Missing Data Imputation Based On Probabilistic Data

C.M. van Kampen
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
c.m.vankampen@student.utwente.nl

ABSTRACT

Missing data is a vast problem in data science. There are different reasons that data might be missing. An example would be when people take a survey, but want to keep some information private. There exist several good methods that try to handle missing data. The most unambiguous method involves deleting the missing values or records containing missing values. However, this is often not preferred as too much information might be lost. Therefore, missing data is usually handled by predicting the missing values using an imputation method. Good imputation techniques exist, but they often introduce a bias into the data. This research attempts to develop an improved imputation method based on probabilistic data. This method will be compared to known methods by a developed evaluation framework. It is assumed this novel method will improve data quality.

Keywords

Missing data, Imputation, Probabilistic data, Uncertainty, Random Forest

1. INTRODUCTION

Since the beginning of data processing, missing data has been a well-known problem. The problem concretes itself by data in a database or dataset that is missing. There are different reasons that data is missing: a survey is not filled out completely (non-response), the data was not inserted into the database, problems with measurement tools, etc. The problem with missing values is that it can introduce a substantial amount of bias [8]. This means that further processing of the data could be declared invalid.

There are different techniques to handle missing data and reduce (almost never completely remove) the unbiasedness. Deletion is the most straightforward example. This method removes all missing values. There are two variations of this method: list-wise and pairwise deletion. The first deletes any record having one or more missing values and the latter removes only the missing values. In some cases, this method is permissible. Though, as the amount of missing values increases, the resulting complete dataset might be too small for relevant research. Furthermore, deletion does not take the associations between variables into consideration.

Multiple other methods exist. These are based around the idea of imputation. This technique tries to determine a plausible value to fill in for the missing field. Most of these methods share the latter problem with deletion which results in a biased dataset.

In this research, an attempt is done to develop an imputation method that reduces the resulting bias and improve the data quality. This is accomplished using a probabilistic data structure.

1.1 Research Questions

This research is centered around the following questions. Both questions are split in sub questions that need to be answered before the main research question can be answered.

RQ 1. To what extent can an imputation technique based on probabilistic data be designed?

RQ 1.1. How should uncertainty be measured and included in the imputed data?

RQ 2. How does the quality of this imputation technique compare to existing alternatives?

RQ 2.1. How should an imputation technique be evaluated?

RQ 2.2. How should the quality of the data be measured?

To come about answering these questions, this research is centered around **RQ 1.**. The goal is to create a probabilistic imputation method. In software development, a testing environment is initially created before developing the actual software. Therefore, this research also started out by creating an evaluation framework (**RQ 2.1.**) to compare several imputation methods and analyze their effectiveness. A set of metrics has been determined that are used in related research and give a good quality measure (**RQ 2.2.**). These metrics are output by the evaluation framework and are used to answer **RQ 2.**. Next, the probabilistic method has been developed by considering different alternatives, each of which is compared to known imputation methods.

This paper will start out by providing a more comprehensive description of the concept of missing data imputation in section 2. Furthermore, related research can be found in section 3 which furnishes a backbone for this research.

31st Twente Student Conference on IT July 5th, 2019, Enschede, The Netherlands.

Then the research method is given in section 4 and results from the evaluation framework (section 4.3) are provided in section 5.

2. BACKGROUND

Missing data is a well-known problem. In this section, the concept of missing data will be explained more comprehensive and it goes more in depth of imputation. Furthermore, a definition for probabilistic data and data quality is given.

2.1 Missing Data Mechanisms

Since there are different reasons data might be missing, there are also different types of missing data (often called mechanisms). We can differentiate three of these mechanisms:

1. **Missing completely at random (MCAR)**
   In MCAR, the probability that data is missing is the same for all records. The probability of missing data does not depend on recorded data, neither on unrecorded data. An example of MCAR is when a measurement tool is accidentally broken.

2. **Missing at random (MAR)** MAR data depends on already recorded data. For example, bankers might be less likely to share their income than a teacher. Unfortunately, there is no approach to prove whether data is MAR [7].

