COMPARISON OF SIMPLE MISSING DATA IMPUTATION TECHNIQUES FOR NUMERICAL AND CATEGORICAL DATASETS

[1]Ramu Gautam and [2]Shahram Latifi

[1,2]Department of Electrical and Computer Engineering, University of Nevada Las Vegas, United States Email: { [1]ramu.gautam@unlv.edu, [2]shahram.latifi@unlv.edu}

Abstract

Almost every dataset has missing data. The common reasons are sensor error, equipment malfunction, human error, or translation loss. We study the efficacy of statistical (mean, median, mode) and machine learning based (k-nearest neighbors) imputation methods in accurately imputing missing data in numerical datasets with data missing not at random (MNAR) and data missing completely at random (MCAR) as well as categorical datasets. Imputed datasets are used to make prediction on the test set and Mean squared error (MSE) in prediction is used as the measure of performance of the imputation. Mean absolute difference between the original and imputed data is also observed. When the data is MCAR, kNN imputation results in lowest MSE for all datasets, making it the most accurate method. When less than 20% of data is missing, mean and median imputations are effective in regression problems. kNN imputation is better at 20% missingness and significantly better when 50% or more data is missing. For the kNN method, k = 5 gives better results than k=3 but k=10 gives similar results to k=5. For MNAR datasets, statistical methods result in similar or lower MSE compared to kNN imputation when less than 25% of instances have a missing feature. For higher missing levels, kNN imputation is superior. Given enough data points without missing features, deleting the instances with missing data may be a better choice at lower missingness levels. For categorical data imputation, kNN and Mode imputation are both effective.

Key Words - Statistical imputation techniques; k-nearest neighbor imputation; sensor data imputation; MCAR; MNAR

1. Introduction

One of the major issues in machine learning is the missing data in the datasets. The datasets may be missing data because of several reasons like equipment malfunction, sensor malfunction, refusal to respond to a question, human error, translation error, and so on. If a significant portion of data is missing from a dataset, it could massively affect the accuracy, precision, and repeatability of a machine learning project. Missing data not only affects machine learning applications, but also data mining applications, audits, accounting, etc. If the dataset has missing data in it, one option is to remove all data points with missing data. This approach is feasible if there is a lot of data points and a very small missing data. In other cases, missing data imputation is beneficial and at times even necessary.

Missing data imputation generally means replacing missing values with a plausible value [1]. Statistical imputation methods have been used to effectively impute missing data [2]. Mean, Median and Most Frequent (Mode) are the three most common statistical methods used for missing data imputation. In mean imputation, the missing feature of a data point is replaced by the mean of that feature of all data points. While some studies show that mean imputation results in parameter estimates with high bias, other studies suggest that the limitation of mean imputation is not significant if the amount of missing data is less than 10% [3][4][5]. Median of the feature and most frequent value of the feature are used to replace missing data in median imputation and mode imputation, respectively. In this work, we compare the performance of k-nearest neighbor (kNN) as an imputation technique with the performance of the statistical methods discussed above. kNN is a machine learning approach, where the absent data is predicted from the k nearest neighbors of the dataset. Various machine learning methods have been studied for missing data imputation and have been shown to be excellent choices [2][6]. kNN is one such method which has been suggested as an effective data imputation technique~ [7][8]. Recently, data imputation research with kNN have focused mainly on classification problems [9][10]. In this research, we study missing data imputation in regression as well as classification problems.

We study two types of scenarios: Missing-Completely-at-Random (MCAR) and Missing-Not-at-Random (MNAR). Data is said to be missing completely at random if all the data points have equal probability of missing data. When the data are missing not at random, the missingness of the data is related to the unobserved data and the data are not equally probable to be missing. Cause and effect is often the basis for data to be MNAR. We study two regression datasets with numerical features and one classification dataset with categorical features. The first numerical dataset has no missing data and the second numerical dataset has <1% of all data missing. 1%, 2%, 5%, 10%, 20%, 50%, 67% and 75% of all features in the datasets are deleted at random to obtain new MAR sets. Since these deletions are random, multiple features of an instance can be deleted. This results in various levels of missingness for each dataset.

