two non-informative attributes and five attributes that are alternative prediction targets (according to the documentation). It is not clear whether McLeish & Cecile (1990) used the exact same attribute set.

mammographic-masses **(Elter et al., 2007)**

961 full-field digital mammograms, to be classified as benign (516) or malignant (445) on the basis of 4 attributes: the patient's age and the shape, margin and density of the masses.

The data was collected at the Institute of Radiology of the University of Erlangen-Nuremberg between 2003 and 2006.

mi **(Golovenkin et al., 2020)**

1700 patients with myocardial infarction (MI). The 8 classes describe whether the patient died, and if so, what the cause of death was. The 111 attributes consist of patient characteristics, comorbidies, test outcomes and symptoms.

This dataset is very imbalanced, as the class with surviving patients contains 1429 records.

The data was collected at the Krasnoyarsk Interdistrict Clinical Hospital between 1992 and 1995. Earlier versions of this dataset were used by e.g. Rossiev et al. (1995).

nomao **(Candillier & Lemaire, 2012)**

34 465 pairs of place records. The task is to predict whether the two records refer to the same place (24 621) or to different places (9844), on the basis of 118 attributes expressing the similarity or difference of the attributes of the two original records.

The data for this dataset was provided by Nomao for the 'Nomao Challenge' of the 2012 Active Learning in Real-world Applications ECML-PKDD Workshop.

primary-tumor **(Cestnik et al., 1987)**

339 cancer patients. The task is to identify the site of the primary tumor out of 21 possibilities, based on 17 attributes. Most attributes are boolean and refer to body parts. Their meaning is slightly unclear, it is possible that they refer to the locations that the cancer has spread to.

Many of the classes are very small. There are six classes with fewer than 5 records. In fact, by design the number of classes is 22, but one class is empty.

The data was collected at the University Medical Centre in Ljubljana by M Zwitter and M Soklic.

secom **(McCann et al., 2008)**

1567 produced wafers at a production line of a semiconductor fabrication plant, 1463 of which passed testing and 104 of which failed, to be classified on the basis of 590 signals.

This dataset was created for the 'Causality Challenge' of the 2008 NIPS Workshop on Causality.

soybean **(Michalski & Chilausky, 1980)**

683 soybean plants, displaying 19 different diseases, to be classified on the basis of 35 symptoms.

Michalski & Chilausky (1980) omitted the four smallest classes, using only 630 records.

thyroid0387 **(Quinlan et al., 1986)**

9172 thyroid patients at St Vincent's Hospital in Sydney between August 1984 and January 1987. The task is to predict the diagnosis out of 18 classes, based on 29 patient characteristics and test scores.

This dataset is strongly imbalanced: 6771 patients have no diagnosis, while there are three classes with fewer than ten records.

The variant of this dataset used by Quinlan et al. (1986) only had 3066 records, and didn't have the 'I131 treatment', 'hypopituitary', 'psych' and 'referral source' attributes.

We have had to preprocess this dataset because a small number of records belonged to multiple classes. When one diagnosis was indicated as being more likely than another, we retained the more likely diagnosis. Otherwise, we resolved this by retaining the most specific class. Furthermore, the provided file already contained missing-indicators, which we have removed to properly evaluate imputation without missing-indicators.

B. Full Classification Results

We list here the results of our experiments in greater detail. Table 7 contains the mean AUROC across five-fold cross-validation and five random states for each classifier, each dataset, each imputation strategy, without and with missing-indicators. Table 8 contains the mean AUROC for CART, GBM and ERT with updated hyperparameter values (as discussed in Subsection 4.1). Table 9 contains the mean AUROC obtained by imputing missing categorical values with the mean, after one-hot encoding (Subsection 4.3).

Table 7: AUROC, Main Experiment. **Bold**: Higher Value (Without or With Missing-Indicators) by at Least 0.001.

