# Evaluating the Influence of Missing Data on Classification Algorithms in Data Mining Applications

**Luciano C. Blomberg[1], Duncan Dubugras A. Ruiz[1]**

[1]PPGCC – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Porto Alegre – RS – Brazil

luciano.blomberg@acad.pucrs.br, duncan.ruiz@pucrs.br

***Abstract.*** *This paper presents an analysis regarding the influence of missing data on datasets when submitted to traditional classification algorithms in data mining applications. For this purpose, we use ten UCI datasets and manipulate them to hold controlled levels of missing data. Our empirical analysis shows that the classification performance decreases after significant insertion of missing values in all datasets tested. Among the analyzed algorithms, Naïve Bayes is the least influenced by missing data, being SMO the next. IBK is the most influenced, presenting the lowest accuracy, predominantly in datasets whose independent variables are continuous.*

## 1. Introduction

Much has been discussed in machine learning literature about the data quality importance to classification models induction. A common quality problem relates to lack of complete data which can be caused by many situations. In health area, for example, this kind of error is usually produced by malfunctioning machines or refusal of patients to answer personal questions (e.g.: income, weight, age), producing large amount of missing data in a medical record. Regardless of the reason or domain area, missing data has generated serious problems in the knowledge extraction, hiding important information about the dataset, skewing results and affecting the accuracy of induced models [Jonsson and Wohlin 2004], [Liu et al. 2005], [Nogueira et al. 2007] and [Zhang et al. 2010].

According to Acuna and Rodriguez [2004], rates of less than 1% missing data are generally considered trivial, and 1-5% considered manageable. However, 5-15% requires sophisticated methods to handle them, and more than 15% may severely impact any kind of interpretation. In this way, a great concern of the scientific community has involved the development of strategies to deal with low quality data.

Even though dealing with missing data is a typical problem of data pre-processing, a considerable amount of machine learning algorithms has implemented its own strategy for dealing with missing data. SMO, for example, replaces the missing values with means or modes. Other machine learning methods are inherently tolerant to incomplete data and, thus, require no specialized mechanisms for handling missing values, such as Naïve Bayes, which simply skips over missing values when computing distributions of attribute [Kalousis and Hilario 2000].

Furthermore, distribution of missing data is another aspect that may influence the effectiveness of classification algorithms. Litle and Rubin [1987] presented three different mechanisms by which the missing data are distributed:

- MCAR (Missing Completely at Random): the probability of an instance to have a missing value for an attribute does not depend on any other value from the dataset.
- MAR (Missing at Random): the probability of an instance to have a missing value for an attribute depends on some other value from the dataset.
- NMAR (Not Missing at Random): the probability that a value is missing does depend on the value(s) of the target variable to be imputed, and possibly also on the values of auxiliary variables.

In NMAR, specially, the use of strategies based on means cannot generate good results. For example, when low income individuals have less probability to inform their income, using means to fill the missing data could create bias in favour of high income individuals. To better describe this situation, we use Tables 1, 2 and 3, where the last of them contains two records with missing data on income attribute (id 3 and 5). These attributes are imputed by averaging the present values (id 1, 2 and 4), thus creating a distortion of the true values.

**Table 1. True Income**

| Id | Income | Class |
|---|---|---|
| 1 | 12000 | Y |
| 2 | 9000 | X |
| 3 | 850 | X |
| 4 | 12000 | X |
| 5 | 675 | Y |

**Table 2. Missing Income**

| Id | Income | Class |
|---|---|---|
| 1 | 15000 | Y |
| 2 | 9000 | X |
| 3 | | X |
| 4 | 12000 | X |
| 5 | | Y |

**Table 3. Imputed Income**

| Id | Income | Class |
|---|---|---|
| 1 | 15000 | Y |
| 2 | 9000 | X |
| 3 | 12000 | X |
| 4 | 12000 | X |
| 5 | 12000 | Y |

In this paper we analyze the missing data influence on five classifiers, as follows: J48 (Weka implementation to C4.5 decision trees method) IBK (Weka implementation to k-nearest neighbors method), MLP (Weka implementation to neural networks method), SMO (Weka implementation to support vector machines method) and Naïve Bayes (Weka implementation to bayesian networks method). For this purpose, we used UCI data sets, such as Acute, Breast, Stalog (Heart), Tic-tac-toe, Blood,Iris, Wine, Hayes, Glass and Teaching, inserting controlled levels of missing data (1% to 50%), according to MCAR mechanism.

