International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Low Percentage Missing Imputation using KNN, NB and DT

Abdullah Hussein Al-Amoodi

Universiti Pendidikan Sultan Idris, Tanjung Malim, Perak, Malaysia

Abstract: ***The objective of this research is to test data imputation for Missing data over 7 cases. Different machine learning algorithms to impute the missing data were tested and evaluated: K-nearest Neighbor (KNN), Naïve Bayes (NB) and Decision Tree (DT). Evaluation was done using t-test for the experiment with different configurations (i.e. 5%, 10% missing). The result of the experiment shows that KNN has scored better results compared with Naïve Bayes and Decision Tree. In conclusion, it is clear that machine learning algorithms can be used for missing data imputation. The implications of this research shows promising potentials for the utilization of KNN.***

Keywords: Missing Data, Imputation, and Machine Learning Imputation

1. Introduction

Missing data is an inevitable occurrence associated with the data collection process [1] especially when the data collected are huge and contains large number of inputs [2]. This issue can cause several drawbacks affecting the findings later on. Among the drawbacks of the missing data comes the possibility of bias findings [3], [4], reducing the sample size [5], excluding data [6] and the inability to understand changes in the data [7]. However, missing data should be taken into account [8] specially when dealing with repeated measurement [9]. The importance of dealing with the missing data should begin during the data collection stage [10], and all suitable environments should be setup in advance to encourage participants to fill up the data efficiently and reduce the ratio of missing occurrences. Missing data thus can be handled by means of statistical procedures [10], by means of machine learning [11], or by elimination.

2. Machine Learning Imputation

This section elaborates for the imputation of 2 different cases for missing imitation using 2 percentages (i.e. 5 and 10%)

2.1 First Case with 5 Percentage

This section includes imputing last 5% of missing data as in in sample case. The elimination of records are from seven attributes (i.e. 1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th}, 6^{th}, and 7^{th}). T-test was used for the comparison. Results are shown in Table 1

Table 1: 5% Comparison 1st iteration

		Decision Tree	KNN	Naïve Bayes
1^{st} Iteration	1ST	0.103105874	0.407315603	0.412581746
	2ND	0.087899271	0.338726097	0.351889969
	3RD	0.070189034	0.074757163	0.100009231
	4TH	0.304684349	0.063528363	0.130200824
	5TH	0.105882692	0.477877506	0.210698982
	6TH	0.029449176	0.163531894	0.065076439
	7TH	0.242767976	0.308840387	0.025377435

As seen from Table 1, seven sample cases were selected for missing data making then imputation; 1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th}, 6^{th}, and 7^{th} with two cases utilized (i.e. ***1^st^*** and ***2^nd^*** iteration). For iteration 1, 1^{st} presented no significance difference by (DT, KNN, and NB), the same result in 2^{nd} presented no significant differences. For 3^{rd}, neither DT, KNN nor NB exhibited any significance difference results. 4^{th} also presented no significance differences, and the same applied for 5^{th} results. However, for 6^{th} results, a significance difference was observed in DT with (*P-value=0.029449176*), and the rest for 6^{th} presented no significance differences. In the last iteration of 7^{th}, NB was the only one with significance difference with (*P-value=0.025377435*).

Table 2: 5% Comparison 2^{nd} iteration

		KNN	Naïve Bayes	Decision Tree
	Decision Tree			
2^{nd} Iteration	1ST	0.103105874	0.366477546	0.39917499
	2ND	0.087899271	0.463847813	0.185608379
	3RD	0.070189034	0.071810193	0.132498362
	4TH	0.304684349	0.04368849	0.365979715
	5TH	0.105882692	0.464188265	0.210698982
	6TH	0.029449176	0.186567759	0.04683767
	7TH	0.242767976	0.429082277	0.05130544

As for the second iteration from Table 2, 1^{st}, 2^{nd}, and 3^{rd} exhibited no significance differences. As for 4^{th}, one significance difference was observed in KNN with (*P-value=0.04368849*). 5^{th} imputation presented no significance difference. For 6^{th}, two significance difference results were observed, in NB with (*P-value=0.04683767*), and in DT with (*P-value=0.029449176*). Finally for 7^{th}, no significance difference were presented.

2.2 Second Case with 10 Percentage

This section includes imputing last 10% of missing data as in in sample case. The elimination of records are from seven cases (i.e. 1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th}, 6^{th}, and 7^{th}). T-test was used for the comparison. Results are shown in
Table **3**

Table 3: 10% Comparison 1^{st} iteration

		Decision Tree	KNN	Naïve Bayes
1^{st} Iteration	1ST	0.276868052	0.462530697	0.427979664
	2ND	0.036096948	0.126956913	0.124287242
	3RD	0.138955963	0.183959441	0.458934482

	4TH	0.057684758	0.30040064	0.5
	5TH	0.010244131	0.15979235	0.205467514
	6TH	0.257437424	0.107619948	0.143787026
	7TH	0.490124765	0.390523699	0.205741847

In the 1st iteration of 10% from

Table **3**, two significance differences were identified when the imputation was done across seven cases. The first was observed on 2ND DT with (*P-value=0.036096948*), and last was 5TH DT with (*P-value=0.010244131*).For the other cases when the imputation was performed on one and two cases, it presented no significance differences.