Classifier	Dataset	Imputation strategy, missing-indicators no/yes					
		Mean/mode		Neighbours		Iterative	
		No	Yes	No	Yes	No	Yes
NN-1	adult	0.857	0.858	0.858	0.858	0.858	0.858
	agaricus-lepiota	1.000	1.000	1.000	1.000	1.000	1.000
	aps-failure	**0.928**	0.926	**0.926**	0.922	**0.928**	0.923
	arrhythmia	0.760	0.760	0.760	0.760	0.760	0.760
	bands	0.836	**0.838**	0.834	**0.847**	0.836	**0.848**
	ckd	**0.993**	0.992	0.989	0.990	**0.992**	0.989
	crx	0.908	**0.909**	0.904	**0.908**	0.909	**0.910**
	dress-sales	0.548	**0.555**	0.540	**0.545**	0.527	**0.531**
	exasens	0.710	**0.726**	0.703	**0.713**	0.717	**0.726**
	hcc	0.699	**0.760**	0.707	**0.745**	0.712	**0.753**
	heart-disease	0.846	0.847	0.841	**0.844**	0.843	**0.846**
	hepatitis	**0.849**	0.841	0.841	**0.850**	0.839	**0.847**
	horse-colic	0.716	**0.733**	**0.738**	0.734	0.726	**0.738**
	mammographic-masses	0.821	**0.827**	0.821	**0.825**	0.824	**0.831**
	mi	0.572	**0.579**	0.564	**0.580**	0.569	**0.579**
	nomao	**0.983**	0.982	0.978	**0.981**	0.983	0.982
	primary-tumor	0.675	**0.687**	0.678	**0.693**	0.676	**0.687**
	secom	0.641	**0.651**	0.641	**0.643**	0.646	**0.653**
	soybean	0.993	0.993	0.992	0.993	0.993	0.993
	thyroid0387	0.876	**0.883**	0.852	**0.876**	0.873	**0.884**
NN-2	adult	0.860	0.861	0.861	0.861	0.861	0.860
	agaricus-lepiota	1.000	1.000	1.000	1.000	1.000	1.000
	aps-failure	0.920	**0.922**	0.918	**0.920**	0.921	0.921
	arrhythmia	0.733	0.733	0.734	0.734	0.733	0.733
	bands	0.830	**0.832**	0.818	**0.835**	0.825	**0.836**
	ckd	**0.995**	0.992	0.990	**0.991**	0.991	0.991
	crx	0.899	0.900	0.898	**0.899**	0.900	0.901
	dress-sales	**0.554**	0.547	**0.541**	0.539	**0.532**	0.527
	exasens	0.709	**0.716**	0.699	**0.706**	0.712	**0.718**
	hcc	0.690	**0.696**	0.695	**0.709**	0.698	**0.705**
	heart-disease	0.831	**0.835**	0.828	**0.837**	0.829	**0.836**
	hepatitis	**0.861**	0.851	0.846	**0.850**	0.860	**0.862**
	horse-colic	0.684	**0.710**	**0.724**	0.706	0.695	**0.704**
	mammographic-masses	0.820	**0.825**	0.821	**0.824**	0.822	**0.828**
	mi	0.561	**0.563**	0.555	**0.560**	0.563	0.563
	nomao	0.980	**0.982**	0.976	**0.980**	0.980	**0.981**
	primary-tumor	0.667	**0.673**	0.670	**0.675**	0.666	**0.677**
	secom	0.607	**0.612**	0.614	**0.617**	0.607	**0.613**
	soybean	0.986	**0.988**	0.987	0.988	0.986	**0.988**

Continued on next page

Table 7: AUROC, Main Experiment. **Bold**: Higher Value (Without or With Missing-Indicators) by at Least 0.001.

Classifier	Dataset	Imputation strategy, missing-indicators no/yes					
		Mean/mode		Neighbours		Iterative	
		No	Yes	No	Yes	No	Yes
	thyroid0387	0.871	**0.878**	0.848	**0.875**	0.866	**0.876**
NN-1-D	adult	0.838	0.838	0.837	**0.839**	0.837	0.838
	agaricus-lepiota	1.000	1.000	1.000	1.000	1.000	1.000
	aps-failure	**0.929**	0.926	**0.927**	0.922	**0.928**	0.923
	arrhythmia	0.764	0.764	0.763	0.763	0.764	0.764
	bands	0.871	**0.875**	0.865	**0.879**	0.870	**0.880**
	ckd	**0.994**	0.992	0.989	0.990	**0.992**	0.989
	crx	0.907	**0.908**	0.905	**0.908**	0.908	**0.909**
	dress-sales	0.544	**0.560**	0.538	**0.545**	0.528	**0.535**
	exasens	0.629	**0.641**	0.625	**0.634**	0.632	**0.640**
	hcc	0.728	**0.786**	0.733	**0.772**	0.738	**0.773**
	heart-disease	0.847	**0.848**	0.843	**0.845**	0.843	**0.847**
	hepatitis	**0.857**	0.853	0.841	**0.855**	0.841	**0.853**
	horse-colic	0.743	**0.751**	**0.762**	0.752	0.749	**0.757**
	mammographic-masses	0.802	**0.806**	0.798	**0.805**	0.803	**0.808**
	mi	0.572	**0.580**	0.564	**0.580**	0.569	**0.579**
	nomao	**0.984**	0.983	0.979	**0.982**	0.984	0.983
	primary-tumor	0.665	**0.676**	0.667	**0.684**	0.665	**0.677**
	secom	0.644	**0.652**	0.644	**0.645**	0.647	**0.655**
	soybean	0.993	0.993	0.992	0.993	0.993	0.993
	thyroid0387	0.878	**0.884**	0.853	**0.878**	0.874	**0.885**
NN-2-D	adult	0.842	0.843	0.842	**0.843**	0.842	0.843
	agaricus-lepiota	1.000	1.000	1.000	1.000	1.000	1.000
	aps-failure	0.920	**0.922**	0.918	**0.921**	0.922	0.922
	arrhythmia	0.735	0.736	0.736	0.736	0.735	0.735
	bands	0.859	**0.861**	0.844	**0.863**	0.850	**0.863**
	ckd	**0.995**	0.993	0.990	**0.991**	0.991	0.991
	crx	0.898	0.899	0.898	**0.900**	0.900	0.901
	dress-sales	0.548	0.548	**0.543**	0.538	**0.534**	0.532
	exasens	0.628	**0.635**	0.623	**0.629**	0.629	**0.634**
	hcc	0.710	**0.723**	0.716	**0.737**	0.719	**0.729**
	heart-disease	0.833	**0.838**	0.830	**0.839**	0.831	**0.839**
	hepatitis	**0.862**	0.856	0.847	**0.852**	0.859	**0.865**
	horse-colic	0.712	**0.731**	**0.745**	0.730	0.719	**0.729**
	mammographic-masses	0.802	**0.805**	0.799	**0.804**	0.802	**0.807**
	mi	0.560	**0.563**	0.556	**0.560**	0.564	0.565
	nomao	0.981	**0.983**	0.977	**0.981**	0.981	**0.982**
	primary-tumor	0.659	**0.666**	0.660	**0.667**	0.657	**0.669**
	secom	0.606	**0.610**	0.612	**0.615**	0.606	**0.611**
	soybean	0.986	**0.988**	0.987	0.988	0.986	**0.988**
	thyroid0387	0.873	**0.880**	0.850	**0.877**	0.868	**0.877**
SVM-L	adult	0.905	**0.906**	0.905	0.906	0.905	0.906
	agaricus-lepiota	1.000	1.000	1.000	1.000	1.000	1.000
	aps-failure	0.966	**0.969**	0.961	**0.969**	0.963	**0.966**
	arrhythmia	0.818	**0.843**	0.819	**0.843**	0.818	**0.843**
	bands	0.796	**0.817**	0.791	**0.809**	0.760	**0.801**
	ckd	1.000	1.000	0.999	1.000	0.999	1.000

