Sakarya University Journal of Science

ISSN 1301-4048 | e-ISSN 2147-835X | Period Bimonthly | Founded: 1997 | Publisher Sakarya University |
http://www.saujs.sakarya.edu.tr/

Title: Effect Of Imputation Methods İn The Classifier Performance

Authors: Pınar Cihan, Oya Kalıpsız, Erhan Gökçe
Recieved: 2019-01-22 09:13:39

Accepted: 2019-08-23 12:22:04

Article Type: Research Article
Volume: 23
Issue: 6
Month: December
Year: 2019
Pages: 1225-1236

How to cite
Pınar Cihan, Oya Kalıpsız, Erhan Gökçe; (2019), Effect Of Imputation Methods İn The Classifier Performance. Sakarya University Journal of Science, 23(6), 1225-1236, DOI: 10.16984/saufenbilder.515716
Access link
http://www.saujs.sakarya.edu.tr/issue/44246/515716

Sakarya University Journal of Science 23(6), 1225-1236, 2019

Effect of Imputation Methods in the Classifier Performance

Pınar Cihan*[1], Oya Kalıpsız[2], Erhan Gökçe[3]

Abstract

Missing values in a dataset present an important problem for almost any traditional and modern statistical method since most of these methods were developed under the assumption that the dataset was complete. However, in the real world, no complete datasets are available and the issue of missing data is frequently encountered in veterinary field studies as in other fields. While the imputation of missing data is important in veterinary field studies where data mining is newly starting to be implemented, another important issue is how it should be imputed. This is because in many studies observations with any variables having missing values are being removed or they are completed by traditional methods. In recent years, while alternative approaches are widely available to prevent the removal of observations with missing values, they are being used rarely. The aim of this study is to examine mean, median, nearest neighbors, MICE and missForest methods to impute the simulated missing data which is the randomly removed with varying frequencies (5 to 25% by 5%) from the original veterinary dataset. Then highly accurate methods selected to impute the original dataset for observation of influence in classifier performance and to determine the optimal imputation method for the original dataset.

Keywords: missing value, multiple imputation, classification, naive bayes, decision tree, machine learning, veterinary

1. INTRODUCTION

In machine learning, classification is one of the most important tasks [1]. Many machine learning algorithms require a complete dataset and missing values lead to big classification error rates [2].

Missing values are frequently encountered in veterinary field studies. In veterinary; missing value is frequently encountered because collecting the animals for weighing, measuring and other operations or taking and analyzing blood samples are so laborious and costly. During scientific studies, it is quite possible that the researchers might be unable to collect the data in a proper way in terms of completeness due to diseases, deaths, erroneous analysis, the inappropriateness of the measured sample, etc.

* Corresponding Author: pkaya@nku.edu.tr
[1] Namık Kemal University, Computer Engineering Department, Tekirdağ, Turkey. ORCID: 0000-0001-7958-7251
[2] Yıldız Teknik University, Computer Engineering Department, Istanbul, Turkey. ORCID: 0000-0001-9553-669X
[3] Kafkas University, Internal Medicine Department, Kars, Turkey. ORCID: 0000-0003-2674-1010

There are a few alternative methods to overcome missing values in datasets: extending the data with new observations, deleting the observations with the missing value from the dataset and conducting predictions about missing value and substituting the missing value with obtained approximate values [3].

In a number of studies in this field, a complete case analysis is conducted where any observations with missing values are omitted (known as listwise deletion). Even if there are alternative approaches that prevent the removal of observations with missing values, they were rarely used in veterinary field studies. Dohoo et al. [4] have tried to determine which set of results was reliable by imputing missing values in mastitis attitudes data through both deletion and imputation techniques. Ser et al. [5], have first created 10 to 20% missing data in a dataset consisting of 47 sheep and lambs. Then, missing values were imputed using the MI method and the performance of this method was assessed.

Some studies conducted in other fields are as follows; Hron et al. [6] used version of kNN to impute the missing values in compositional data. Liao et al. [7] used four versions of kNN to estimate the missing values in large phenomic data. Tutz and Ramzan [8] used a weighted kNN to impute the missing values in several datasets. Xia et al. [9] used Adjusted Weight Voting Random Forest (AWVRF) for handling missing values. Schmitt et al. [10] compare Mean, K-nearest neighbors (KNN), fuzzy K-means (FKM), singular value decomposition (SVD), bayesian principal component analysis (bPCA) and multiple imputations by chained equations (MICE) imputation methods to impute missing values in four real datasets

The objective of this study is to examine mean, median, kNN, MICE and missForest methods in completing missing data within the veterinary dataset, their influence in classifier accuracy and to determine the optimal imputation method for the dataset. First, five different imputation methods were compared according to criteria of Root Mean Square Error (RMSE), classification error (SCE) and execution time in completing the missing values which were created from 5 to 25% by 5% non-missing part of the original dataset and the most unsuccessful methods were eliminated. Secondly, missing values in the original dataset were imputed using these successful imputation methods and the influence of these methods to the classification performance was observed. Neonatal lambs were classified according to diagnosis using naïve bayes (NB) and decision tree (DT) methods. Accuracy, kappa, recall and precision criteria were taken into consideration during the comparison.