Data was deleted conditionally (e.g., SO2 readings are deleted if greater than 0.95) to obtain MNAR datasets. The range to satisfy the conditions were varied to obtain

different levels of missingness. In addition to that, 1% of all features in MNAR datasets are randomly deleted to mimic real-world datasets. The range of data deleted is varied to obtain various levels of missingness. For MNAR, the missingness used is the percentage of instances with at least one missing feature. For both MCAR and MNAR, we study the accuracies of imputation techniques and the effect of those imputation in machine learning applications.

The categorical dataset has around 5% missing data. We compare the area under the ROC curve (AUC) for datasets obtained from Mode and kNN imputation with AUC for the datasets resulting from deleting the instances with missing data.

2. Materials and Methods

Gas Turbine CO and NOx Emission Data Set from UCI Machine Learning Repository contains 36733 data points, each data point containing 11 sensor measures [11]. This dataset does not have any missing data points. The Gas Turbine CO and NOx Emission Data Set was collected from a gas turbine over a period of one hour in Turkey with the intention of studying CO and NOx emission [11][12]. The attributes are ambient temperature, ambient pressure, ambient humidity, air filter difference pressure, gas turbine exhaust pressure, turbine inlet temperature, turbine after temperature, compressor discharge pressure and turbine energy yield. The target variables are carbon monoxide concentration and nitrogen oxides concentration. Every attribute in this dataset is numerical and this dataset does not contain any missing data. The Beijing Multi-Site Air-Quality Data dataset contains hourly air pollutant concentration over four years [13]. It has data from 12 air quality monitoring site, and each monitoring site's data is self-sufficient to be treated as a separate dataset. Fourteen variables from the datasets are considered: year, month, day, hour, station, temperature, pressure, dew point, rain, wind speed, PM2.5 level, PM10 level, NO2 concentration and SO2 concentration. SO2 and NO2 concentration are selected separately as target variables. This dataset has less than 10% data points with a missing feature and less than 1% of all data is missing. The Autism Screening Adult Data Set is a classification dataset from UCI Machine Learning Repository with 21 attributes and 704 instances which is considered in this study~\cite{thabtah2017autism}. There are ten behavioral features and ten individual characteristics that have proved to be effective in detecting Autistic Spectrum Disorder [14].

The prediction accuracy of the imputation methods is first expressed in terms of mean absolute error in prediction of missing values. The imputed values and their original values before deletion are compared and the absolute difference between each imputed and original value is averaged over all imputation to get the mean absolute error. The data is min-max normalized before imputation, so the mean absolute error is also obtained in the normalized range. The fractions of missing data that were predicted with less than 0.1 and 0.25 errors compared to the original data are also obtained. We introduce the terms L10 and L25 to indicate the percentage of imputed data with Absolute Normalized Error (ANE) less than 0.1 and 0.25 respectively. i.e.,

$$\text{ANE} = |\text{ original value (normalized)} - \text{predicted value}| \tag{1}$$

$$\text{L10} = \frac{\#\text{ of imputed data with ANE} < 0.1}{\#\text{ of total imputed data}} * 100\% \tag{2}$$

$$\text{L25} = \frac{\#\text{ of imputed data with ANE} < 0.25}{\#\text{ of total imputed data}} * 100\% \tag{3}$$

The effect of missing data imputation techniques in machine learning applications is the focus of this research. The MLPRegressor and SVR from scikit-learn were selected as the machine learning models of choice for regression problems. SVR with radial basis function kernel and linear kernel were selected. MLPRegressor is a multilayer perceptron regressor and can be customized to suit what is needed. Gas Turbine CO and NOx Emission Data Set was used to select the best model. The first three years' data (2011-2013) was used as the ground truth, 2014 data as validation set and 2015 data as test set. We use the ground truth and validation set to choose the best machine learning model for this research. The model with the lowest average MSE for NOx and CO prediction is selected. We tested two Support Vector Regression methods (SVRs) and a 3-layer neural network, a 5-layer neural network and 10-layer neural network. The SVRs have Radial Basis Function (RBF) and linear kernels, respectively and the neural networks have 8 nodes in each hidden layer. The results are shown in Fig. 1.