Continued on next page

Table 7: AUROC, Main Experiment. **Bold**: Higher Value (Without or With Missing-Indicators) by at Least 0.001.

Classifier	Dataset	Imputation strategy, missing-indicators no/yes					
		Mean/mode		Neighbours		Iterative	
		No	Yes	No	Yes	No	Yes
	crx	**0.922**	0.920	0.920	0.920	0.922	0.921
	dress-sales	**0.598**	0.593	**0.594**	0.588	0.591	**0.597**
	exasens	0.762	**0.780**	0.761	**0.769**	0.761	**0.780**
	hcc	**0.757**	0.738	**0.781**	0.756	**0.746**	0.733
	heart-disease	0.866	0.865	0.866	**0.867**	0.867	0.868
	hepatitis	**0.848**	0.824	**0.857**	0.831	**0.856**	0.833
	horse-colic	**0.790**	0.784	**0.798**	0.784	**0.770**	0.762
	mammographic-masses	0.865	**0.867**	0.862	**0.865**	0.864	0.864
	mi	0.641	**0.666**	0.639	**0.669**	0.636	**0.671**
	nomao	0.986	**0.988**	0.986	**0.988**	0.985	**0.988**
	primary-tumor	0.769	0.769	**0.772**	0.770	0.778	0.777
	secom	0.626	**0.629**	**0.671**	0.659	**0.631**	0.628
	soybean	0.999	0.999	0.999	0.999	0.999	0.999
	thyroid0387	0.957	**0.965**	0.951	**0.963**	0.939	**0.957**
SVM-G	adult	0.895	**0.897**	0.896	0.896	0.896	0.897
	agaricus-lepiota	1.000	1.000	1.000	1.000	1.000	1.000
	aps-failure	0.967	0.968	0.960	**0.965**	0.965	**0.966**
	arrhythmia	0.848	0.848	0.848	0.848	0.848	0.848
	bands	0.855	**0.865**	0.858	**0.870**	0.857	**0.869**
	ckd	1.000	1.000	1.000	1.000	1.000	1.000
	crx	0.926	**0.927**	0.924	**0.927**	0.926	**0.928**
	dress-sales	0.618	**0.620**	**0.620**	0.619	0.607	**0.612**
	exasens	0.772	**0.780**	0.767	**0.780**	0.773	**0.780**
	hcc	0.778	**0.790**	0.785	**0.793**	0.770	**0.783**
	heart-disease	**0.865**	0.864	0.863	**0.864**	0.864	0.864
	hepatitis	0.893	0.892	**0.888**	0.887	**0.893**	0.890
	horse-colic	0.768	**0.771**	0.784	**0.786**	0.767	**0.769**
	mammographic-masses	0.840	**0.845**	0.838	**0.841**	0.839	**0.842**
	mi	0.635	**0.643**	0.637	**0.645**	0.639	**0.648**
	nomao	0.991	0.992	0.988	**0.991**	0.989	**0.991**
	primary-tumor	0.762	**0.765**	0.764	**0.767**	0.766	0.767
	secom	**0.699**	0.694	**0.702**	0.698	**0.689**	0.685
	soybean	0.999	0.999	0.999	0.999	0.999	0.999
	thyroid0387	0.976	**0.978**	0.965	**0.977**	0.966	**0.970**
LR	adult	0.905	**0.906**	0.906	0.906	0.906	0.906
	agaricus-lepiota	1.000	1.000	1.000	1.000	1.000	1.000
	aps-failure	0.971	**0.979**	0.971	**0.980**	0.967	**0.978**
	arrhythmia	0.860	0.860	0.860	0.860	0.859	0.860
	bands	0.819	**0.833**	0.811	**0.830**	0.808	**0.828**
	ckd	1.000	1.000	1.000	1.000	1.000	1.000
	crx	0.924	0.923	0.923	0.923	0.924	0.924
	dress-sales	0.620	0.620	0.619	**0.624**	0.614	**0.620**
	exasens	0.774	**0.783**	0.768	**0.775**	0.773	**0.782**
	hcc	**0.778**	0.760	**0.796**	0.774	**0.772**	0.755
	heart-disease	0.867	**0.868**	0.867	**0.869**	0.867	**0.869**
	hepatitis	**0.863**	0.856	**0.871**	0.862	**0.870**	0.862
	horse-colic	**0.789**	0.786	**0.793**	0.786	**0.769**	0.764