To insert missing data we used a combination of the INT() and RAND() functions of the tool Microsoft Excel, so that the address of the cell to be artificially produced was defined by randomly generating of values to row and column.

For better understanding, let us to consider a blood dataset example with 11x5 (row x col). Assuming that it is not desirable to insert missing data in the first row (header) and either in the last column (class attribute), the following formula would be necessary to produce 10 % of missing values into R, F, M and T attributes:

=INT ((RAND() * (11-1)) + 2): Generating values from 2 to 11 for the row.
=INT ((RAND() * (5-1)) + 1): Generating values from 1 to 4 for the column.

Once generated random values for row and column, the columns values were replaced by letters, and concatenated with the rows values, thus producing the cell address to be removed, as shown in Figure 1.

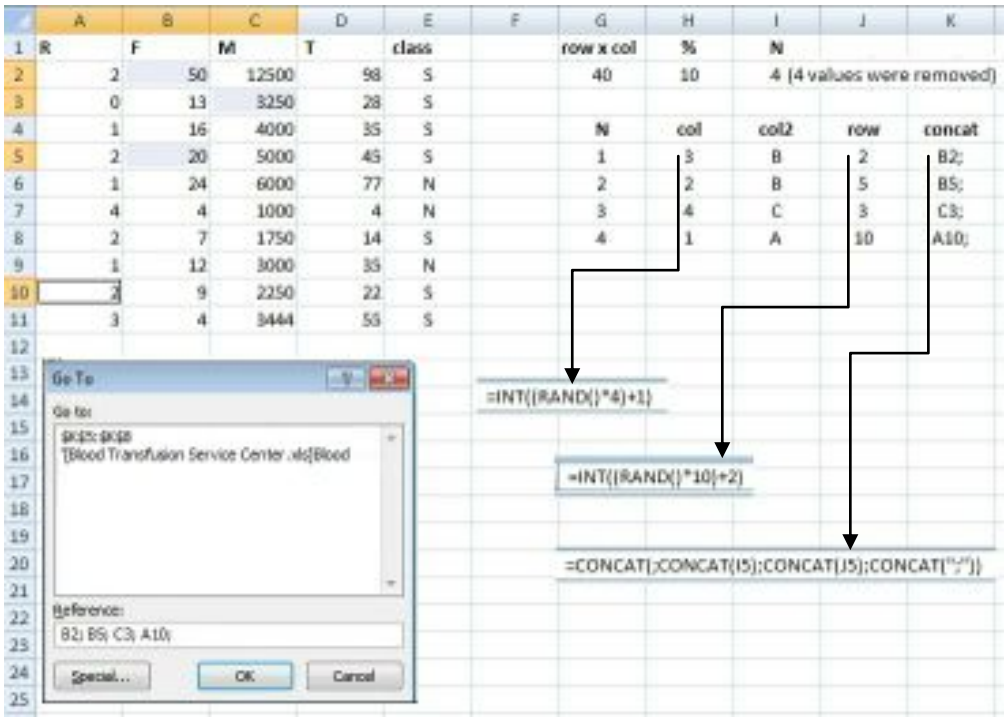


**Figure 1. Inserting missing data**

## 2. Internal Strategies for Missing Data Treatment in Classification Algorithms

As most statistics or machine learning methods, classifiers in data mining are designed to assume that data are complete. However, datasets produced by real world applications frequently include noisy, incomplete and inconsistent data. This fact has highlighted the importance of developing new strategies to handle missing data, or the implementation of robust algorithms to deal with this kind of data quality problem.

J48 algorithm is a traditional decision tree method to classification in data mining applications. Its internal strategy to handle missing data involves splitting the instance into pieces, using a numeric weighting method, and sending part of it down each branch in proportion to the number of training instances going down that branch. Eventually, the various parts of the instance will reach a leaf node each, and the decisions at these leaf nodes must be recombined using the weights that have percolated to the leaves [Witten and Frank 2005].