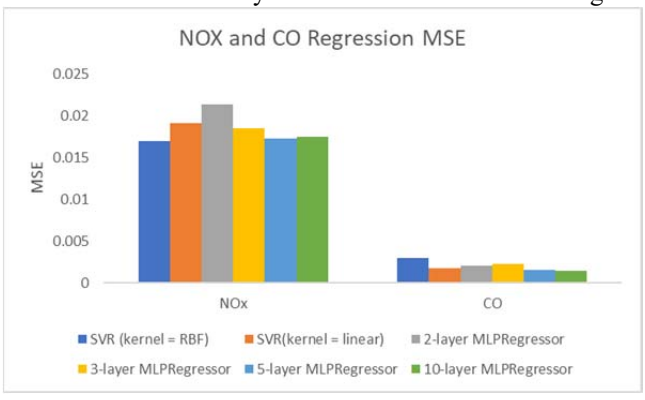


Fig 1. NOX and CO prediction MSE for various models

The SVR with RBF kernel resulted in highest mean-squared error (MSE) for CO but lowest MSE for NOX. 5-layer neural network resulted in the lowest mean-squared error (MSE) for CO and for NOX, it resulted in second lowest MSE with MSE very close to that of SVR. 5-layer MLPRegressor model will be used in this work.

For the classification problem, two models were considered: MLPClassifier from scikit-learn (a neural

network) and SVC from scikit-learn (a support vector machine). The neural netwok has five hidden layers. Constant learning rate with initial learning rate 0.0001 was selected. Adam was chosen for weight optimization. The area under the ROC curve (AUC) was taken as the parameter to select the better model between SVC and the neural network described above. The ROC curves and corresponding AUCs for are shown in Fig. 2. Fig. 2 shows that SVC results have higher AUC compared to the neural network. Thus, the SVC is selected as the model for this experiment.

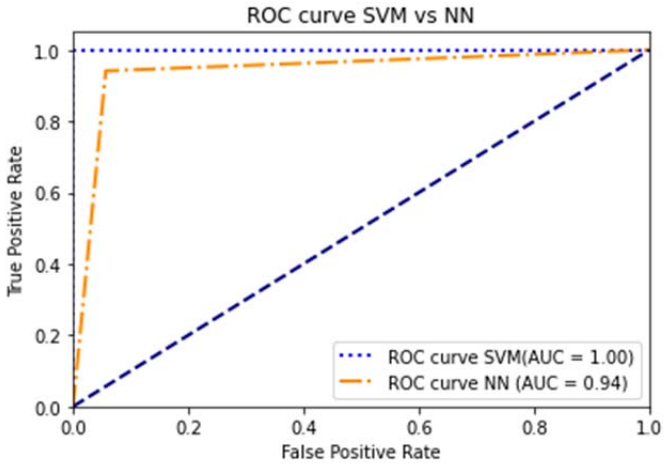


Fig. 2. ROC curve and AUC for ground truth and validation set.

3. Results and Discussion

3.1 Missing Completely at Random

3.1.1. Gas Turbine CO and NOx Emission Dataset

In this work, first we selected Gas Turbine CO and NOx Emission Data Sets. Nine attributes were then selected with CO and NO2 level as target variables. Min-Max scaler was used for feature scaling.

From the training set, 1%, 2%, 5%, 10%, 20%, 50%, 67% and 75% of each feature were deleted randomly to obtain three new datasets. This resulted in datasets with respectively 5%, 9%, 26%, 46%, 73%, 98%, 100%, and 100% of all instances with at least one missing feature. All these datasets were imputed using mean, median, mode, and KNN methods. Three values of k were selected for KNN (k=3,5, and 10). At the same time, all instances with missing data were removed from each dataset thus obtained to obtain 'missing-deleted' datasets at each level of missingness. Mean Absolute Error (MAE), and percentage of imputed data points with Absolute Normalized Error (ANE) less than 0.1 and 0.25 respectively (L10 and L25) for each imputation method shown in Fig. 3 at various missing levels.