3. **Missing not at random (MNAR)** Data that is MNAR only depends on the missing value. If someone is ashamed by their income, they might not want to share this information. It is possible to transform MNAR data into MAR data by looking for potential causes of missingness [7]. This is useful as a lot of methods that handle missing data work better in an MAR situation.

2.2 Probabilistic Data and Uncertainty

Uncertainty is an important aspect to consider when processing data. Especially in the context of missing data imputation where the resulting dataset should resemble the real dataset as close as possible. One problem with the resulting dataset using known imputation techniques is that the imputed values are assumed to be real. This means that any further processing of the data is going to be biased if the imputed values are biased. Therefore, this research tries to discover an imputation method based on probabilistic data.

Probabilistic or uncertain data can be represented by including multiple plausible values where each value has a probability that together sum to 1. This results in a set of possible worlds: all the possible combinations of the plausible values [19].

By imputing with probabilistic data, the resulting dataset includes an uncertainty measure making it less bias.

Finally, it is good to consider that there are two different kinds of uncertainty. Tuple-level uncertainty is about the correctness of a tuple or record. If the record is not correct, it should be excluded from the dataset. The other kind is attribute-level uncertainty. At this level, the uncertainty is about which value from a set of plausible values should be used for a single attribute in a record. In this research, the latter kind of uncertainty is considered.

2.3 Quality of Data

When data is imputed, the quality needs to be measured to check whether the imputation technique used, is a good one. First a definition of quality of data is necessary. For data to be of good quality, it needs to be realistic and, moreover, unbiased. All imputation techniques deal with the realistic property by imputing with data close to known values. The unbiasedness constraint is not always met. This constraint requires that there is a standard error in the imputed data. When this error is too small, the precision is overestimated and might result in detecting an association where none exists [12]. Too large of an error and the results are not realistic anymore. In case of single imputation, the error is often too small or too large to be unbiased. Therefore a technique, called multiple imputation, has been created which aims to be unbiased [12].

2.4 Imputation

2.4.1 Single Imputation

As mentioned, imputation is a technique to predict plausible values for missing data. Single imputation is based around a single rule. An example is mean imputation. This method imputes every missing value with the mean of the corresponding attribute. A problem with this is that every missing value is assigned the same value which distorts the distribution and it adds no extra information [10][12]. Furthermore, this method does not take into account, the associations between the imputed attribute and other attributes in the dataset. This results in the fact that mean imputation is biased in case of MAR data and is therefore not preferred [3]. An alternative to mean imputation is regression. A regression model is built around a target variable with missing values. The model is estimated based on known values of the variable and of variables that are related to the target variable by an association. The model is then used to impute missing values. As opposed to mean imputation, regression preserves the distribution [10]. Though, it can also not be considered a good imputation method. The problem of regression imputation is that no error is included, resulting in too realistic values [3]. Stochastic regression tries to deal with this problem by adding the average regression variance to the imputed data. Nevertheless, this error is often not accurate enough for the data to be completely unbiased [3].

2.4.2 Multiple Imputation

Multiple imputation is a technique that solves this problem. It is (approximately) unbiased and was introduced by Rubin (1987). It consists of three steps:

1. **Imputation**
   In multiple imputation, missing values are imputed m times, resulting in m (>1) possible datasets.

2. **Analysis**
   The m datasets are individually analyzed, resulting in m analyzed datasets.

3. **Pooling**
   From the m analyzed datasets, a single dataset is determined by combining values into one.

Multiple imputation is widely used and can be used with all three cases of missing data [12]. As multiple imputation works on m plausible datasets, uncertainty is taken

into account. Moreover, a higher m will decrease the biasedness of the data [15][18]. In practice, an m of 20 will be most effective. Multiple imputation by chained equations (MICE), developed by Van Buuren [18], is the most used method.

Multiple imputation assumes the input data to be MAR. Researchers have found the technique to be very powerful. However, there are cases in which a different imputation technique introduces less bias, because the data is MNAR [10][17].

3. RELATED WORK

Considerable research has been done in the field of missing data handling. In this section, some research will be referenced and used as a backbone for this paper.