The rest of the study is organized as follows. The second section introduces the imputation methods, datasets, principle of the analysis and evaluation measures criteria. The third section is dedicated to frequency of missingness in dataset, the imputation methods performance, the classification performance and makes a comparison. The last section provides our conclusions.

2. MATERIAL AND METHOD

2.1. Imputation Methods

We compared five commonly used imputation methods that are namely, mean, median, kNN, MICE and MissForest imputation methods. All methods implemented in R programming. Briefly, mean imputation: Mean consists of replacing the missing value for a given variable by the mean of all known values of that variable [11]. Median imputation: Replacing the missing value for a given variable by the median of all known values of that variable [11]. kNN [12]: Algorithm use distances measure such as Euclidean distance for computes the distance between the data point. The missing values are imputed by the average of the non-missing k-nearest neighbors. MICE [13] is an iterative algorithm: First, missing values are estimated using only complete data. Next, missing values are imputed using the complete data and the imputed values from the last iteration. Now, as multiple imputations create multiple predictions for each missing value; they take into account the uncertainty in the imputation and give the best standard errors. If there is not much information on the given data used to prepare the model, the imputations will be highly variable, leading to high standard errors in the analysis.

MissForest [14]: This method can be used to both impute continuous and categorical values. Given the dataset used to train the random forest model and later this model used to predict the missing values. It yields an out-of-bag imputation error estimate without the need of a test set.

2.2. Dataset

In this study used dataset was collected from 347 lambs in the two sheep flocks in Kars/Turkey. A unique ear tag number was used for registered each lamb. Blood samples, gender, birth weight, parity, health status, etc. information (given in Table 1) are recorded with this ear tag. Clinical examinations were performed as previously defined by our authors [15]. The health status of lambs was regularly monitored by daily visits to farms at the neonatal period and lambs were classified as unhealthy if have any symptoms such as mastitis, pneumonia, enteritis, etc.

2.2.1. Variable Selection from Dataset

Because of placental structure in lambs, the passage of many crucial substances primarily the antibodies from the dam to the lamb does not occur. All substances required for the prevention of diseases in lambs and their normal development are available within the first milk/colostrum produced by the dams after birth. Therefore, taking sufficient colostrum is very important and its inadequacy may be determined by various blood parameters such as IgG measured within 24 hours after the birth. Particularly, the diseases developing at the neonatal period are directly associated with insufficient intake of substances in colostrum. However, it is clear that this effect fades at the post-neonatal period and factors such as the physical and environmental conditions of the plant, vaccination becomes more effective [16]. Therefore, the disease classification shall be performed on neonatal lambs. In order to perform these analyses in an accurate way, features not associated with the disease status or those with direct relation were removed from the dataset. Eventually, 347 samples, 14 features and 1 class label were used in the study. The information and abbreviations of these features are given in Table 1.

Table 1. Dataset features and abbreviations

Features	Abbreviation	Type
Immunoglobulin G	IgG	Numeric { 19 – 5302 }
Gamma Glutamil Trasnfereraz	GGT	Numeric { 38 – 7517 }
Laktoferrin	LT	Numeric { 354 – 2194 }
Total Protein	TP	Numeric { 21 – 117 }
Albumin	ALB	Numeric { 32 – 51 }
Birthweight	BW	Numeric { 2260 – 5900 }
Body weight 28 day after birth	WG28	Numeric { 4364-14016 }
Average Daily Weight Gain	MDG28	Numeric { 17 – 340 }
Healthy status of dams	AH	Nominal { healthy / ill }
Dam's age	AGE	Numeric { 1 - 6 }
Dam's parity	PARITY	Numeric { 1 - 5 }
Type of birth	TWIN	Nominal { twin / single}
Gender	GENDER	Nominal { male / female}
Farm	FARM	Nominal { farm1 / farm2}
Neonatal healthy status of lamb	NGH	Nominal { healthy / ill }

2.3. Principle of the Analysis

Figure 1 shows the general principles of the analysis. The study consists of two-stage. At the first stage; a complete dataset consisting of 259 samples was obtained by removing 88 samples containing missing values from an original dataset consisting of 347 samples. We have implemented a varying percentage of missing values (from 5% to 25% by 5%) generated under an MCAR [17] assumption on the obtained complete dataset. These simulated missing values were imputed using the 5 methods. After measuring 3 evaluation criteria (RMSE, SCE, execution time), they were used to evaluate the differences between original values and replaced ones, the influence of imputed values through the RMSE, SCE criteria, and execution times in seconds, respectively. Two imputation methods with the least successful statistics were removed. In the second stage; missing values within the original dataset were imputed through these more successful three imputation methods. Then, NB and DT algorithms used to diagnosis classification were performed. The classifier performances of accuracy, kappa, recall and precision were compared and the influence of imputation methods on classifiers was observed.