Table 4: 10% Comparison 2nd iteration

		Decision Tree	KNN	Naïve Bayes
2nd Iteration	1ST	0.40531919	0.170359565	0.430735986
	2ND	0.102078752	0.368683082	0.209312416
	3RD	0.096702504	0.355238577	0.105169512
	4TH	0.054500683	0.47392108	0.064642625
	5TH	0.338756765	0.338283967	0.223678697
	6TH	0.053657276	0.155926943	0.152043784
	7TH	0.191277083	0.013199548	0.014841645

As seen from Table 4, the remaining two significance differences were identified when the imputation was done on seven attributes. The first was observed on M15 DT with (*P-value=0.036096948*), and last was M24 DT with (*P-value=0.010244131*).For the other cases when the imputation was performed on one and two cases, it presented no significance differences.

3. Discussion

The aim of this section is to determine across the previous cases which of the MLs scored highest by having least significant differences (i.e. using T-test) when compared with the original data before artificial missing making procedure. This process is applied to all the cases identified in next sub sections

3.1 5% Comparison

This section introduce the achievement results of the 5%.

Imputation for the all machine learning algorithms used (i.e. Decision Tree, K-Nearest Neighbor, and Naïve Bayes). The section also discusses how each of the MLs scored in different attempts. The results of the comparison (i.e. using T-test) are listed in Table 5. It should be noted here that the total of imputation attempts is (*n=28*) resulted from counting iterations starting with first all the way till seventh.

Table 5: 5% Overall Comparison

ML	1st	2nd	3rd	4th	5th	6th	7th	28
Decision Tree	1	2	2	3	5	5	6	24
KNN	1	2	3	4	5	6	7	28

Naïve Bayes	1	2	3	3	5	6	6	26
1st Iteration								
ML	1st	2nd	3rd	4th	5th	6th	7th	28
Decision Tree	1	2	2	3	5	5	6	24
KNN	1	2	3	4	5	6	6	27
Naïve Bayes	1	2	2	4	5	6	6	26
2nd Iteration								

In Table 5, DT has scored (*n=24/28*) with no significance differences for the total imputations cases of first iteration, similarly for iteration 2, it scored same results (*n=24/28*). For KNN, the total score is (*n=28/28*) with no significance differences for case A, as for the other case B, it scored (*n=27/28*) for the number of no significant differences. KNN presented better imputation results for both iterations. For NB, it scored the same for both cases (*n=26/28*) for the number of imputations where no significance differences were identified. The achievement scores of 5% is illustrated in Figure 1.

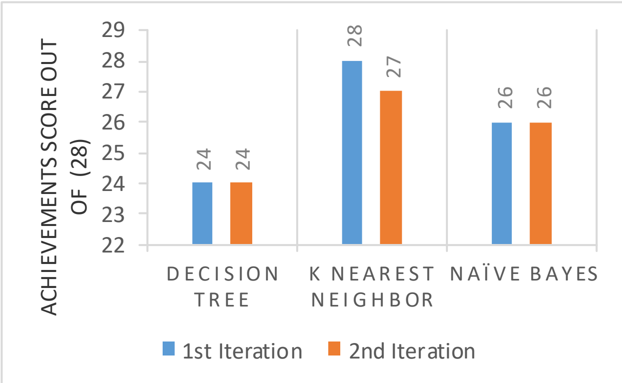

Figure 1: 5% Achievement Scores across 3 MLs

It is clear as presented in Figure 1, that there was no apparent significant difference whether in iterations 1 or 2, particularly in cases of NB and DT. However, for the case of KNN, the result is better with one point in the 1st iteration. Therefore, KNN is preferable over the remaining two experimented algorithms. This outstanding achievement of KNN is promising and suggest its capability to handle missing data

3.2 10% Comparison

This section introduce the achievement results of the 10% imputation for the all machine learning algorithms used (i.e. Decision Tree, K-Nearest Neighbor, and Naïve Bayes). The section also discusses how each of the MLs scored in different attempts. The results of the comparison (i.e. using T-test) are listed in

Table **6**. It should be noted here that the total of imputation attempts is (*n=28*) resulted from counting iterations starting with first all the way till seventh.