Continued on next page

Table 7: AUROC, Main Experiment. **Bold**: Higher Value (Without or With Missing-Indicators) by at Least 0.001.

Classifier	Dataset	Imputation strategy, missing-indicators no/yes					
		Mean/mode		Neighbours		Iterative	
		No	Yes	No	Yes	No	Yes
	mammographic-masses	0.866	**0.868**	0.863	**0.865**	0.865	0.865
	mi	0.654	**0.685**	0.645	**0.685**	0.650	**0.688**
	nomao	0.986	**0.988**	0.986	**0.988**	0.985	**0.988**
	primary-tumor	0.773	**0.776**	0.772	**0.775**	0.780	**0.783**
	secom	**0.686**	0.678	**0.687**	0.680	**0.676**	0.673
	soybean	0.999	0.999	0.999	0.999	0.999	0.999
	thyroid0387	0.970	**0.974**	0.966	**0.974**	0.967	**0.974**
MLP	adult	0.890	0.890	**0.891**	0.889	**0.891**	0.890
	agaricus-lepiota	1.000	1.000	1.000	1.000	1.000	1.000
	aps-failure	0.928	**0.942**	0.931	**0.943**	0.931	**0.942**
	arrhythmia	0.831	**0.846**	0.831	**0.845**	0.831	**0.845**
	bands	0.871	**0.879**	0.873	**0.885**	0.868	**0.882**
	ckd	1.000	1.000	1.000	1.000	1.000	1.000
	crx	0.902	**0.906**	0.901	**0.905**	0.900	**0.905**
	dress-sales	0.549	**0.553**	0.560	**0.561**	0.544	0.545
	exasens	0.759	**0.762**	0.746	**0.755**	0.757	**0.763**
	hcc	0.778	**0.781**	0.791	**0.796**	0.777	**0.781**
	heart-disease	**0.819**	0.815	**0.816**	0.811	**0.818**	0.816
	hepatitis	0.861	0.861	**0.870**	0.865	**0.872**	0.866
	horse-colic	0.714	**0.744**	0.727	**0.756**	0.719	**0.734**
	mammographic-masses	**0.845**	0.840	**0.841**	0.836	**0.847**	0.840
	mi	0.659	**0.695**	0.656	**0.697**	0.660	**0.697**
	nomao	0.991	0.991	0.987	**0.990**	0.990	0.991
	primary-tumor	0.768	**0.782**	0.765	**0.778**	0.769	**0.785**
	secom	0.693	**0.701**	0.699	**0.704**	0.686	**0.697**
	soybean	0.999	0.999	0.999	0.999	0.999	0.999
	thyroid0387	0.986	**0.988**	0.979	**0.987**	0.980	**0.986**
CART	adult	0.776	0.775	0.776	0.775	**0.776**	0.774
	agaricus-lepiota	1.000	1.000	1.000	1.000	1.000	1.000
	aps-failure	0.855	**0.858**	**0.858**	0.857	0.854	**0.857**
	arrhythmia	**0.712**	0.710	**0.712**	0.702	**0.714**	0.702
	bands	**0.716**	0.713	0.697	**0.716**	0.706	**0.717**
	ckd	0.971	**0.977**	0.981	**0.985**	**0.979**	0.978
	crx	**0.818**	0.812	**0.813**	0.810	**0.815**	0.809
	dress-sales	0.524	**0.548**	0.526	**0.529**	**0.534**	0.532
	exasens	**0.618**	0.616	**0.618**	0.608	0.621	**0.626**
	hcc	0.593	**0.603**	**0.619**	0.617	**0.614**	0.601
	heart-disease	0.702	0.703	0.701	0.700	0.703	**0.706**
	hepatitis	**0.660**	0.657	**0.691**	0.673	**0.703**	0.700
	horse-colic	**0.695**	0.673	**0.700**	0.663	**0.680**	0.676
	mammographic-masses	**0.748**	0.744	0.747	0.746	0.744	**0.746**
	mi	0.572	0.572	0.549	**0.574**	0.557	**0.571**
	nomao	0.935	0.935	0.922	**0.925**	0.926	0.927
	primary-tumor	0.621	0.621	0.625	**0.627**	0.622	**0.623**
	secom	0.547	**0.552**	0.555	**0.558**	**0.542**	0.538
	soybean	0.975	**0.977**	0.973	0.974	0.971	**0.973**
	thyroid0387	0.879	**0.897**	0.828	**0.883**	0.836	**0.881**

Continued on next page

Table 7: AUROC, Main Experiment. **Bold**: Higher Value (Without or With Missing-Indicators) by at Least 0.001.