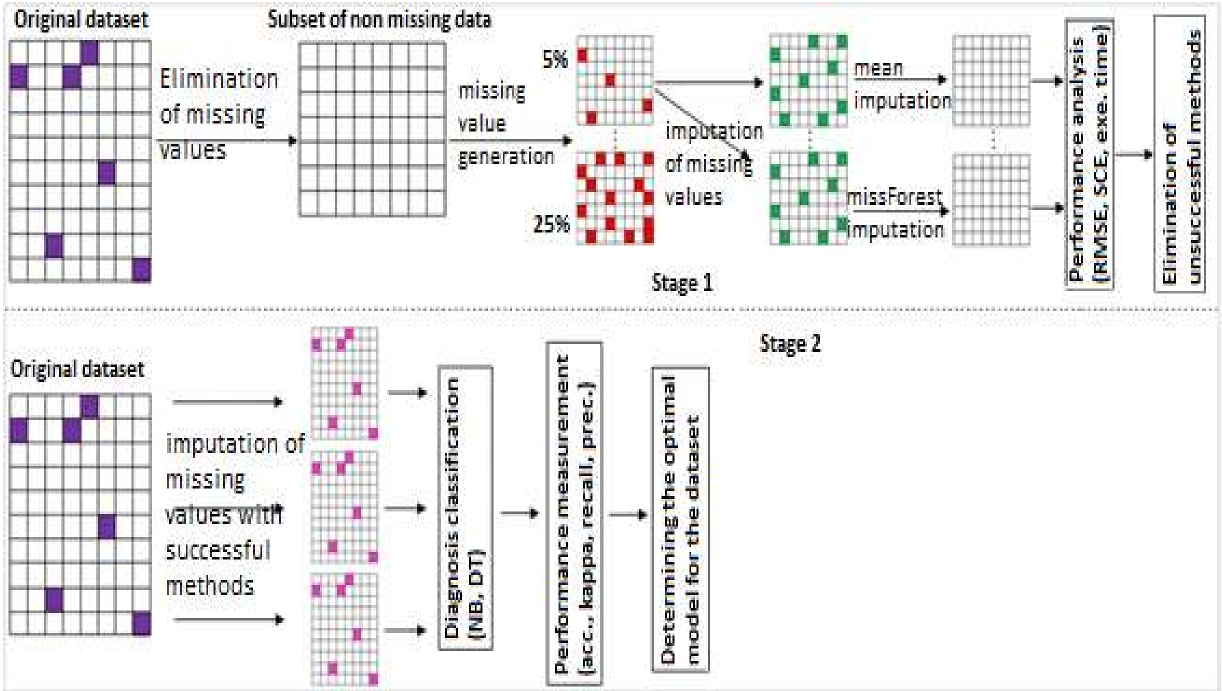

Figure 1. Principle of the method

2.4. Evaluation Metrics

At the first stage of this study, RMSE, SCE and execution time criteria were used, in order to compare the performances of imputation techniques in imputing the missing values within datasets which contain various percentages of missing values.

Root mean square error (RMSE): It measures the difference between the actual value and the estimated value. The smallest RMSE value is always desirable. Basically, the RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(t_{orig}^{m} - t_{reco}^{m}\right)^2} \quad (1)$$

Where t_{orig} and t_{reco} are the m^{th} vectors whose elements are the original values and the reconstructed values, respectively. M denotes the amount of missing value was used.

Supervised classification error (SCE) rate: After imputing the missing data through supervised classification, this criterion measures the difference between the current subgroups and those which were generated after missing data imputation and assesses if the discriminative or predictive capability is maintained. The approach used for supervised classification is NB algorithm. Classification error defined as:

$$SCE = \frac{\sum False\ positive \quad \sum False\ negative}{\sum Total\ population} \tag{2}$$

Execution time: This criterion indicates the time duration as missing values in datasets are completed by imputation methods. The difference in system time between the start and end of the method gives us this criterion. Particularly in big sized datasets (for example containing videos, images, etc.) the execution time is an important factor and completion in a short time is a desired aspect.