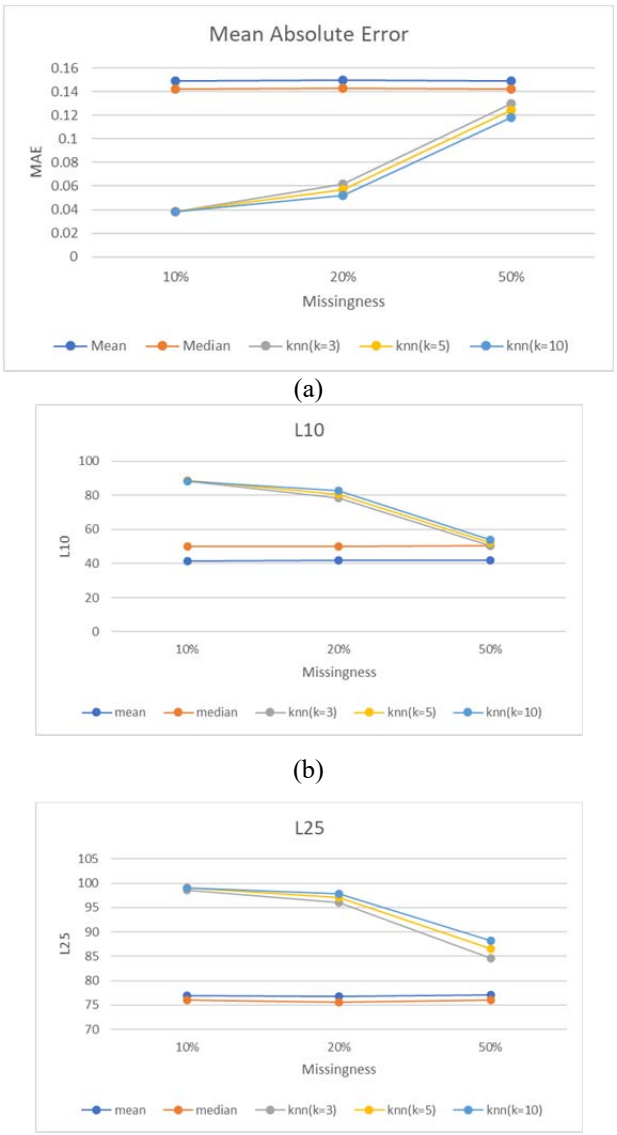


Fig. 3. (a) Mean absolute errors (b) L10 scores (c) L25 scores

Fig. 3 shows that kNN results in significantly lower Mean Absolute Error at all levels of missingness. Since the statistical methods are just replacing the missing values of a feature with the mean and median of the non-missing values of that feature, they will have almost a constant MAE. In case of kNN imputation, the number of neighbors does not cause a huge change in MAE, but the advantage of taking higher number neighbors in terms of lower MAE, L10 score and L25 score becomes more apparent with the increase in missingness. kNN imputation shows marginal improvement in prediction accuracy when the number of neighbors is increased from 3 to 10, so the higher number of neighbors were not tested. It is clear from the Fig. that kNN is good at imputation method, but this does not paint the whole picture. We want to study the benefits of using imputation methods, if any, and compare the performance of these imputation methods among themselves and against deleting

altogether the instances with missing data.

For this purpose, the datasets obtained from imputation were used to train the neural network discussed above and predictions were made on the test set with CO and NOx as target variables. Fig. 4 shows the resulting mean squared errors for NOx regression when the mean-imputed dataset, median-imputed dataset, KNN-imputed datasets with k = 3, 5, and 10, and missing-deleted dataset at various levels of missingness are used to train the model.

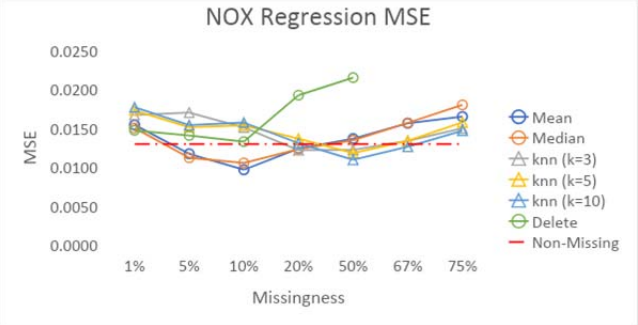


Fig. 4. NOX regression MSE for various imputation methods

The results show that for less than 20% missingness, the mean and median imputation show the best results compared to the other imputation techniques. This observation for mean imputation is consistent with literature [3], [4]. Deleting the data points with missing features resulted in lower MSE compared to kNN methods and similar MSE to using original dataset when the missingness is less than 10%. This can be attributed to the fact that the dataset has over 36000 instances. At 20% missingness, the MSE for statistical methods and kNN methods converges while deleting the missing instances results in significantly higher MSE. The advantage of kNN imputation methods becomes apparent at 50% missingness and the trend continues for 67% and 75% missingness. At those levels of missingness, 100% of the datapoints have at least one missing feature, so imputation is necessary. The difference in MSE for k=3,5 and 10 is not significant, but k=10 results in consistently lower MSE at higher missingness.