Classifier	Dataset	Imputation strategy, missing-indicators no/yes					
		Mean/mode		Neighbours		Iterative	
		No	Yes	No	Yes	No	Yes
RF	adult	0.890	0.890	0.890	0.891	0.891	0.890
	agaricus-lepiota	1.000	1.000	1.000	1.000	1.000	1.000
	aps-failure	0.988	0.989	0.988	0.989	0.988	0.988
	arrhythmia	0.883	**0.884**	0.885	0.885	**0.886**	0.883
	bands	0.893	**0.896**	0.886	**0.898**	0.896	0.896
	ckd	1.000	1.000	1.000	1.000	1.000	1.000
	crx	0.932	0.931	**0.934**	0.932	0.931	0.931
	dress-sales	0.591	**0.606**	0.583	**0.602**	0.582	**0.597**
	exasens	0.701	0.701	0.689	**0.694**	0.698	**0.701**
	hcc	0.803	**0.816**	0.813	0.813	0.794	**0.806**
	heart-disease	0.861	**0.864**	0.862	**0.866**	0.864	**0.866**
	hepatitis	0.882	**0.887**	**0.890**	0.887	**0.888**	0.886
	horse-colic	**0.800**	0.791	**0.811**	0.809	0.793	0.792
	mammographic-masses	0.812	**0.821**	0.815	**0.819**	0.812	**0.820**
	mi	0.687	0.687	0.676	**0.681**	**0.687**	0.679
	nomao	0.994	0.994	0.991	**0.992**	0.993	0.993
	primary-tumor	0.749	**0.758**	0.730	**0.761**	0.748	**0.761**
	secom	**0.722**	0.710	**0.719**	0.713	**0.722**	0.710
	soybean	0.999	0.999	0.999	0.999	0.999	0.999
	thyroid0387	0.994	0.994	0.988	**0.991**	0.988	**0.990**
ERT	adult	0.846	0.847	0.847	0.847	0.846	0.847
	agaricus-lepiota	1.000	1.000	1.000	1.000	1.000	1.000
	aps-failure	0.989	0.989	0.989	0.988	0.989	0.989
	arrhythmia	0.885	**0.889**	0.881	**0.885**	0.881	**0.885**
	bands	0.889	0.890	0.874	**0.890**	0.885	**0.892**
	ckd	1.000	1.000	1.000	1.000	1.000	1.000
	crx	**0.913**	0.911	0.916	0.915	**0.912**	0.910
	dress-sales	0.572	**0.600**	0.563	**0.594**	0.560	**0.589**
	exasens	**0.633**	0.632	0.622	**0.626**	0.624	**0.630**
	hcc	0.783	**0.799**	0.776	**0.804**	0.771	**0.796**
	heart-disease	0.858	**0.861**	0.862	**0.865**	0.861	0.861
	hepatitis	**0.871**	0.861	0.876	**0.877**	**0.882**	0.871
	horse-colic	**0.793**	0.780	**0.818**	0.796	**0.790**	0.780
	mammographic-masses	0.793	**0.801**	0.791	**0.800**	0.793	**0.801**
	mi	**0.689**	0.683	0.661	**0.683**	0.676	**0.686**
	nomao	0.994	0.993	0.991	**0.992**	0.993	0.993
	primary-tumor	0.702	**0.718**	0.698	**0.717**	0.704	**0.721**
	secom	**0.718**	0.713	**0.716**	0.705	0.706	**0.716**
	soybean	0.999	0.999	0.999	0.999	0.999	0.999
	thyroid0387	0.981	0.982	0.972	**0.979**	0.972	**0.979**
ABT	adult	0.915	0.915	0.915	0.915	0.915	0.915
	agaricus-lepiota	1.000	1.000	1.000	1.000	1.000	1.000
	aps-failure	0.987	0.987	0.987	0.987	0.986	0.987
	arrhythmia	**0.634**	0.632	0.634	0.633	**0.634**	0.632
	bands	0.806	0.806	0.793	**0.809**	0.805	**0.807**
	ckd	1.000	1.000	0.999	1.000	0.998	**1.000**
	crx	0.905	0.906	0.907	0.906	**0.909**	0.905

Continued on next page

Table 7: AUROC, Main Experiment. **Bold**: Higher Value (Without or With Missing-Indicators) by at Least 0.001.