Other algorithms just ignore attributes with missing values, such as MLP and Naïve Bayes. Though MLP can solve complex and non-linear problems and it has been used widely, it cannot handle missing data directly [Chang and Shin 2006]. Naïve Bayes, on the other hand, optimizes the model over the whole dimensionality, and is capable of learning even in presence of some missing values [Shi and Liu 2011].

Regarding IBK, this algorithm conventionally calculates distances between examples as if all of their values were known. In general, if the value of a given attribute A is missing in tuple $X_1$ and/or in tuple $X_2$, we assume the maximum possible difference. Suppose that each of the attributes has been mapped to the range [0,1]. For nominal attributes, we take the difference value to be 1 if either one or both of the corresponding values of A are missing. If A is a numeric and missing from both tuples $X_1$ ad $X_2$, then the difference is also taken to be 1. If only one value is missing and other (which we will call v′) is present and normalized, then we can take the difference to be either |1-v′| or |0-v′|, whichever is greater [Han and Kamber 2011]. SMO implements the sequential minimal optimization algorithm in order to train a support vector classifier doing global replacement of missing values with means or modes [Witten and Frank 2005].

An important question related to the efficacy of internal strategies on classification algorithms regards to the characteristics of the datasets analyzed. As it is known, some characteristics such as number of instances, number of attributes, classes' quantity and balancing of these classes are able of affect the performance of any classification algorithm. When we handle datasets with missing values, other characteristics such as attribute correlation, mechanisms of distributing missing data and the percentage of these data in the whole dataset makes this task even more complex. This way, to apply an ideal procedure to compare the performance of classification algorithms becomes an extremely onerous task.

## 3. Empirical Analysis

The analysis presented in this paper are based on individual and average accuracy of five traditional classification algorithms applied on datasets with missing data. For this purpose, we used J48, SMO, MLP, IBK and Naïve Bayes algorithms, and applied them on ten datasets from the UCI Repository, as shown in Table 4.

**Table 4. Experimental Datasets**

| Datasets | Instances | Non-terminal attributes | Attribute Characteristics | Number of Classes |
|---|---|---|---|---|
| Acute | 120 | 7 | Categorical, Integer | 2 |
| Tic-tac-toe | 958 | 9 | Categorical | 2 |
| Stalog | 270 | 13 | Categorical, Real | 2 |
| Hayes | 160 | 5 | Categorical | 3 |
| Teaching | 151 | 5 | Categorical, Integer | 3 |
| Wine | 178 | 13 | Integer, Real | 3 |
| Glass | 214 | 10 | Real | 7 |
| Iris | 150 | 4 | Real | 3 |
| Breast | 699 | 10 | Integer | 2 |
| Blood | 748 | 4 | Real | 2 |

Aiming to make more controlled experiments, we sought to analyze only datasets with specific profile of size and dimensionality. In this way, only complete

datasets with up to 958 instances and 13 non-terminal attributes were analyzed. Missing data were artificially inserted on these datasets in the following percentages: 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45% and 50%. For that, MCAR mechanism was used.

In order to evaluate the performance of the classification algorithms, we analyzed the accuracy behavior with the increase of missing data percentage for each dataset, evaluating them individually and on the prediction's average. This way, we applied the default parameters from machine learning tool Weka 3.7.4 and ten-fold cross-validation, producing the following results, presented in Figures 2 to 11.