At the second stage of the study, diagnosis classification was performed using NB and DT algorithms, in order to observe the effectiveness of imputation methods and determine the optimal imputation method for the original dataset. Classification accuracy, kappa, recall and precision criteria were used in order to analyze the performances of classifiers. A confusion matrix [18] is used to indicate classification results together on a table where the above measurements may be calculated. A sample confusion matrix is shown in Table 2.

Table 2. A sample confusion matrix for two classes

		Actual	
		Positives (ill)	**Negatives (Healthy)**
Predicted	**Positives (ill)**	*TP (True Positive)*	*FP (False Positive)*
	Negatives (Healthy)	*FN (False Negative)*	*TN (True Negative)*

Classification accuracy: This is the simplest performance measure. It is the proportion of accurately classified samples obtained through any paired classification to the number of all samples. This study indicates how many ill and healthy neonatal lambs may be accurately estimated.

$$\text{Accuracy (Acc)} = \frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

Recall: Shows the percentage of actually ill lambs which may be estimated by the new method. Namely, this criterion gives the rate of accurate predictability for ill lambs.

$$\text{Recall} = \frac{\sum True\ positive}{\sum Condition\ positive} = \frac{TP}{TP+FN} \tag{4}$$

Precision: Proportion of real ill lambs that are positive (ill) according to the test result.

$$\text{Precision} = \frac{\sum True\ positive}{\sum Test\ outcome\ positive} = \frac{TP}{TP+FP} \tag{5}$$

Kappa statistic: It compares the measurement system with random estimation. High rate of agreement indicates the possibility of more accurate ratings. Poor rates of agreement denote that the ratings may be used in a limited way [19].

$$\text{Kappa} = \frac{(P_{observed} - P_{chance})}{(1 - P_{chanc})} \tag{6}$$

Where $P_{observed}$ is proportion of units classified in which the raters agreed and P_{chance} is proportion of units for which one would expect agreement by chance.

3. RESULTS

3.1. Analysis of Missing Values in Dataset

The dataset used in the study comprises 347 samples, 14 features and 1 label. Some features within the dataset contain missing values. Figure 2(A) shows the percentages of missing values in these features, and Figure 2(B) shows the total missing values of these features.

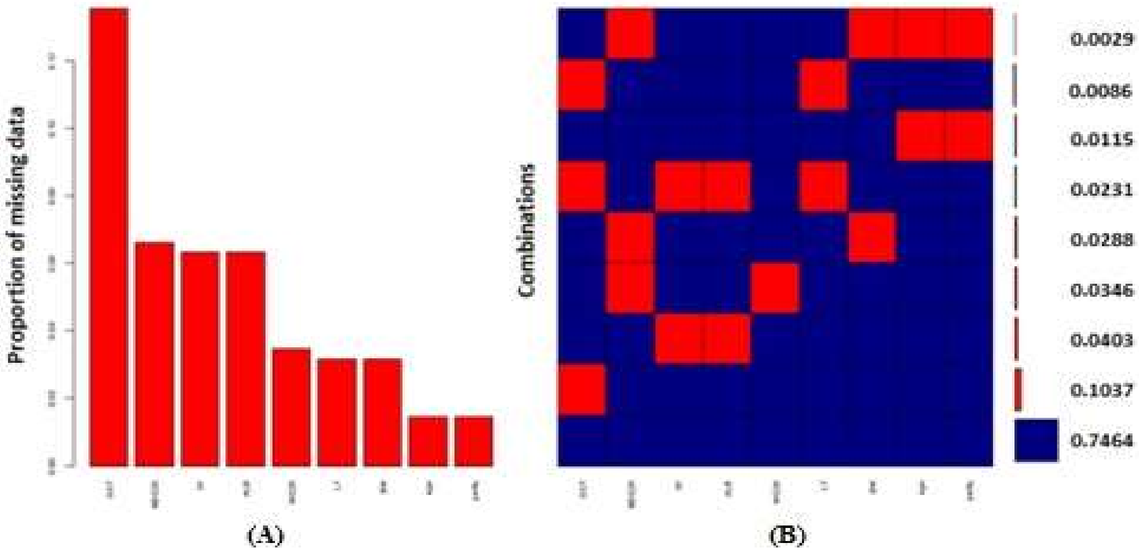

Figure 2. Histogram of missing values and missing rate [20]

When the missing value percentages of features within the dataset are examined (Figure 2(A)) the GGT looks as the one with most missing values, followed by the MDG28, TP, ALB, WG28, LT, BW, age and parity feature. It is obvious that the IgG, AH, twin, gender and farm features do not contain any missing value. When we examined features with missing values, it was clear that 4 out of 5 blood sample features (IgG, GGT, TP LT and ALB) contained missing values and GGT had more than twice missing values in comparison to other blood samples.