Classifier	Dataset	Imputation strategy, missing-indicators no/yes					
		Mean/mode		Neighbours		Iterative	
		No	Yes	No	Yes	No	Yes
	dress-sales	**0.590**	0.582	**0.584**	0.578	0.587	**0.589**
	exasens	0.720	0.720	0.705	**0.717**	**0.713**	0.711
	hcc	0.715	**0.724**	**0.739**	0.735	**0.708**	0.687
	heart-disease	0.860	0.860	0.857	**0.861**	**0.861**	0.858
	hepatitis	0.797	**0.804**	0.824	**0.830**	0.805	**0.814**
	horse-colic	**0.753**	0.752	**0.749**	0.742	**0.735**	0.729
	mammographic-masses	0.856	**0.857**	0.855	0.856	0.854	0.855
	mi	0.555	**0.572**	0.572	**0.586**	0.573	0.572
	nomao	0.987	0.987	0.985	0.986	0.986	0.986
	primary-tumor	**0.661**	0.660	**0.670**	0.668	0.668	**0.671**
	secom	0.670	0.670	0.661	0.661	0.663	0.663
	soybean	0.863	**0.871**	0.777	**0.850**	0.855	**0.865**
	thyroid0387	0.685	0.685	0.704	**0.707**	0.712	**0.714**
GBM	adult	0.921	0.921	0.921	0.921	0.921	0.921
	agaricus-lepiota	1.000	1.000	1.000	1.000	1.000	1.000
	aps-failure	0.989	0.988	0.988	0.989	0.988	0.988
	arrhythmia	0.873	**0.874**	**0.880**	0.875	**0.879**	0.878
	bands	0.869	0.870	0.857	**0.871**	0.870	**0.873**
	ckd	1.000	1.000	0.997	0.997	0.998	0.998
	crx	0.932	0.932	0.930	0.931	0.929	**0.931**
	dress-sales	**0.612**	0.606	0.597	**0.601**	**0.612**	0.609
	exasens	0.725	0.725	0.720	**0.724**	0.723	**0.725**
	hcc	0.759	**0.780**	0.762	**0.773**	**0.747**	0.742
	heart-disease	0.872	0.872	0.869	0.870	0.873	0.872
	hepatitis	**0.837**	0.828	0.837	**0.838**	0.854	0.854
	horse-colic	**0.793**	0.789	**0.794**	0.789	**0.798**	0.789
	mammographic-masses	0.850	**0.853**	0.847	**0.851**	0.846	**0.853**
	mi	**0.664**	0.663	0.659	**0.663**	0.654	**0.661**
	nomao	0.991	0.991	0.989	**0.990**	0.991	0.991
	primary-tumor	0.760	**0.763**	0.762	0.762	**0.754**	0.752
	secom	0.708	**0.710**	**0.717**	0.716	0.708	**0.711**
	soybean	0.999	0.999	0.998	0.999	0.998	0.998
	thyroid0387	0.885	**0.918**	0.904	**0.915**	**0.866**	0.860

Table 8: AUROC, Additional Experiment for Mean/Mode Imputation and Classifiers with Adjusted Hyperparameter Values. **Bold**: Higher Value (Without or With Missing-Indicators) by at Least 0.001.

Dataset	Classifier, missing-indicators no/yes					
	CART		GBM		ERT	
	No	Yes	No	Yes	No	Yes
adult	0.844	0.844	0.927	0.927	0.847	0.847
agaricus-lepiota	0.991	0.992	1.000	1.000	1.000	1.000
aps-failure	0.859	0.859	0.988	0.988	0.991	0.991
arrhythmia	**0.749**	0.748	0.850	**0.852**	0.897	**0.899**
bands	0.749	**0.759**	0.855	**0.857**	0.890	0.890

Continued on next page

Table 8: AUROC, Additional Experiment for Mean/Mode Imputation and Classifiers with Adjusted Hyperparameter Values. **Bold**: Higher Value (Without or With Missing-Indicators) by at Least 0.001.

Dataset	Classifier, missing-indicators no/yes					
	CART		GBM		ERT	
	No	Yes	No	Yes	No	Yes
ckd	0.976	0.976	0.997	0.996	1.000	1.000
crx	0.897	0.897	0.934	0.933	0.914	0.914
dress-sales	0.568	**0.570**	0.608	**0.614**	0.572	**0.602**
exasens	0.723	**0.732**	0.755	**0.757**	0.626	0.626
hcc	0.577	**0.588**	0.737	**0.745**	0.791	**0.808**
heart-disease	0.777	0.777	0.870	**0.871**	0.861	**0.862**
hepatitis	**0.626**	0.578	**0.812**	0.809	**0.877**	0.873
horse-colic	**0.742**	0.724	**0.789**	0.783	**0.799**	0.782
mammographic-masses	0.823	0.823	0.857	**0.859**	0.795	**0.802**
mi	0.586	**0.592**	**0.650**	0.639	**0.702**	0.695
nomao	0.916	0.916	0.994	0.994	0.994	0.994
primary-tumor	0.703	**0.707**	0.766	0.767	0.705	**0.714**
secom	0.500	0.500	**0.684**	0.677	0.746	**0.747**
soybean	0.990	0.991	0.999	0.999	0.999	0.999
thyroid0387	0.909	0.909	0.904	**0.918**	0.987	0.988

Table 9: AUROC, Additional Experiment for Imputation of Categorical Attributes (Mode Imputation or Mean Imputation after One-hot Encoding). extbfBold: Higher Value by at Least 0.001.