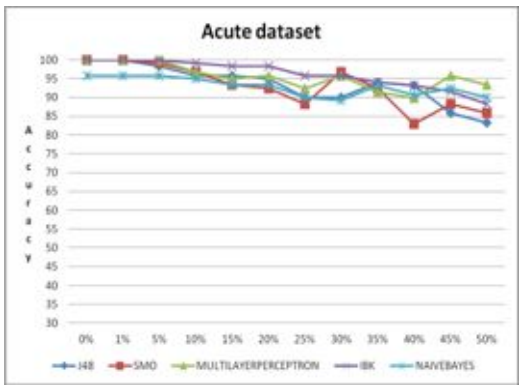


**Figure 2. Results of Acute dataset**

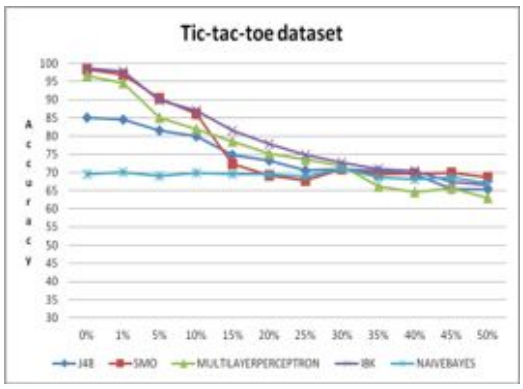


**Figure 3. Results of Tic-tac-toe dataset**

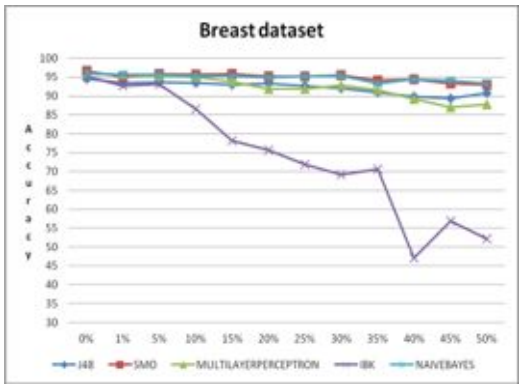


**Figure 4. Results of Breast dataset**

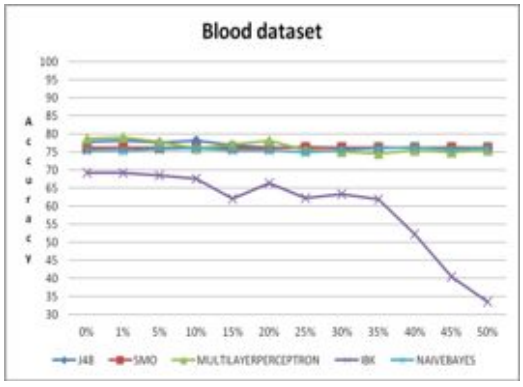


**Figure 5. Results of Blood dataset**

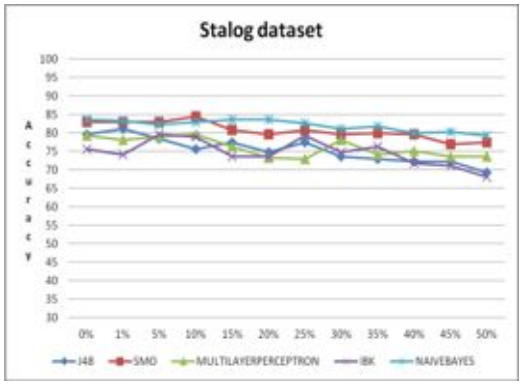


**Figure 6. Results of Stalog dataset**

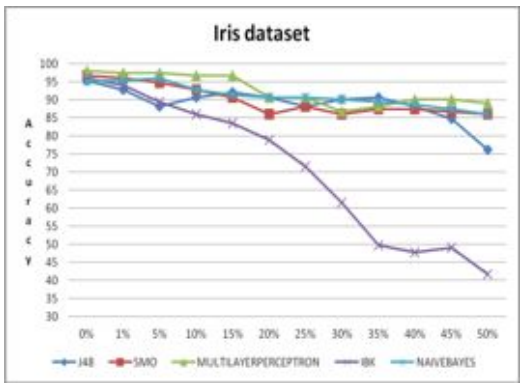


**Figure 7. Results of Iris dataset**

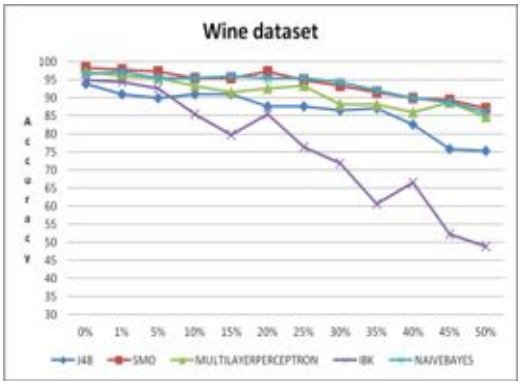


**Figure 8. Results of Wine dataset**

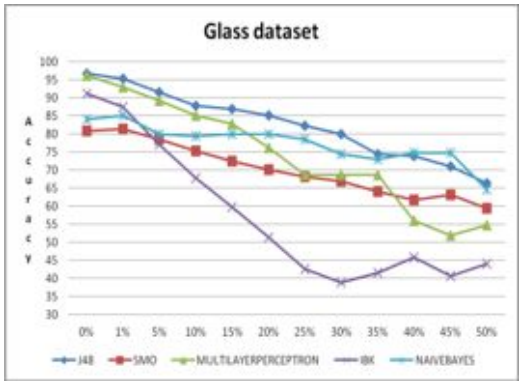


**Figure 9. Results of Glass dataset**

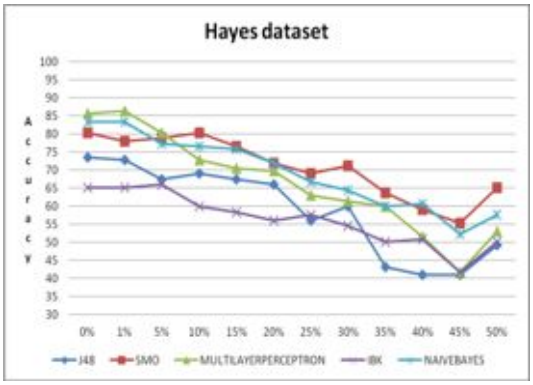


**Figure 10. Results of Hayes dataset**

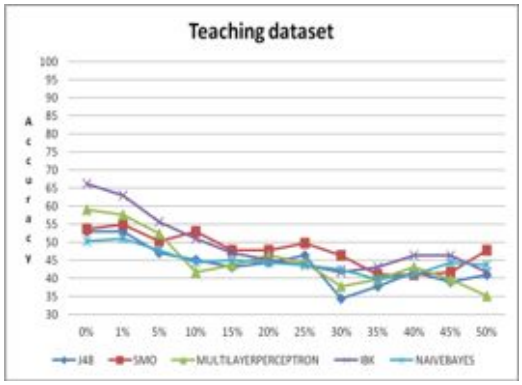


**Figure 11. Results of Teaching dataset**

The individual graphs illustrated in Figures 2 to 11 show that all classifiers presented loss of performance in the prediction's accuracy with the increase of missing data percentage. As we can see, it was not identified an algorithm that presents the best performance in all aspects. In most datasets, SMO and Naïve Bayes presented better performance than other classifiers, considering the higher missing data percentage analyzed.