Same process was repeated for CO emission regression. The results are shown in Fig. 5.

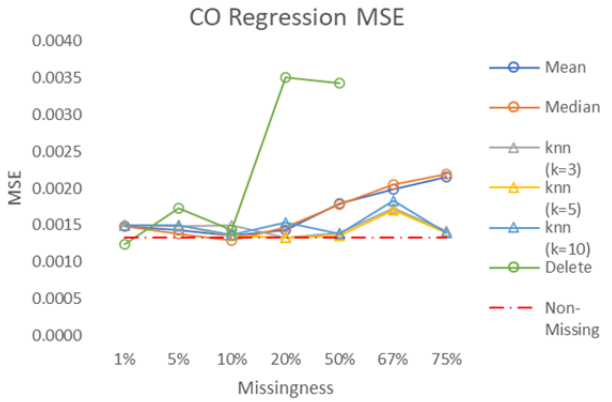


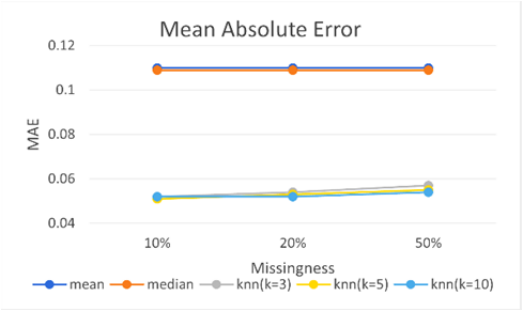
Fig. 5. CO regression MSE for various imputation methods

Similar to the NOX regression results, mean and median imputation result in lowest MSE at 10% and less missingness. KNN imputation is significantly more effective at higher levels of missingness, especially when 50% or more of the data is missing. KNN imputations show similar MSE over three levels of missingness. Except for 1% missingness, deleting the instances with missing data resulted in higher MSE compared imputing with any imputation method. The difference is significant at 20% missingness and beyond.

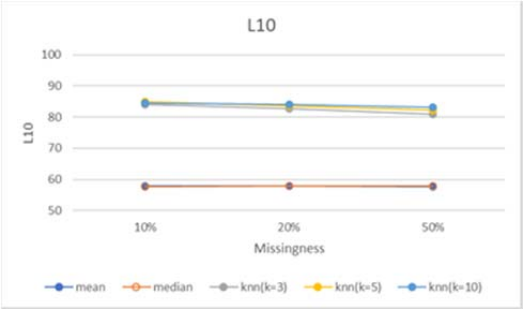
3.1.2 Gas Turbine CO and NOx Emission Dataset

Out of several sites, the Aotizhongxin and Changpin site data from Beijing Multi-Site Air Quality Data dataset available at UCI Machine Learning Repository were selected. SO2 and NO2 levels were selected as target variables. In an attempt to compare the efficacy of imputation with the dataset with no imputation, instances with missing data were removed. The dataset obtained as such was split into training set and test set at 80/20 distribution.

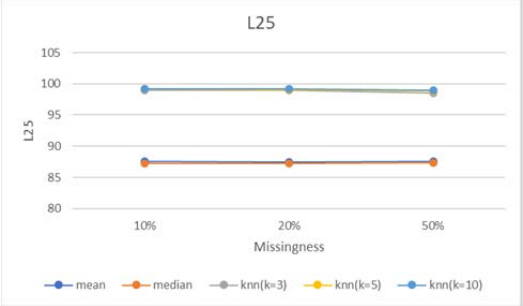
The training set was subjected to data deletion. From the training set, 1%, 2%, 5%, 10%, 20%, 50%, 67% and 75% of each feature were deleted randomly to obtain three new datasets. This resulted in datasets with respectively 5%, 9%, 26%, 46%, 73%, 98%, 100%, and 100% of all instances with at least one missing feature. From each of those datasets, the instances with missing features were deleted to obtain the missing-deleted datasets. The datasets with various levels for missingness were then imputed using mean, median and KNN (k=3, 5, and 10) imputation techniques. Mean absolute error between imputed and original values and L10 and L25 scores are obtained for 10%, 20% and 50% missingness. The results are shown in Fig. 6.