To determine data quality of an imputed dataset, a set of metrics has to be defined. In a lot of research within data science, the root mean square error (RMSE) is employed as a performance measure [2][16][20]. Furthermore, for prediction and classification problems, often a confusion matrix is constructed. From such a matrix, the accuracy, precision and recall can be determined [11].

Overall, a lot of research that has been done compares several imputation methods. Continuously, the conclusions drawn (almost) always conclude that multiple imputation is superior to single imputation methods [9][12][20]. Therefore, multiple imputation by chained equations (MICE) is often the preferred method to use.

Furthermore, some research has been done with machine learning methods [5][14]. For example, a random forest classifier can be trained to predict missing values [11][13]. This has showed to be a very effective imputation method.

All in all, it is shown that several imputation methods exist and one can be preferred over the other under different conditions such as the type of missing data. Generally, multiple imputation is one of the better and most used methods.

From this literature, it can be found that no research has been done after a probabilistic imputation method. Therefore, a novel imputation method is proposed in this paper which imputes with probabilistic data. This would result in a dataset where the uncertainty measures are integrated. It is foreseen that this should improve the data quality and the validity of additional data processing. Consequently, this research aims to create aforementioned method and will prove whether this increases the data quality.

4. METHODS OF RESEARCH

4.1 Dataset and Preprocessing

4.1.1 Dataset

The dataset that is used in this research is the *Adult Census Income* dataset taken from the *UCI Machine Learning Repository* [1]. The data is extracted from the *1994 US Census bureau database*. The data provides information on people and their occupation. The last attribute in the dataset is a class stating whether the person in question makes more or less than $50.000 a year. More information on the attributes can be found in Table 3 in Appendix A.

4.1.2 Preprocessing

The dataset consists of 32.561 records; 2.399 of these records contain missing values marked by a question mark as seen in Table 1. As missing values cannot be resolved, evaluation of imputation techniques cannot be done. Therefore, the records with missing data are removed and the remaining 30.162 records are considered the complete dataset. This is still a large enough dataset to get valid results out of the research.

Table 1. Original Adult Census Income Dataset

	age	workclass	fnlwgt	education
0	90	?	77053	HS-grad
1	82	Private	132870	HS-grad
2	66	?	186061	Some-college
3	54	Private	140359	7th-8th
4	41	Private	264663	Some-college

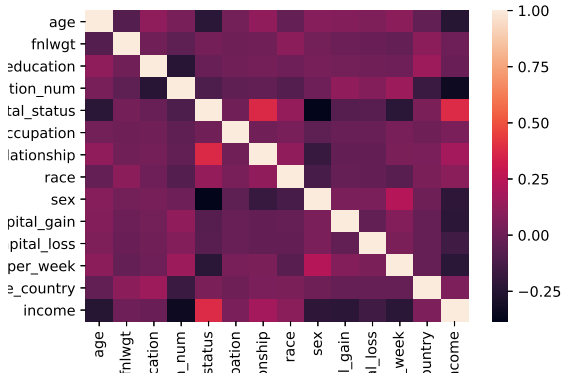

Figure 1. Heatmap of pairwise correlation.

The next part is to analyze the complete dataset to better understand it. From the pairwise correlation of columns in Figure 1, it can be concluded that e.g. the relationship and marital status have a high correlation value. I.e. information about someone's relationship gives an indication of their marital status. This means that these attributes are better for making predictions. Therefore, the comparison is based around the imputation of these attributes.

Furthermore, this research is centered around a probabilistic imputation method. Therefore, imputing a categorical attribute is more straightforward as a continuous probabilistic distribution is hard to read from and compare to. Hence, the MCAR mechanism is applied on one of the categorical attributes before comparing the imputation methods.

Finally, for smooth processing, the categorical (nonnumerical) values are replaced by a number that uniquely identifies the respective category.

4.2 Probabilistic Imputation

This research is primarily focused on finding an imputation method based on probabilistic data.

The representation of probabilistic data is the first consideration that needs to be made. For example, an extra column can be included with the probabilities and each record with missing values is copied for each plausible value. Though, in this research a different approach is

taken. For each missing value a list of plausible values with their respective probabilities is imputed instead of a single crisp value. An example of what this looks like can be found in Table 2.