On the other hand, IBK was the algorithm with the worst performance in at least 5 analyzed datasets (Glass, Wine, Iris, Blood and Breast). In these datasets, the non-terminal attributes were predominant continuous. Another interesting aspect is the MLP performance, which in most of the datasets (Hayes, Wine, Iris, Acute, Stalog and Teaching) was superior or equivalent to J48, even ignoring missing data. In Table 5, we present the prediction's average values of the classifiers, considering all datasets. These values can also be observed in the graph showed in Figure 12.

**Table 5. Prediction's Average**

| Classifiers | 0% | 1% | 5% | 10% | 15% | 20% | 25% | 30% | 35% | 40% | 45% | 50% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| J48 | 85 | 84 | 81 | 81 | 80 | 79 | 77 | 75 | 74 | 73 | 70 | 69 |
| SMO | 86 | 86 | 84 | 84 | 80 | 79 | 78 | 78 | 76 | 74 | 74 | 75 |
| MLP | 89 | 88 | 85 | 82 | 81 | 79 | 77 | 76 | 74 | 72 | 71 | 71 |
| IBK | 85 | 84 | 81 | 77 | 72 | 71 | 68 | 64 | 62 | 59 | 56 | 53 |
| Naïve Bayes | 83 | 83 | 81 | 81 | 81 | 80 | 79 | 78 | 77 | 76 | 76 | 74 |