On the other hand, dividing each variable into certain ranges Figure 2(B) shows the percentage of missing values for all features in that range. This figure indicates that approx. 75% of samples in this dataset do not contain missing values, 10% miss only the GGT feature, 3% miss four blood samples (GGT, TP, LT, and ALB) together, 3% miss only the age feature. Eventually, missing percentage of the GGT is 2 to 10 times higher than that of other features.

The major part of missing features derives from the GGT. So, it is important to estimate this feature with minimal error, as otherwise it shall be a big source of problem for future analyses. Because IgG levels of <1000 mg/dL taken from ruminants such as lambs or calves with colostrum at the 24th hour after the birth is an important risk factor for the development of diseases in the neonatal period. This is defined as inadequate passive colostral immunity. Utilization of IgG level as a means of detecting passive colostral immunity is limited in farm examination programs, as its measurement is time-consuming and complicated, as it requires comprehensive laboratory conditions and advanced equipment, as it is laborious to send samples routinely to veterinary diagnostic laboratories, and because multiple sample analyses are exhaustive and non-economic. Therefore, detection of passive immunity through indirect test methods such as GGT enzyme activity and total protein (TP) level which are economic, fast and more practicable on field for individual flocks and using direct tests only as validation methods is considered as a better approach. Besides, GGT and TP have a big importance in terms of their adaptability on the field. For these concerns, the estimation of missing values for the GGT feature with a low error rate to determine the association between GGT enzyme activity and IgG levels with high performance is an important issue.

3.2. Evaluation of Imputation Techniques

After generation of missing value from the complete part of the original dataset in percentages of 5%, 10%, 15%, 20% and 25%, the missing values were imputed using mean, median, kNN, MICE and missForest imputation methods. RMSE was calculated for imputed datasets. The whole process was replicated 10 times and the results are summarized using the box-and-whiskers plots (median/IQR/min-max, including outliers) in Figure 3.

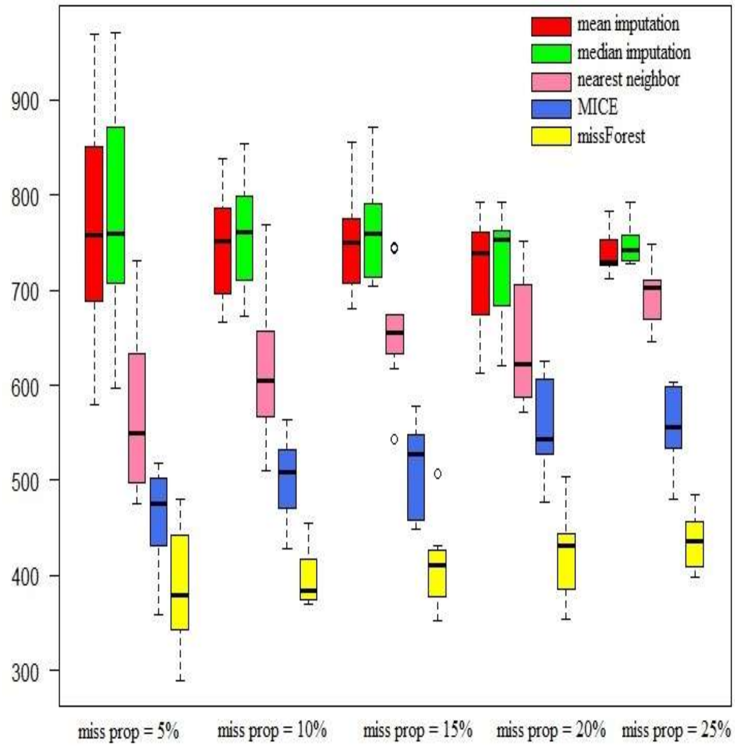

Figure 3. RMSE of imputation methods

RMSE value expresses the error level, so, lower levels indicate the success of the method. Figure 3 shows that the RMSE value of imputed datasets using mean and median method were higher than imputed datasets using kNN, MICE and missForest methods.

According to average RMSE results the most successful method has been missForest (avg. RMSE = 412), followed by respectively MICE (avg. RMSE = 519), kNN (avg. RMSE = 640), mean (avg. RMSE = 747) and median (avg. RMSE = 757) methods. Moreover, the average RMSE value of the missForest method is approx. half of the same for mean, median and kNN methods, so we may infer that it has imputed missing data with approx. 50% lower error rate in comparison to those three methods.

As the missing rate in datasets increase, it is obvious that the average RMSE values of the methods rise near to linear. Therefore, we may conclude those mean and median methods are the most unsuccessful methods in imputing missing values in our dataset for both their high averaged RMSE values and their high deviation from the average.

After imputing the missing values in datasets by 5 different methods, diagnosis classification was performed using NB method. The classification errors and standard deviation values are given in Table 3.