Table 2. Probabilistic Imputation applied on MCAR amputed Adult Census Income Dataset

	age	workclass	fnlwgt	education
0	82	Private	132870	HS-grad
1	54	(Private, 0.7) (Self-emp-not-inc, 0.1) (Local-gov, 0.1) (State-gov, 0.1)	140359	7th-8th
2	41	Private	264663	Some-college
3	34	Private	216864	HS-grad
4	38	(Private, 0.8) (Local-gov, 0.2)	150601	10th

Imputing with probabilistic data is not difficult. However, the method to discover the probabilities is. There are multiple different possible techniques to do so and some give better results than others. In this research three techniques are discovered and used to design two probabilistic imputation methods.

4.2.1 Frequentist Probability

The first technique is the simplest. After the complete dataset has been amputated with the MCAR mechanism, a new complete dataset is created by removing the missing value records. From this dataset, the frequency of each value is considered and taken as a probability. If n_t is the total number of values and n_x is the number of times value x occurs, then the probability $P(x)$ is defined as:

$$P(x) \approx \frac{n_x}{n_t} \tag{1}$$

The missing value is then imputed with every possible value for which the probability is greater than 0.

4.2.2 Decision Tree Classification

A more advanced technique that is used in a lot of situations is decision tree classification. A decision tree is a tree structure where the nodes are the attributes. Based on the values of an attribute a branch is picked and the next node is tested. A leaf node represents the class to be decided upon.

An example of a decision tree can be found in Figure 2. A decision has to be made whether to play badminton outside. If it's for example sunny and the humidity is normal, then the decision is to play badminton.

A decision tree is constructed based on information gain. In the example *outlook* gives the highest information gain. In other words, when the *outlook* is known, the set of possible values for the other attributes is smallest. Therefore, less other attributes are required to be known in order to make a decision about the class label. When the *outlook* is sunny, *humidity* gives the highest information gain.

4.2.3 Random Forest Classification

Decision trees are a good technique to use with missing data imputation. Though, they can overfit on the training data leading to high variance. This means that the results of classifying unseen data tend to be of low quality. An

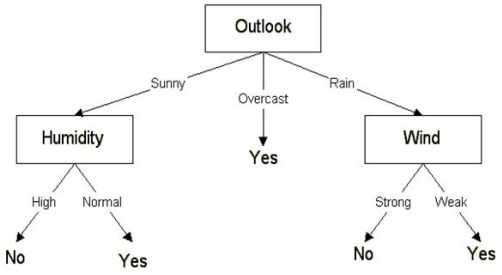

Figure 2. Decision tree example.

improvement upon decision trees is when multiple trees are combined into a forest. Each classification tree individually is called a "weak learner" while assembled they form a "strong learner".

One such a forest is a random forest [4]. It has been used in other research based around missing data imputation [11]. A random forest is constructed as follows; for T trees:

1. Select N records at random from the dataset.
2. Then for each node in the decision tree, m variables are picked at random from the predictor variables. The variable with the highest information gain is then selected as the next node.

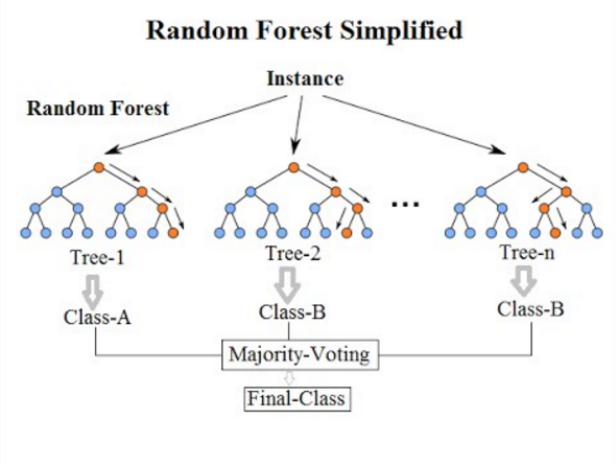

Figure 3. Visualization of random forests.

When the random forest is constructed, it is time to run through the data. Each decision tree makes a decision for a record and then the majority of the voting is decided to be the class label as seen in Figure 3.