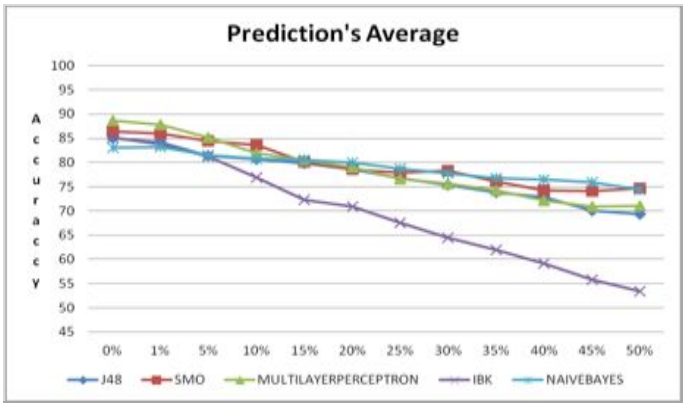


**Figure 12. Prediction's average**

The results presented in Figure 12 show that in average IBK algorithm had the worst performance among the five classifiers. This difference is accentuated from 10% of missing data. However, in percentages up to 5% of missing data, the IBK algorithm has shown competitive results with others algorithms, such as J48 and Naïve Bayes.

Among the five classifiers, SMO and Naïve Bayes algorithms are the most accurate to missing data. SMO algorithm is the one with better performance (in average) for high missing data percentage. However, Naïve Bayes algorithm shows equivalent performance to SMO after 15% of missing data percentage.

In order to obtain the variation of the prediction's average for all of them classifiers with similar performance, we use the population variance measure. Considering that for each classifier we apply the prediction's average of all datasets taking $N$ variations according to the missing data percentage (in this case, 12), the population variance of $yi$ is given by:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(yi - \mu)^2$$

where $i = 1, 2, ...., N$, and $\mu$ represents the mean of prediction's average for each classifier. The population variances obtained are shown in Table 6.

**Table 6. Population variances**

| Classifiers | μ | σ² |
|---|---|---|
| Naïve Bayes | 79.07 | 7.59 |
| SMO | 79.50 | 19.02 |
| J48 | 77.31 | 24.55 |
| MLP | 78.64 | 36.19 |
| IBK | 69.37 | 107.52 |

Regarding to the variation of these performances, major differences were not identified between the algorithms, except by IBK classification algorithm, that was the worst of them. Among the five classifiers, Naïve Bayes algorithm was the least influenced by the increase in missing data percentage followed by SMO and J48.

As is well known, there are a vast number of studies in the literature around to treatment of missing data and their effects on the predictive accuracy to classification algorithms. However, most of these studies are focused on preprocessing strategies for

dealing with missing data (e.g. case deletion, imputation) [Acuna and Rodriguez 2004] [Song et al. 2008] [Farhangfar et al. 2008] [Su et al. 2008] [Luengo and Herrera 2012]. This way, there is a few number of studies that evaluate the influence of internal strategies on datasets exclusively incomplete and with controlled percentage of missing data.

Among the most related studies, we have highlighted the work of [Liu, Lei and Wu 2005], in which the influence of missing data was investigated to six representative classifiers, considering for that, 10 UCI datasets. Although the Liu, Lei and Wu (2005) study has converged with this work in the definition of the most and least sensitive method to handling missing data (k-Nearest Neighbours and Naive Bayesian classifiers respectively), more details on the insertion process of missing data as well as an objective criterion (nonvisual) for evaluation of sensitivity variation are needed.

## 4. Final Remarks

Missing data is a common problem in datasets produced by real world applications. High percentages of missing data may influence considerably the process of extraction of knowledge, reducing the accuracy of classification algorithms.

Although part of the classification algorithms implement some internal strategy for dealing with missing data, there is no universal method for all possible configurations for a dataset.

This paper evaluates the performance of five traditional classification algorithms from public datasets with controlled levels of missing data artificially inserted. In order to make more controlled experiments, we selected only complete datasets with up to 958 instances, 13 non-terminal attributes according to MCAR mechanism.

Among the five classifiers, Naïve Bayes algorithm is the least influenced by the increase in missing data percentage, being SMO the next. Regarding to prediction's accuracy, SMO is the one with better results for high missing data percentage. Naïve Bayes is the next, especially, after 15% of missing data percentage. IBK algorithm is the one with the worst performance among them.