Table 3. Classification error of the five imputation methods.

Method	5%	10%	15%	20%	25%	Average
mean	0.150±0.030	0.113±0.036	0.146±0.046	0.152±0.040	0.146±0.034	0.141±0.037
median	0.141±0.032	0.140±0.033	0.157±0.029	0.133±0.028	0.150±0.043	0.144±0.033
kNN	0.141±0.031	0.154±0.028	0.132±0.031	0.149±0.015	0.125±0.040	0.140±0.029
MICE	0.119±0.024	0.151±0.038	0.155±0.025	0.139±0.031	0.151±0.041	0.143±0.032
missForest	0.137±0.035	0.122±0.029	0.135±0.033	0.139±0.034	0.104±0.024	**0.127±0.031**

While the classification error was 0.121 when the dataset without missing data was classified by the NB classification method, average classification errors of datasets imputed with mean, median, kNN, MICE and missForest imputation methods were 0.141 ± 0.037, 0.144 ± 0.033, 0.140 ± 0.029, 0.143 ± 0.032, and 0.127 ± 0.031 respectively. We have obtained the result that classification errors of imputed datasets using different imputation methods and classification error of the real dataset were quite close. Nevertheless, it is obvious that datasets imputed by the median, MICE and mean imputation methods were those which were classified with the highest error rate and which were the farthest to the real dataset. The dataset imputed by the missForest imputation method was concluded to be the most successful method with the lowest classification error.

When classification errors of imputation methods on average was considered the most successful method has been missForest (error = 0.127 ± 0.031), followed by kNN (error = 0.140 ± 0.029), MICE (error = 0.143 ± 0.032), mean (error = 0.141 ± 0.037) and median (error = 0.144 ± 0.033) methods.

The execution time of imputation method is also a significant factor in completing the missing values within big datasets. While estimation of nearest possible values to real ones by the method, it is also desirable to complete these operations in a short time. Therefore, execution times of methods were calculated and the results were given in Figure 4.

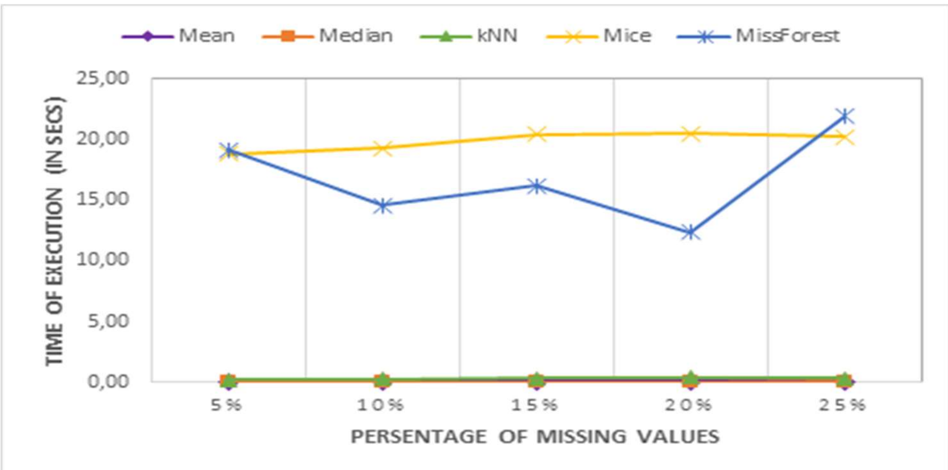

Figure 4. Execution time (secs) of imputation methods

When datasets containing 5-25% missing values were imputed through mean, median, kNN, MICE and missForest imputation methods, the average execution times of these methods were 0.05, 0.03, 0.3, 19.8, and 16.8 seconds respectively. The operation time is quite short for the mean and median methods, because it only places the feature's mean/median value. However, the error rates for datasets imputed with these methods were high (Figure 3 and Table 3).

Because a random forest is created at estimation of missing data through missForest method, and because the MICE method uses regression to estimate missing values, their execution times were longer than other methods. While execution time is important in big datasets, we have focused on success by ignoring execution time, because the dataset used in this study was not an extremely big one and the accurate estimation of the diagnosis was more important.

3.3. Effect in the Classifier Accuracy

Considering the performance of imputation methods and the criteria RMSE, SCE and execution time in imputing datasets, the missing values in an original dataset consisting of 347 samples were imputed through MICE, kNN and missForest imputation methods. Missing values in the original dataset were imputed by the above three methods and the optimal imputation method was determined for the dataset by examining the effect on their classification performances. Neonatal lambs were classified according to diagnosis using NB and DT methods which are frequently used within the literature. While classifying, 70% of the dataset was allocated to trainset, and 30% to test set purposes. During training at model creation, 10-fold cross-validation was performed and the obtained model was tested through a 30% test set. The whole process was replicated 10 times and the accuracy, kappa, recall, precision results are summarized using the box-and-whiskers plots (median/IQR/min-max, including outliers) in Figure 5.