Since each decision tree is random to some extent, random forests don't tend to overfit. Therefore, random forest classification is a good method for handling missing data.

4.3 Evaluation Methodology

To evaluate the aforementioned imputation method, an evaluation framework has been designed. The imputation method is compared against several known methods by determining the quality of the resulting dataset. The methods that are compared include: mean imputation, multiple imputation by chained equations (MICE), probabilistic imputation by frequency (PBF) and probabilistic

imputation by prediction (PBP). The PBP method uses a classifier; two classifiers are tested in this research: decision trees and the random forest classifier.

4.3.1 Metrics

First of all, a set of metrics needed to be determined which qualify the imputation methods. A good example which is often used in statistical classification is called a confusion matrix. It is a cross table constructed from the possible values for an attribute. A row represents the instances of the predicted value and the column the actual value. From a confusion matrix, a couple of metrics can be determined. The accuracy is the actual percentage of correctly predicted values. Furthermore, the precision and recall are calculated. The precision for a specific value is the percentage of true positives out of all predicted positives. The recall, on the other hand, is the percentage of true positives out of the actual positives. From the recall and precision another metric can be calculated that summarizes both called the F-measure: $\frac{2*precision*recall}{precision+recall}$; the F-measure is also used by Pantanowitz and Marwala [10]. These metrics are specifically used to determine how well the predictions of the imputation method are.

Another metric that is used in a lot of research is the root mean square error (RMSE) [2][16][20]. This is often a better measure than the confusion matrix. It is calculated as the square root of the average of squared differences between predicted and actual value.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y_j})^2} \tag{2}$$

The RMSE represents the average magnitude of the prediction error.

Furthermore, a metric is used that expresses the quality of the probabilistic distributions that the novel imputation method imputes with. It is called the Kullback-Leibler divergence [6] (or KL divergence) and it determines how close the predicted probability distribution is to the actual distribution. The formula for this metric looks as follows:

$$D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) \log(\frac{P(x)}{Q(x)}) \tag{3}$$

Here P and Q represent the probability of the imputed and actual dataset respectively.

The actual distribution consists of a probability of 1 for the actual value and 0 probability for the rest of the possible values. Thus, the probability for the actual dataset is 1. Considerably, the non-probabilistic imputation methods either impute with the correct value in which case the KL divergence is 0, i.e. there is no difference between the distributions, or it is completely incorrect. The probabilistic imputation method (almost) never picks one value as the only possible value, and thus it will lose score when a non-probabilistic method predicts correctly. Though, the probabilistic imputation method will score very well compared to other methods when these methods falsely predict.

Clearly, $D_{KL}(P \parallel Q)! = D_{KL}(Q \parallel P)$ does not (always) hold. The KL divergence is asymmetric. In order to better measure the similarity between P and Q, a symmetric form is constructed as also suggested by Deng [6]. Moreover, the actual metric that is determined looks as follows:

$$D_s(P \parallel Q) = \frac{D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P)}{2} \tag{4}$$

All in all, the metrics that will be measured conclude: RMSE, (symmetric) KL divergence, accuracy, precision and recall.

4.3.2 Simulation Technique

The evaluation framework consists of two evaluation techniques. The first being the most important. This technique runs 1000 simulations such that the result is more acceptable and valid. It is based on the method used by Schmitt [16]. It starts out with the original dataset (without missing values) and it generates missing values at random with a chance of 0.10. All imputation methods will then impute the dataset and the metrics are determined from the resulting dataset. In the end, the metrics are averaged over all simulations. The same process can be found in Figure 4.