(a)



(b)



(c)

Fig. 6. (a) Mean Absolute Errors (b) L10 Scores (c) L25 Scores

The results are similar to the previous dataset, with kNN imputation resulting in lowest mean absolute error in missing data prediction.

Imputations were made for the datasets with various levels of missingness discussed above. The instances with missing values were removed to obtain missing-deleted sets. These datasets and the non-missing dataset before deletion were used to train a machine learning model, and predictions were made on the test set. As discussed in section 2, the 5-layer neural network was trained using the datasets obtained from imputation or deletion of incomplete instances and prediction was made on the test set. Fig. 7 shows the NO2 regression mean squared error for all cases at the specified levels of missingness.

Fig. 7. NO2 regression MSE for various imputation methods

Fig. 8 shows SO2 regression mean squared error for all cases at the specified levels of missingness.

Fig. 8. SO2 regression MSE for various imputation methods

For both NO2 and SO2 regression, statistical imputation methods result in slightly lower MSE, but the difference is not significant. Deleting the instances with missing features results in similar performance in terms of resulting MSE in prediction compared to imputing the missing data. Imputing the missing data lowers the MSE by a significant amount compared to delete approach for all imputation techniques. Beyond 20% missingness, all three kNN imputations result in significantly lower MSE compared to mean and median imputation. Beyond 50% missingness, all instances of data have at least one missing feature, so deleting is not an option. The plots show that imputation must be done when the missingness is 20% or higher.

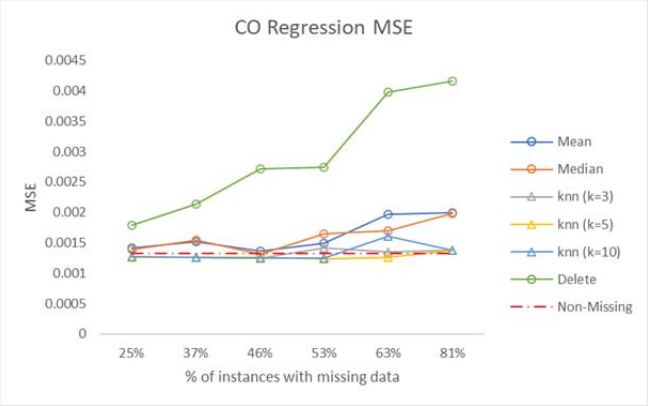
3.2 Missing Not at Random

Gas Turbine CO and NOx Emission Dataset and Beijing Multi-Site Air Quality Data Dataset were used to study the imputation of MNAR data. Each sensor reading was subjected to deletion when the mean-max normalized value reached a certain range. This range was varied to obtain various levels of missingness. In this section, missingness

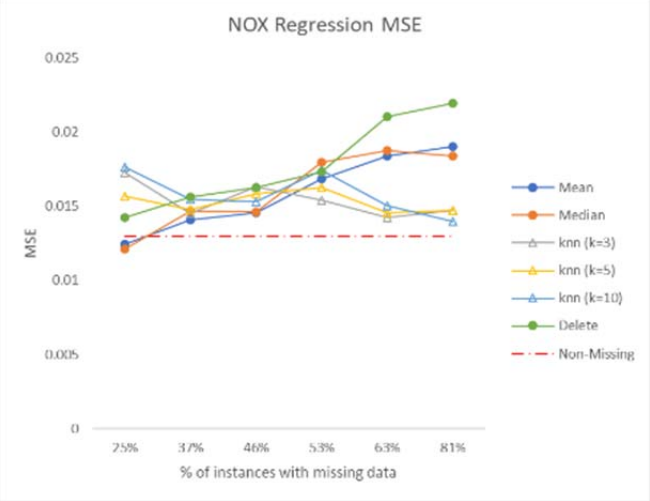
refers to the percentage of all datapoints with at least one missing feature. The datasets thus obtained was imputed using mean, median, and kNN imputation (k=3, 5, and 10).

3.2.1 Gas Turbine CO and NOx Emission Dataset

After imputation, resulting datasets were used to train the neural network and make prediction on the test set as discussed above. The resulting MSE in NOX and CO regression are shown in Fig. 9.



(a)



(b)

Fig. 9. CO and NOX regression MSE for various imputation methods

For both CO and NOx regression, kNN imputation methods result in the lowest MSE when at least 50% of the instances have missing data. At lower missingness however the results are different for CO and NOx regression. This can be attributed to the fact that when the data is missing not at random, regression of one variable may suffer more than regression of another variable based on which feature is missing. KNN imputation with k=5 shows consistently low MSE for both variables compared to other imputation methods, and more importantly when compared to deleting the instances with missing data.