An important question concerns the variability in the behavior of algorithms on different datasets. There are different reasons for that. As has been reported in the literature, there is a number of particularities in a dataset that makes it more susceptible to a given algorithm and less to others. Number of instances and attributes, class attribute balance, data correlation, and noise presence are examples of such features [Hulse et al. 2010], [Espinosa et al. 2011].

As future work, we intend to evaluate the performance of traditional regression algorithms, as well as the development of a new experimental protocol that considers the correlation of the attributes with missing data with the class attribute.

## Acknowledgements

This study received support from the National Council of Scientific and Technological Development (CNPq), in accordance with the research grant MCT/CNPq70/2009.

## References


Acuna, E. and Rodriguez, C. (2004) “The treatment of missing values and its effect in the classifier accuracy”, In Classification, Clustering and Data Mining Applications, p. 639-648.

Chang, W. and Shin, J. (2006). “Missing data handling in multi-layer perceptron.” In Proceedings of the 10th WSEAS international conference on Computers, Stevens Point, Wisconsin, USA, p. 640-645.

Espinosa, R., Zubcoff, J. and Mazón, J.N. (2011). “A set of experiments to consider data quality criteria in classification techniques for data mining”. In Proceedings of the 2011 international conference on Computational science and its applications (ICCSA'11), Berlin, Heidelberg, p. 680-694.

Farhangfar, A., Kurgan, L., Dy, J. (2008) “Impact of imputation of missing values on classification error for discrete data”. In Pattern Recognition. 41(12), p.3692-3705.

Han, J. and Kamber, M., (2011). “Data Mining: concepts and techniques”. Morgan Kaufmann, San Francisco, USA.

Hulse, J.V., Khoshgoftaar, T.M. and Napolitano, A. (2011). "Evaluating the Impact of Data Quality on Sampling," In Journal of Information & Knowledge Management (JIKM). 10(03), p. 225-245.

Jonsson, P. and Wohlin, C. (2004) “An Evaluation of k-Nearest Neighbour Imputation Using Likert Data”, In Proceedings of the Software Metrics, 10th International Symposium. IEEE Computer Society, Washington, DC, USA, p.108-118.

Kalousis, A. and Hilario, M. (2000) “Supervised knowledge discovery from incomplete data”. Cambridge,UK, In Proceedings of the 2nd International Conference on Data Mining 2000, WIT Press.

Litle, R.J.A. and Rubin, D.B (1987) “Statistical analysis with missing data, Wiley Series in probability and statistics”, Wiley, New York.

Liu, P., Lei, L. and Rubin, D.B. (2005) “A Quantitative Study of the Effect of Missing Data in Classifiers”, In Proceedings of the The Fifth International Conference on Computer and Information Technology. IEEE Computer Society, Washington, DC, USA, p.28-33.

Luengo, J., Garcia, S., Herrera, F. (2012) “On the choice of the best imputation methods for missing values considering three groups of classification methods”. In Knowl. Inf. Syst. 32(1), p.77-108.

Nogueira, B.M., Santos, T.R.A. and Zarete, L.E. (2007) “Comparison of Classifiers Efficiency on Missing Values Recovering: Application in a Marketing Database with Massive Missing Data”, In Computational Intelligence and Data Mining, p. 66-72.

Shi, H. and Liu, Y. (2011). “Naïve bayes vs. support vector machine: resilience to missing data.” In Proceedings of the Third international conference on Artificial intelligence and computational intelligence, Springer-Verlag, Berlin, Heidelberg, p. 680-687.

Song, Q., Shepperd, M., Chen, X., Liu, J. (2008) “Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation”. In Journal of Systems and Software. 81(12), p.2361-2370.

Su, X., Khoshgoftaar, T.M. and Greiner, R. (2008) “Using imputation techniques to help learn accurate classifiers”. In 20th IEEE International Conference on Tools with Artificial Intelligence. IEEE Computer Society, Washington, DC, USA, p.437–444.

Witten, I. and Frank, E. (2005) “Data mining: pratical machine learning tools and techniques”,Morgan Kaufmann, San Francisco.

Zhang, S., Wu, X. and Zhu, M (2010) “Efficient missing data imputation for supervised learning”, In Proceedings of IEEE ICCI, p. 672-679.