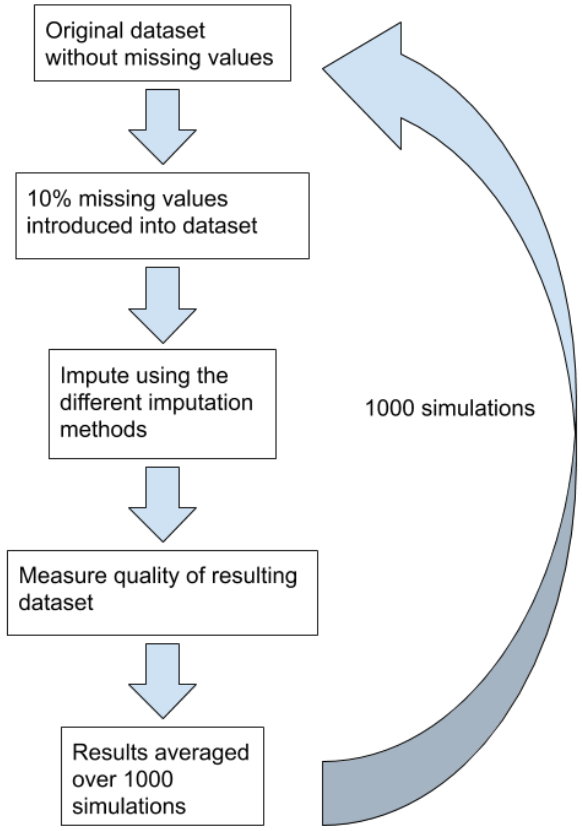

Figure 4. The simulation process [16].

4.3.3 Neural Network Classification

The second evaluation technique is a neural network. In real world situations an imputed dataset would be further processed. An example would be to classify the data. In the dataset, used in this research, a class is also available: whether a person makes more or less than $50.000. Therefore, to test whether the resulting dataset could be used

in future research, a neural network is constructed that classifies that data.

5. COMPARISON AND RESULTS

After running the simulations and the neural network on the dataset, the results are output. These can be found in Appendix B.

Some small notes have to be made first. For the probabilistic imputation technique it is difficult to determine a confusion matrix directly from the probabilistic data. Therefore, the data is reduced to the value with the highest probability after which the confusion matrix can be determined. Note that in the case of the PBF method this would result in the same confusion matrix as with a most frequent imputation method. This same reduced dataset is used to determine the RMSE. These restrictions make that the accuracy, precision and recall are not as accurate as they should be, but they should still be close enough to make sense.

5.1 Simulation Technique

From the results, a few conclusions can be drawn when looking at the known methods. The mean imputation method is considerably scoring worse than the rest. This was to be expected beforehand, but as it was unknown how well the probabilistic imputation method would work, this method had to be included. As the mean imputation method imputes with the mean of an attribute taken over the whole dataset, the accuracy is (significantly) low. It does often occur that something holds for more people when it holds for a large group, but it cannot be assumed to hold for everyone. Hence, in a lot of cases, the mean value is not the actual value. The multiple imputation method is scoring substantially better than mean imputation (a single imputation method). As multiple imputation imputes the dataset multiple times and combines the results, the resulting dataset will be less biased. This is a considerable improvement in data quality.

The first probabilistic imputation method that was constructed is: probabilistic imputation by frequency (PBF). This is one of the more straightforward probabilistic imputation methods. However, from the results it can be determined that it is scoring quite well. The RMSE is very high, meaning there are some large errors found in the dataset. Though, these errors are taken from categorical values. It determines a large error when two categories have a high difference in categorical index. However, who is to say when two categories lie far apart from each other. Still, the results are taken over a 1000 simulations, so the results should be trustworthy. It turns out that the most frequent value (in case of this dataset) is quite accurate; the PBF method has a higher accuracy than mean imputation.

The predictive probabilistic imputation method (PBP) seems to work very well. The metrics show a high quality of data for both classifiers. The random forest classifier is a (small) improvement compared to the decision tree as a random forest results in less overfitting. Though, both classifiers show a high accuracy, precision and recall. More importantly, the RMSE is very close to 0 and even lower than the RMSE that MICE is showing.

Looking more into the KL divergence, probabilistic imputation also shows to be valuable. The KL divergence in case of the mean and multiple imputation methods is already very low. This means that the imputed dataset shows (probabilistic) distributions close to the actual. Frequentist probability scores worse than expected. The reason for this is most probably that the probability distribution is more spread out. In other words, the correct value could have the highest probability (most frequent value), but a significant amount of probability is also spread across the rest of the possible values. The predictive probability imputation method has a more advanced technique to determine the probability distributions. Therefore, the correct value is given a higher probability than with PBF. This translates into the result as a low KL divergence. The probabilistic distribution is very similar to the distribution of the actual dataset.