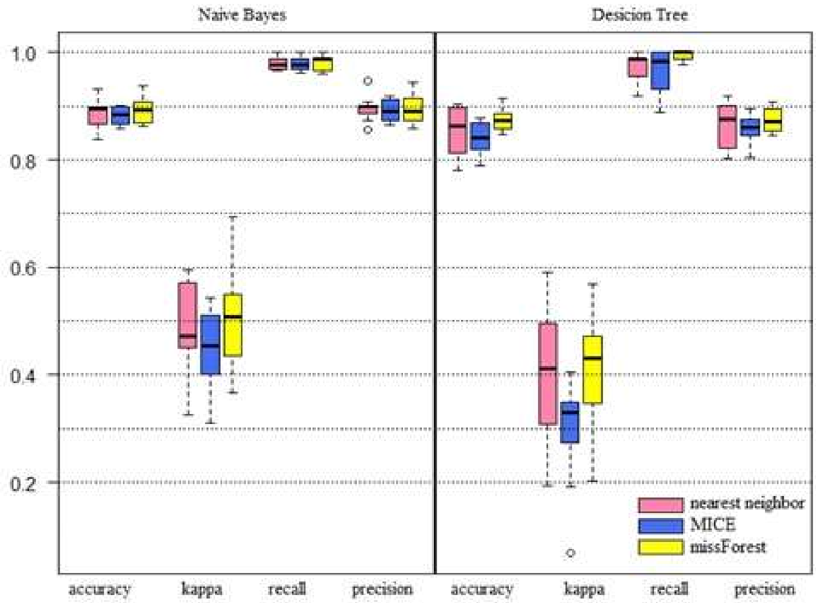

Figure 5. NB and DT classification algorithms results

As shown in Figure 5, the classification performances of datasets imputed by kNN, MICE and missForest methods are quite close to each other. The accuracy criterion indicates how accurately the number of healthy/ill neonatal lambs were estimated. According to this criterion, the classification performance of the dataset imputed by missForest was observed to be more successful than both NB and DT and other imputation methods. Even though the accuracy criterion is a straightforward and important criterion at classification, it should not be assessed alone. Because estimation of ill lambs is crucial in decision support systems along with accurate classification, in general. Therefore, when we examined recall criterion, it was clear that ill lambs were identified with higher success for the dataset imputed by missForest method in both NB and DT methods. Kappa criterion indicates whether the classification has a chance of success. We can say, the higher this value, the less random was the classification. In our study, it is obvious that missForest method had the highest kappa value for both NB and DT classification methods. On the other way, the precision criterion indicates how good healthy lambs were identified. For both NB and DT methods, it is obvious that the dataset imputed by the kNN method had a higher precision value.

4. CONCLUSION

The used dataset [20] in this study involved 301 ewes and 347 lambs born on two Akkaraman crossbreed sheep farms located in Kars, Turkey. As about approximately 87% of the sheep population in Turkey consist of the fat-tailed breeds mainly Akkaraman, also in veterinary field data mining application is very few and the data mining applications in this area are increasing rapidly. So this dataset has big importance both veterinary and computer science filed. The decrease in the number of ovine animals and livestock products impoverishes the people living in rural parts of our country. Raising the profitability or efficiency is a precondition for the prevention of impoverishment in this sector. This necessitates raising the demand for sheep breeding. Raising the demand for sheep breeding necessitates the reduction of disease and mortality rates. Therefore, estimations and analyses about the diagnosis of the animal is an important issue.

As the missing value within the dataset deprives the opportunity for analyses, they should be imputed. Many methods are being used for the imputation of missing values. The generation of values which are close to real values by the available methods for imputation shall influence the success of analyses positively. So, we have performed an evaluation of five imputation methods for imputing missing values in veterinary data and implemented the classifier accuracy.

At the first stage of the study, missing values were created in datasets in percentages of missing values from 5 to 25% by 5%, after imputing those datasets by mean, median, kNN, MICE and missForest imputation methods, the missing data imputation performances were compared taking the RMSE, SCE and execution time criteria into consideration. According to average RMSE results, the most successful method has been missForest method. When classification errors of imputation methods on average were considered the most successful method has been missForest method. On the other hand, when execution times of imputation methods were examined, the median imputation method is faster than others. The execution time was ignored because our dataset was not too big and also because an accurate estimation of ill lambs was more important. Eventually, at the first stage of the study, mean and median imputation methods were found to be the worst methods to imputing datasets containing 5-25% missing data. Mean and median methods reduce the standard error which invalidates most hypothesis tests. Also, it introduces a wrong representation of the relationship of the variable with other variables in the dataset.