Classifier	Dataset	Without missing-indicators		With missing-indicators	
		Mode	Mean	Mode	Mean
NN-1	adult	0.857	0.858	0.858	0.858
	agaricus-lepiota	1.000	1.000	1.000	1.000
	bands	0.836	**0.839**	0.838	**0.843**
	ckd	0.993	**0.997**	0.992	**0.994**
	crx	0.908	**0.909**	0.909	0.909
	dress-sales	**0.548**	0.533	**0.555**	0.539
	horse-colic	0.716	**0.737**	0.733	**0.737**
	mammographic-masses	0.821	**0.831**	0.827	**0.828**
	nomao	0.983	0.984	0.982	0.982
	primary-tumor	0.675	**0.679**	0.687	**0.693**
	soybean	0.993	0.993	0.993	0.993
	thyroid0387	0.876	**0.878**	0.883	**0.885**
NN-2	adult	0.860	0.861	0.861	0.861
	agaricus-lepiota	1.000	1.000	1.000	1.000
	bands	0.830	0.829	0.832	**0.834**
	ckd	0.995	**0.997**	0.992	**0.994**
	crx	0.899	0.898	0.900	0.900
	dress-sales	**0.554**	0.548	**0.547**	0.531
	horse-colic	0.684	**0.688**	0.710	**0.719**
	mammographic-masses	0.820	**0.824**	0.825	0.825
	nomao	0.980	0.981	0.982	0.982
	primary-tumor	0.667	**0.669**	0.673	0.674
	soybean	0.986	0.986	0.988	0.988

Continued on next page

Table 9: AUROC, Additional Experiment for Imputation of Categorical Attributes (Mode Imputation or Mean Imputation after One-hot Encoding). extbfBold: Higher Value by at Least 0.001.

Classifier	Dataset	Without missing-indicators		With missing-indicators	
		Mode	Mean	Mode	Mean
	thyroid0387	0.871	**0.872**	0.878	0.879
NN-1-D	adult	0.838	0.838	0.838	0.838
	agaricus-lepiota	1.000	1.000	1.000	1.000
	bands	0.871	**0.874**	0.875	0.876
	ckd	0.994	**0.997**	0.992	**0.994**
	crx	0.907	0.908	0.908	0.908
	dress-sales	**0.544**	0.537	**0.560**	0.544
	horse-colic	0.743	**0.763**	0.751	**0.756**
	mammographic-masses	0.802	**0.810**	0.806	0.807
	nomao	0.984	0.985	0.983	0.983
	primary-tumor	0.665	**0.669**	0.676	**0.681**
	soybean	0.993	0.993	0.993	0.993
	thyroid0387	0.878	**0.880**	0.884	**0.887**
NN-2-D	adult	0.842	0.843	0.843	0.843
	agaricus-lepiota	1.000	1.000	1.000	1.000
	bands	**0.859**	0.857	0.861	**0.862**
	ckd	0.995	**0.997**	0.993	**0.994**
	crx	0.898	0.898	0.899	0.900
	dress-sales	**0.548**	0.543	**0.548**	0.535
	horse-colic	0.712	**0.716**	0.731	**0.739**
	mammographic-masses	0.802	**0.806**	0.805	0.806
	nomao	0.981	0.982	0.983	0.983
	primary-tumor	0.659	**0.661**	0.666	0.667
	soybean	0.986	0.986	0.988	0.988
	thyroid0387	0.873	**0.874**	0.880	0.881
SVM-L	adult	0.905	0.905	0.906	0.906
	agaricus-lepiota	1.000	1.000	1.000	1.000
	bands	0.796	**0.797**	0.817	0.817
	ckd	1.000	1.000	1.000	1.000
	crx	0.922	0.921	0.920	0.920
	dress-sales	**0.598**	0.590	0.593	0.593
	horse-colic	0.790	**0.794**	0.784	0.784
	mammographic-masses	0.865	0.866	0.867	0.867
	nomao	**0.986**	0.984	0.988	0.988
	primary-tumor	0.769	0.769	0.769	0.769
	soybean	0.999	0.999	0.999	0.999
	thyroid0387	0.957	0.957	0.965	0.965
SVM-G	adult	0.895	**0.896**	0.897	0.897
	agaricus-lepiota	1.000	1.000	1.000	1.000
	bands	0.855	0.856	0.865	**0.867**
	ckd	1.000	1.000	1.000	1.000
	crx	0.926	0.925	0.927	0.927
	dress-sales	**0.618**	0.609	**0.620**	0.614
	horse-colic	0.768	**0.774**	0.771	**0.774**
	mammographic-masses	0.840	**0.843**	**0.845**	0.843
	nomao	0.991	0.991	0.992	0.992
	primary-tumor	0.762	**0.764**	0.765	0.766

Continued on next page

Table 9: AUROC, Additional Experiment for Imputation of Categorical Attributes (Mode Imputation or Mean Imputation after One-hot Encoding). extbfBold: Higher Value by at Least 0.001.