5.2 Neural Network Classification

In Table 5 in Appendix B the results of the neural network classification are shown. A note has to be made explaining why the results are so low. The constructed neural network is quite simple. However, the results can still be compared validly. From the results, it is difficult to truly take any conclusions. Nonetheless, it can still be determined that multiple imputation again shows to work better than mean (single) imputation. Furthermore, probabilistic frequentism has results which would be equal to a most frequent imputation technique; the value with highest probability is taken as the neural network cannot analyze the probability distributions in the dataset. The predictive probabilistic imputation method shows results equivalent to MICE.

6. DISCUSSION

In data science, missing data is a known problem. There exist different ways to handle missing data and each method has their advantages and disadvantages. Deletion is usually not desirable and therefore imputation is often used. Upon single imputation, an improvement has been made when multiple imputation was discovered. Nonetheless, there are also disadvantages to such methods. Therefore, in this research a novel imputation method based on probabilistic data has been created.

The method comes with two alternatives: frequentist probability and predictive probability using a classifier. The frequentist probability method is the most straightforward and is calculates the probability of a plausible value by its frequency in the complete dataset. The predictive probabilitistic imputation method predicts the probabilities for which any classifier can be taken. However, in this research the decision tree and random forest classifier were tested. The random forest classifier in general shows less overfitting and is therefore often preferred over a decision tree classifier.

From the results, it can be resolved that the proposed imputation technique has potential. The PBF method is a simpler technique and is not preferable, but the PBP method shows very good results for both classifiers that are used. The random forest classifier is still better than a single decision tree though. At this point, MICE has been one of the better and most used methods for missing data imputation. Nonetheless, judging from the results, the predictive probabilistic imputation method would work just as good or even better. The most important aspect of probabilistic imputation is that the resulting dataset is not assumed to be the real dataset, but the probabilities

for each value are still available. This way, further processing on the data can also conclude its results with a certain probability.

7. CONCLUSION

An incomplete dataset should be handled correctly to eliminate missing values and generate a complete dataset. There exist several imputation methods that predict a plausible value to fill in the missing data. From research, it has been concluded that multiple imputation has been one of the most successful imputation methods. However, there are still some issues with such a method. One of these is that the resulting dataset is assumed to be real and further research might conceive biased results because of this reason. Therefore, in this research, a novel imputation method based on probabilistic data has been developed (**RQ 1.**). Two methods were determined which measure the probability distribution in different ways (**RQ 1.1.**). The first method bases the probability of a plausible value on its frequency in the complete dataset. The second, which is an improvement, makes use of a predictive classifier. A decision tree classifier and a random forest classifier have been tested.

The probabilistic imputation method had to be qualified by an evaluation framework (**RQ 2.1.**). A simulation technique was developed that tried out mean imputation, MICE and the probabilistic imputation method on a dataset affected by the MCAR mechanism. Additionally, a neural network was setup to create a real life data processing situation and discover how well the imputed dataset could be further processed. These evaluation techniques generate a set of measurements based on the following metrics: RMSE, KL divergence, accuracy, precision and recall (**RQ 2.2.**). From the end results it has been determined that the probabilistic imputation method could work as good, or even better, than MICE (**RQ 2.**). Especially, the predictive alternative (PBP) generated very good results. The random forest classifier showed substantially better results as it tends to overfit less on the training data. Likewise, the probabilistic data introduced into the dataset could be used as an uncertainty measure in further research on the data.

7.1 Future Work

To really take advantage out of this novel imputation method, other techniques might be discovered that could measure uncertainty. Furthermore, there should be a better way to construct a confusion matrix for a probabilistic dataset. The current method takes the values out of the dataset with the highest probability to compose a non-probabilistic dataset. However, all probabilistic data should be taken into account to get a more valid measure out of it.

8. REFERENCES

[1] Adult census income. https://www.kaggle.com/uciml/adult-census-income, October 2016. UCI Machine Learning.

[2] L. Asiedu. Comparison of imputation methods for missing values in longitudinal data under missing completely at random (mcar) mechanism. *African Journal of Applied Statistics*, 4:241–258, January 2017.