3.2.2 Beijing Multi-Site Air Quality Data Dataset

After not-randomly deleting the data at various amount, the resulting datasets were imputed using mean, median and kNN (k=3,5,10) imputation methods. The datapoints with missing data were removed to obtain missing-deleted sets at each missingness level. All these datasets were used to train the neural network and predictions were made on the test set. The mean square errors in SO2 and NO2 regression are shown in Fig. 10.

(a)

(b)

Fig 10. SO2 and NO2 regression MSE for various imputation methods

3.3 Categorical Dataset

3.3.1 Autism Screening Adult Dataset

Autism screening Adult Data Set from UCI Machine Learning Repository is a binary classification dataset. The data set contains categorical as well as numerical attributes and there are 704 instances and 21 attributes in this dataset. In this section, we are going to explore the efficacy of data imputation techniques in categorical data, namely mode-imputation and KNN imputation.

The dataset was split into instances missing at least one feature and instances not missing any feature. 20% of instances with no missing points were taken as the test set and 20% of remaining non-missing instances were taken as the validation set. The selection was completely random.

The remaining non-missing instances are taken as ground truth in this section. The combination of this set and the instances with missing values was the training set. As discussed in section 2, SVM will be used as the machine learning model in this experiment.

Since the training set already consists missing data, one additional dataset with 10% missing data was obtained. All the categorical features were deleted at random, and imputation was performed using Mode and KNN imputation methods. K=5 was taken. Fig. 11 shows the ROC curve and AUC for mode imputed data at original missingness (<5% of total instances, < 1% of total data) and Fig. 12 shows the ROC curve and AUC for 10% missingness.

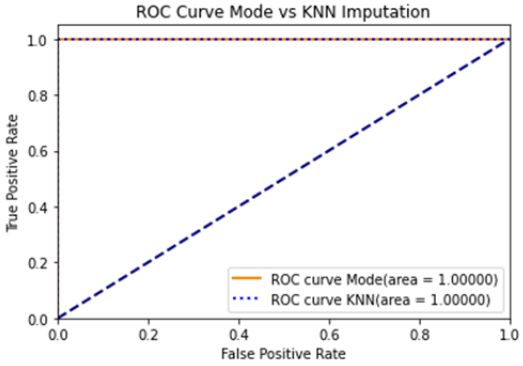


Fig. 11. ROC curves at <5% missingness.

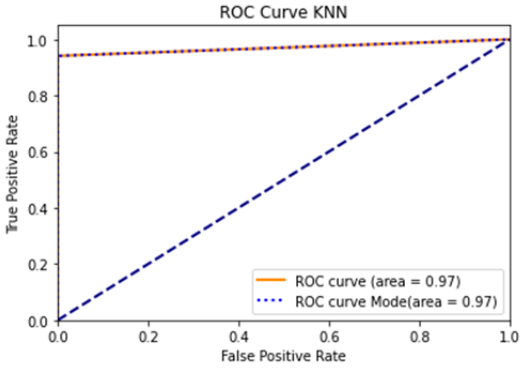


Fig. 12. ROC curves at 10% missingness.

For this dataset, both KNN imputation and mode imputation result in similar AUC for SVM. AUC drops as the number of missing features increases, but the ratio is consistent for both methods.

In all our experiments, KNN has consistently performed highly very effectively as an imputation tool, especially on datasets in which a large fraction of data is missing. This is, however, not the complete story, as we need to address the "KNN and the curse of dimensionality." When the number of features increase, the vector space for the data points increases exponentially. As the number of features approach higher numbers (say 100-1000), the data instances are very sparsely scattered in the vector space. For N samples with d features, the edge length of the smallest hypercube that encloses k nearest neighbors of a point is given by:

$$l = \left(\frac{k}{N}\right)^{1/4} \tag{4}$$

As the dimension 'd' increases, the pairwise distances between the test point and all the other data points approach a common value. This equation tells us that to counter this curse, N must increase exponentially with a linear increase in d, meaning the number of samples has to increase significantly. Therefore, KNN is not a good machine learning model for high-dimensional datasets.