Classifier	Dataset	Without missing-indicators		With missing-indicators	
		Mode	Mean	Mode	Mean
LR	soybean	0.999	0.999	0.999	0.999
	thyroid0387	0.976	0.976	0.978	0.978
	adult	0.905	0.906	0.906	0.906
	agaricus-lepiota	1.000	1.000	1.000	1.000
	bands	**0.819**	0.814	0.833	0.832
	ckd	1.000	1.000	1.000	1.000
	crx	0.924	0.924	0.923	0.924
	dress-sales	**0.620**	0.611	0.620	0.620
	horse-colic	**0.789**	0.788	0.786	0.787
	mammographic-masses	0.866	0.867	0.868	0.868
	nomao	**0.986**	0.984	0.988	0.988
	primary-tumor	0.773	0.773	0.776	0.776
MLP	soybean	0.999	0.999	0.999	0.999
	thyroid0387	0.970	0.970	0.974	0.974
	adult	0.890	**0.891**	0.890	0.890
	agaricus-lepiota	1.000	1.000	1.000	1.000
	bands	0.871	**0.874**	0.879	**0.882**
	ckd	1.000	1.000	1.000	1.000
	crx	0.902	0.902	0.906	0.906
	dress-sales	**0.549**	0.540	**0.553**	0.549
	horse-colic	0.714	**0.727**	0.744	**0.749**
	mammographic-masses	0.845	0.844	0.840	0.841
	nomao	0.991	0.991	0.991	0.991
	primary-tumor	0.768	**0.769**	**0.782**	0.781
CART	soybean	0.999	0.999	0.999	0.999
	thyroid0387	0.986	0.986	0.988	0.988
	adult	0.844	0.844	0.844	0.844
	agaricus-lepiota	0.991	0.991	0.992	0.991
	bands	**0.749**	0.744	**0.759**	0.757
	ckd	0.976	0.977	0.976	**0.977**
	crx	0.897	**0.899**	0.897	**0.899**
	dress-sales	0.568	0.568	**0.570**	0.568
	horse-colic	**0.742**	0.728	**0.724**	0.723
	mammographic-masses	0.823	0.822	**0.823**	0.821
	nomao	0.916	0.916	0.916	0.916
	primary-tumor	0.703	**0.739**	0.707	**0.738**
RF	soybean	0.990	**0.995**	0.991	**0.995**
	thyroid0387	0.909	0.909	0.909	0.909
	adult	0.890	0.891	0.890	0.890
	agaricus-lepiota	1.000	1.000	1.000	1.000
	bands	0.893	**0.895**	**0.896**	0.890
	ckd	1.000	1.000	1.000	1.000
	crx	0.932	0.933	0.931	0.930
	dress-sales	**0.591**	0.589	**0.606**	0.589
	horse-colic	0.800	**0.802**	0.791	**0.795**
	mammographic-masses	0.812	**0.823**	0.821	**0.822**
	nomao	0.994	0.994	0.994	0.994

Continued on next page

Table 9: AUROC, Additional Experiment for Imputation of Categorical Attributes (Mode Imputation or Mean Imputation after One-hot Encoding). extbfBold: Higher Value by at Least 0.001.

Classifier	Dataset	Without missing-indicators		With missing-indicators	
		Mode	Mean	Mode	Mean
	primary-tumor	0.749	**0.753**	0.758	0.759
	soybean	0.999	0.999	0.999	0.999
	thyroid0387	0.994	0.994	0.994	0.994
ERT	adult	0.847	**0.848**	0.847	0.847
	agaricus-lepiota	1.000	1.000	1.000	1.000
	bands	0.890	**0.893**	**0.890**	0.889
	ckd	1.000	1.000	1.000	1.000
	crx	0.914	0.914	0.914	0.914
	dress-sales	0.572	**0.589**	**0.602**	0.591
	horse-colic	0.799	**0.806**	0.782	**0.785**
	mammographic-masses	0.795	**0.804**	0.802	0.801
	nomao	0.994	0.994	0.994	0.994
	primary-tumor	0.705	**0.711**	0.714	0.713
	soybean	0.999	0.999	0.999	0.999
	thyroid0387	0.987	0.987	0.988	0.987
ABT	adult	0.915	0.915	0.915	0.915
	agaricus-lepiota	1.000	1.000	1.000	1.000
	bands	0.806	0.806	**0.806**	0.805
	ckd	1.000	1.000	1.000	1.000
	crx	0.905	0.906	**0.906**	0.904
	dress-sales	**0.590**	0.582	**0.582**	0.579
	horse-colic	0.753	**0.763**	0.752	**0.764**
	mammographic-masses	0.856	0.857	0.857	0.858
	nomao	0.987	0.987	0.987	0.987
	primary-tumor	**0.661**	0.640	**0.660**	0.639
	soybean	**0.863**	0.859	0.871	**0.873**
	thyroid0387	0.685	0.685	0.685	0.685
GBM	adult	0.927	0.927	0.927	0.927
	agaricus-lepiota	1.000	1.000	1.000	1.000
	bands	0.855	0.855	**0.857**	0.854
	ckd	0.997	0.997	0.996	0.996
	crx	0.934	0.934	0.933	0.934
	dress-sales	**0.608**	0.606	**0.614**	0.608
	horse-colic	0.789	**0.792**	0.783	**0.788**
	mammographic-masses	0.857	0.857	**0.859**	0.858
	nomao	0.994	0.994	0.994	0.994
	primary-tumor	0.766	**0.770**	0.767	**0.769**
	soybean	0.999	0.999	0.999	0.999
	thyroid0387	0.904	**0.907**	**0.918**	0.916