[3] A. Baraldi and C. Enders. An introduction to modern missing data analyses. *Journal of School Psychology*, 48(2010):5–37, 2009.

[4] G. Biau and E. Scornet. A random forest guided tour.

[5] D. N. Davis and M. Rahman. Missing value imputation using stratified supervised learning for cardiovascular data. *Journal of Informatics and Data Mining*, 1, January 2016.

[6] J. Deng, Y. Wang, J. Guo, Y. Deng, J. Gao, and Y. Park. A similarity measure based on kullbackâĂŞleibler divergence for collaborative filtering in sparse data. *Journal of Information Science*, October 2018.

[7] C. Enders. *Applied missing data analysis*. The Guilford Press, New York, NY, US, 2010.

[8] J. W. Graham. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60(1):549–576, 2009.

[9] H. Kang. The prevention and handling of the missing data. *Korean J. Anesthesiol*, 64:402–406, 2013.

[10] P. Lodder. To impute or not impute: ThatâĂŹs the question. *University of Amsterdam, Amsterdam, The Netherlands*, 2013.

[11] A. Pantanowitz and T. Marwala. Evaluating the impact of missing data imputation through the use of the random forest algorithm. *School of Electrical and Information Engineering, University of the Witwatersrand, Private Bag X3, Wits, 2050, Republic of South Africa*.

[12] A. Pedersen, E. Mikkelsen, D. Cronin-Fenton, et al. Missing data and multiple imputation in clinical epidemiological research. *Psychological methods*, 7(2):147, 2017.

[13] J. Poulos and R. Valle. Missing data imputation for supervised learning. *University of California, Berkeley*, August 2018.

[14] M. Richman, T. Trafalis, and I. Adrianto. Missing data imputation through machine learning algorithms. *Artificial Intelligence Methods in the Environmental Sciences*, pages 153–169, January 2009.

[15] J. Schafer and J. Graham. Missing data: our view of the state of the art. 2002.

[16] P. Schmitt, J. Mandel, and M. Guedj. A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics*, 6:224, 2015.

[17] J. Sterne et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 2009.

[18] S. van Buuren. *Flexible Imputation of Missing Data*. CRC/Chapman & Hall, Boca Raton, FL, US, 2 edition, 2018.

[19] M. van Keulen. Probabilistic data integration. *Faculty of EEMCS, University of Twente, Enschede, The Netherlands*, 2018.

[20] C. Yozgatligil, S. Aslan, C. Iyigun, and I. Batmaz. Comparison of missing value imputation methods in time series: The case of turkish meteorological data. *Theoretical and Applied Climatology*, 112, April 2012.

APPENDIX

A. DATASET ATTRIBUTES

Table 3. Dataset Attributes

Name	Type	Categorical values
Age	Continuous	-
Workclass	Categorical	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Fnlwgt (final weight)	Continuous	-
Education	Categorical	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
Education num	Continuous	-
Marital status	Categorical	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Categorical	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
Relationship	Categorical	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Race	Categorical	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Sex	Categorical	Female, Male
Capital gain	Continuous	-
Capital loss	Continuous	-
Hours per week	Continuous	-
Native country	Categorical	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
Income	Categorical	>50K, <=50K

B. RESULTS

Table 4. Results averaged over 1000 simulations

	RMSE	Symmetric KL-divergence	Accuracy	Precision	Recall
mean	0.082786	0.003873	0.913951	0.945555	0.914594
mice	0.075339	0.001119	0.953263	0.967634	0.922565
pbf	0.108020	0.002235	0.946618	0.985327	0.914605
pbp:decision_tree	0.062567	0.000231	0.978631	0.936916	0.939743
pbp:random_forest	0.054183	0.000261	0.982776	0.977503	0.941090

Table 5. Neural network classification results

	Accuracy	Precision	Recall	F-measure
mean	0.858066	0.814437	0.796517	0.805377
mice	0.864896	0.822560	0.809524	0.815990
pbf	0.859890	0.817208	0.798488	0.807740
pbp:decision_tree	0.864399	0.821413	0.810083	0.815709
pbp:random_forest	0.864697	0.821721	0.810772	0.816209