Table 6: 10% Overall Comparison

No of Attributes	Iterations Count							Total
	1st	2nd	3rd	4th	5th	6th	7th	28
Decision Tree	1	2	3	3	4	4	5	22

KNN	1	2	3	4	5	5	7	27
Naïve Bayes	1	2	3	4	5	5	7	27
1st Iteration								
No of Attributes	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**	**28**
Decision Tree	1	2	3	3	4	4	5	22
KNN	1	2	2	4	5	5	7	26
Naïve Bayes	1	2	3	4	2	5	7	24
2nd Iteration								

As seen in

Table **6**, DT has scored (*n=22/28*) with no significance differences for the total imputations cases presented over 1st iterations. As for other case (i.e. iteration 2), it scored same results (*n=22/28*). For KNN, the total score is (*n=27/28*) for no significance differences imputations counts for 1st iteration, as for 2nd iteration, it scored (*n=26/28*) for the number of no significant differences. KNN presented better imputation results for both iterations. For NB, it scored (*n=27/28*) for 1st iteration. However, for the second one, NB scored (n=24/28) less number of imputations where no significance differences were identified. The achievement scores of 10% is illustrated in Figure 2.

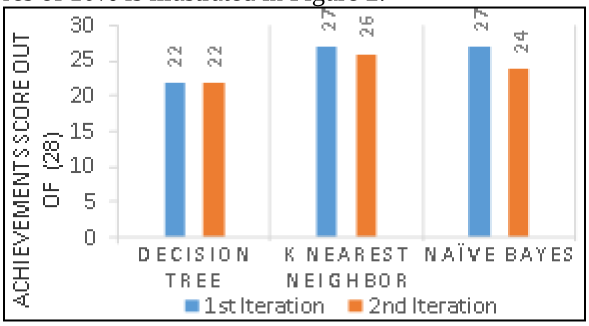

Figure 2: 10% Achievement Scores Across 3 MLs

It is clear as presented in Figure 2 shows better results for both KNN and NB in the 1st iteration. However, for DT same level of achievement was maintained which is way lower than NB and KNN. Based on these results. KNN is preferable over the remaining two experimented algorithms.

3.3 Conclusion

This study was aimed to test machine learning imputation of missing data in two cases of artificially missing data and compare their results using t-test (i.e. 5% missing and 10% missing. The results of the significance differences for t-test varies across different machine learning algorithms and different iterations, however when looking at the achievement score (i.e. non-significant values), it was clear that KNN was the highest, followed slightly by NB. The last one was decision Tree in both iterations across all configurations

It is clear that KNN was most suitable, but different samples (i.e. 1st, 2nd, 3rd, 4th, 5th, 6th, and 7th samples) of data might be suited with other MLs. Therefore, it seems proper to use KNN in other cases of missing data to enable more data findings.

References

[1] P. D. Allison, *Missing data* vol. 136: Sage publications, 2001.

[2] N. L. Crookston and A. O. Finley, "yaImpute: an R package for kNN imputation," *Journal of Statistical Software. 23 (10). 16 p.*, 2008.

[3] R. Miller, "Childhood Health and Prenatal Exposure to Seasonal Food Scarcity in Ethiopia," *World Development*, 2017.

[4] M. Vandecandelaere, S. Vansteelandt, B. De Fraine, and J. Van Damme, "The effects of early grade retention: Effect modification by prior achievement and age," *Journal of school psychology*, vol. 54, pp. 77-93, 2016.

[5] S. P. Singh, C. Winsper, D. Wolke, and A. Bryson, "School mobility and prospective pathways to psychotic-like symptoms in early adolescence: a prospective birth cohort study," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 53, pp. 518-527. e1, 2014.

[6] R. A. Gordon, A. C. Colaner, M. L. Usdansky, and C. Melgar, "Beyond an "Either–Or" approach to home-and center-based child care: Comparing children and families who combine care types with those who use just one," *Early childhood research quarterly*, vol. 28, pp. 918-935, 2013.

[7] T. M. Derrington, M. Kotelchuck, K. Plummer, H. Cabral, A. E. Lin, C. Belanoff, *et al.*, "Racial/ethnic differences in hospital use and cost among a statewide population of children with Down syndrome," *Research in developmental disabilities*, vol. 34, pp. 3276-3287, 2013.

[8] J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, *et al.*, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial intelligence in medicine*, vol. 50, pp. 105-115, 2010.

[9] L.-W. Chen, I. M. Aris, J. Y. Bernard, M.-T. Tint, A. Chia, M. Colega, *et al.*, "Associations of maternal dietary patterns during pregnancy with offspring adiposity from birth until 54 months of age," *Nutrients*, vol. 9, p. 2, 2016.

[10] P. Royston, "Multiple imputation of missing values," *The Stata Journal*, vol. 4, pp. 227-241, 2004.

[11] K. Lakshminarayan, S. A. Harp, R. P. Goldman, and T. Samad, "Imputation of Missing Data Using Machine Learning Techniques," in *KDD*, 1996, pp. 140-145.

Author Profile

Abdullah Hussein Al-Amoodi received the BSc. and MSc. degree in Information and Communication Technology from Limkokwing University, Malaysia and Master of Computer Networking from same university. Now he's a PhD Candidate at Universiti Pendidikan Sultan Idris Malaysia. His Area of Interest includes Data Science, Machine Learning and Artificial intelligence