At the second stage, NB and DT algorithm used to diagnosis classification, in order to observe the effect of kNN, MICE and missForest imputation methods on the success of disease classification and to determine the optimal method for the original dataset. To evaluate the classification success the frequently used accuracy, kappa, recall and precision criteria in literature were used. Through NB classification the most successful imputation method was found to be the missForest method. Through the DT classification model the most successful imputation method was again found to be the missForest method. Thus, according to our findings, the missForest imputation method is optimal for our veterinary dataset.

In the veterinary field; collecting the animals for weighing, measuring and other operations or taking and analyzing blood samples is a laborious and costly process. So, generally, a limited number of animal data were used in conducted studies. This study has special importance because it contains multiple features and numerous samplings in comparison to the veterinary field. While researchers might impute missing data by these tried and true imputation methods, they might also try out different imputation methods that would be compatible with the dataset considering that the performance of methods might differ in different datasets.

Acknowledgments

The dataset used in this study were collected from TUBITAK (Project Code; TOVAG 108 O 847).

5. REFERENCES

[1] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques: Elsevier, 2011.

[2] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," Neural Computing and Applications, vol. 19, pp. 263-282, 2010.

[3] J. L. Schafer, Analysis of incomplete multivariate data: Chapman and Hall/CRC, 1997.

[4] I. R. Dohoo, C. R. Nielsen, and U. Emanuelson, "Multiple imputation in veterinary epidemiological studies: a case study and simulation," Preventive veterinary medicine, vol. 129, pp. 35-47, 2016.

[5] G. Ser and S. Keskin, "Examining of Multiple Imputation Method in Two Missing Observation Mechanisms," JAPS, Journal of Animal and Plant Sciences, vol. 26, pp. 594-598, 2016.

[6] K. Hron, M. Templ, and P. Filzmoser, "Imputation of missing values for compositional data using classical and robust methods," Computational Statistics & Data Analysis, vol. 54, pp. 3095-3107, 2010.

[7] S. G. Liao, Y. Lin, D. D. Kang, D. Chandra, J. Bon, N. Kaminski, et al., "Missing value imputation in high-dimensional phenomic data: imputable or not, and how?," BMC bioinformatics, vol. 15, p. 346, 2014.

[8] G. Tutz and S. Ramzan, "Improved methods for the imputation of missing data by nearest neighbor methods," Computational Statistics & Data Analysis, vol. 90, pp. 84-99, 2015.

[9] J. Xia, S. Zhang, G. Cai, L. Li, Q. Pan, J. Yan, et al., "Adjusted weight voting algorithm for random forests in handling missing values," Pattern Recognition, vol. 69, pp. 52-60, 2017.

[10] P. Schmitt, J. Mandel, and M. Guedj, "A comparison of six methods for missing data imputation," Journal of Biometrics & Biostatistics, vol. 6, p. 1, 2015.

[11] P. Cihan, E. Gökçe, and O. Kalıpsız, "A review of machine learning applications in veterinary field," Kafkas Univ Vet Fak Derg, vol. 23, pp. 673-680, 2017.

[12] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, et al., "Missing value estimation methods for DNA microarrays," Bioinformatics, vol. 17, pp. 520-525, 2001.

[13] S. Van Buuren, H. C. Boshuizen, and D. L. Knook, "Multiple imputation of missing blood pressure covariates in survival analysis," Statistics in medicine, vol. 18, pp. 681-694, 1999.

[14] D. J. Stekhoven and P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," Bioinformatics, vol. 28, pp. 112-118, 2011.

[15] E. HM, "An epidemiological study on neonatal lamb health," Kafkas Üniversitesi Veteriner Fakültesi Dergisi, vol. 15, 2009.

[16] K. AH and E. HM, "Risk Factors Associated with Passive Immunity, Health, Birth Weight And Growth

Performance in Lambs: III. The Relationship among Passive Immunity, Birth Weight Gender, Birth Type, Parity, Dam," Kafkas Üniversitesi Veteriner Fakültesi Dergisi, vol. 19, 2013.

[17] R. J. Little and D. B. Rubin, Statistical analysis with missing data vol. 333: John Wiley & Sons, 2014.

[18] E. Alpaydin, Introduction to machine learning: MIT press, 2009.

[19] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: the kappa statistic," Fam Med, vol. 37, pp. 360-363, 2005.

[20] P. Cihan, "Determination of diagnosis, prognosis and risk factors in animal diseases using by data mining methods," Ph.D. dissertation, Comp. Eng., Yildiz Technical Univ., Istanbul, 2018.