4. Results and Discussion

For numerical datasets with data missing completely at random, mean imputation can be effective at lower level of missingness but relatively not suitable if the fraction of instances with missing features is high. At less than 10% missingness, removing the instances with missing features can be an option if the total number of instances is high. As the number of instances with missing data increases, data imputation is necessary for accurate machine learning application. KNN is extremely effective as a data imputation tool, especially when the level of missingness in a dataset is high. For numerical datasets with data missing not at random (MNAR), simple imputation methods can be effective for some features and not-so-effective for others. KNN imputation with k=5 results in consistently low MSE for all missing levels and different variables and can be a good imputation technique, but further research needs to be done for effective imputation of MNAR data. Future work will focus on introducing an imputation method for MNAR datasets based on statistical and machine-learning-based simple imputation techniques. KNN is also highly effective as a data imputation tool for categorical data, as long as there are not too many features and there are enough data points.

References

1. D. B. Rubin, "Inference and missing data," Biometrika, vol. 63, no. 3, pp. 581–592, 1976.
2. J. M. Jerez et al., "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," Artificial Intelligence in Medicine, vol. 50, no. 2, pp. 105–115, 2010
3. J. W. Graham, S. M. Hofer, S. I. Donaldson, D. P. MacKinnon, and J. L. Schafer, "Analysis with missing data in prevention research.," 1997.
4. N. Tsikriktsis, "A review of techniques for treating missing data in OM survey research," Journal of operations management, vol. 24, no. 1, pp. 53–62, 2005.
5. M. R. Raymond, "Missing data in evaluation research," Evaluation & the health professions, vol. 9, no. 4, pp. 395–420, 1986.
6. A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of Performance of Data Imputation Methods for Numeric Dataset," Applied Artificial Intelligence, vol. 33, no. 10, pp. 913–933, Aug. 2019

7. G. E. Batista and M. C. Monard, “A study of K-nearest neighbour as an imputation method.,” His, vol. 87, no. 251–260, p. 48, 2002.
8. G. E. Batista and M. C. Monard, “An analysis of four missing data treatment methods for supervised learning,” Applied Artificial Intelligence, vol. 17, no. 5–6, pp. 519–533, May 2003
9. A. Choudhury and M. R. Kosorok, “Missing Data Imputation for Classification Problems,” Feb. 2020.
10. A. Ngueilbaye, H. Wang, D. A. Mahamat, and S. B. Junaidu, “Modulo 9 model-based learning for missing data imputation,” Applied Soft Computing, vol. 103, p. 107167, May 2021
11. D. Dua and C. Graff, “UCI Machine Learning Repository.” 2017. [Online]. Available: http://archive.ics.uci.edu/ml
12. H. Kaya, P. Tüfekci, and E. Uzun, “Predicting co and no x emissions from gas turbines: novel data and a benchmark pems,” Turkish Journal of Electrical Engineering & Computer Sciences, vol. 27, no. 6, pp. 4783–4796, 2019.
13. S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, and S. X. Chen, “Cautionary tales on air-quality improvement in Beijing,” Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 473, no. 2205, p. 20170457, 2017.
14. F. A. Thabtah, “Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment,” Proceedings of the 1st International Conference on Medical and Health Informatics 2017, 2017.

Ramu Gautam is a PhD student in Electrical and Computer Engineering Department at University of Nevada Las Vegas. Currently his research is in computer vision, focusing on 3D and 4D biological images. He has a master’s degree in Nanotechnology and a bachelor's degree in Electronics and Communication Engineering. He likes to go hiking and play table tennis in his free time.

Shahram Latifi is a Professor of Electrical Engineering at the University of Nevada, Las Vegas. Dr. Latifi is the co-director of the Center for Information Technology and Algorithms (CITA) at UNLV. He has designed and taught undergraduate and graduate courses in the broad spectrum of Computer Science and Engineering in the past four decades. He has given keynotes and seminars on machine learning/AI and IT-related topics all over the world. His research has been funded by NSF, NASA, DOE, DoD, Boeing, Lockheed, and Cray Inc. Dr. Latifi is the recipient of several research awards, the most recent being the Barrick Distinguished Research Award (2021). Dr. Latifi was recognized to be among the top 2% researchers around the world in December 2020, according to Stanford top 2% list (publication data in Scopus, Mendeley). He is an IEEE Fellow and a Registered Professional Engineer in the State of Nevada.