*Research Journal of Recent Sciences* ISSN 2277-2502
Vol. **15(3),** 68-71, July **(2026)** *Res. J. Recent Sci.*

# A Quantitative analysis of Machine Learning approaches to missing value imputation

**Priya S.**
Department of Computer Science, Government First Grade College, Domlur Shanthi Nagar, Bengaluru, India
priya5mithul@gmail.com

**Available online at: www.isca.in, www.isca.me**
Received 14th April 2026, revised 18th May 2026, accepted 5th June 2026

## Abstract

*Missing data is a persistent challenge in real-world datasets and can significantly reduce the reliability and predictive performance of machine learning models. Accurate imputation strategies are therefore essential for effective data pre processing. This study presents a quantitative evaluation of machine learning–based imputation techniques, with a focused comparative analysis of Naïve Bayes and K-Nearest Neighbor (KNN) methods. The algorithms are assessed using accuracy, root mean square error (RMSE), and computational efficiency across datasets with varying proportions of missing values. Experimental observations indicate that KNN achieves superior estimation accuracy, while Naïve Bayes demonstrates faster execution and better scalability in high-dimensional environments*.

**Keywords:** Missing data, imputation, Naïve Bayes, K-Nearest Neighbor, machine learning, data preprocessing.

## Introduction

Data Mining (DM) is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouse, or other information repositories[1]. There are a lot of functions of DM, such as description, association analysis, classification and prediction, clustering analysis etc. Among all, classification and prediction are widely used in many fields. However, in real-world datasets, there are many problems in data quality such as incompleteness, redundancy, inconsistency, noise data etc. All these serious data quality problems affect the performance of DM algorithms[2].

Incomplete data commonly arises from sensor malfunction, human error, non-response in surveys, or transmission failures. The presence of missing values can bias statistical analysis, degrade model accuracy, and reduce generalization capability. Traditional statistical imputation techniques—such as mean, median, and regression substitution—are computationally simple but often fail to capture complex relationships among variables.

Machine learning–based imputation methods provide a more adaptive solution by leveraging patterns within the data. Among these, Naïve Bayes and K-Nearest Neighbor remain widely used due to their simplicity, interpretability, and effectiveness. This research aims to quantitatively compare these two approaches and identify the contexts in which each method performs best.

Missing data is a quite common problem of data quality in real-life datasets. Then, what is the reaction of classifier to missing data? For the large amount of missing data in real-world datasets, several methods have been proposed. For example, case deletion, mean imputation and model prediction. The basic idea of model prediction is using prediction model, built on known data, to predict and fill in the missing data. The performance of the missing treatment relies on the prediction model. Then, which classifier will be suitable for handling missing data? So we have chosen two predictive classifiers and compared to find which classifier is least sensitive and most sensitive to missing data.

**Classification and prediction: Classifier:** Classification means constructing a classifying function or model from the known data. Such function or model can also be called 'classifier', which can classify the records in database into given classes, thus can predict the unknown variables under some given conditions[3]. Classifiers differ greatly in prediction accuracy, training time and number of leaves (Decision Trees). There is not a classifier which performs best in all aspects[4]. Prediction accuracy of classifiers can be affected by the factors as follows[3].

**Number of records in training subset:** Classifier needs to learn from training set. Therefore, larger training set makes the classifier more reliable. But the training time is also become longer.

**Data quality:** Problems such as noise data, missing data, data inconsistency etc. bring a lot of wrong information which will lead to wrong classify. It is impossible to build a convictive classifier with incompleteness or wrong data.

**Attribute quality:** Attributes provide information for classifying. The prediction accuracy can be improved by including more attributes. However, more attributes means calculating more attribute combinations and more training time.

It is essential to choose attributes which are valuable for classification.

**Characteristics of the records to be predicted:** If Characteristics of the records to be predicted are different from records in training set, it may lead to high incorrect rate.

**Missing treatment methods using classifier:** In the popular methods for missing data handling, using classifier to predict and fill in missing data is a set of fast developing methods. Among the so many classifiers, which one is suitable and how to use is still in the research. Methods have been used are introduced as follow:

**K-Nearest Neighbor Imputation (KNN):** This method uses k-nearest neighbor algorithms to estimate and replace missing data. The main advantages of this method are that:

It can estimate both qualitative attributes (the most frequent value among the k nearest neighbors) and quantitative attributes (the mean of the k nearest neighbors).

It is not necessary to build a predictive model for each attribute with missing data, even does not build visible models. Efficiency is the biggest trouble for this method. While the k-nearest neighbor algorithms look for the most similar instances, the whole dataset should be searched. However, the dataset is usually very huge for searching. On the other hand, how to select the value "k" and the measure of similar will impact the result greatly.

**Bayesian Iteration Imputation (BII):** Naive Bayesian Classifier is a popular classifier, not only for its good performance, but also for its simple form. It is not sensitive to missing data and the efficiency of calculation is very high. Bayesian Iteration Imputation uses Naive Bayesian Classifier to impute the missing data. It is consisted of two phases:

Decide the order of the attribute to be treated according to some measurements such as information gain, missing rate, weighted index, etc.;

Using the Naive Bayesian Classifier to estimate missing data. It is an iterative and repeating process. The algorithms replace missing data in the first attribute defined in phase one, and then turn to the next attribute on the base of those attributes which have be filled in. Generally, it is not necessary to replace all the missing data (usually 3~4 attributes) and the times for iterative can be reduced[7].

## Methodology

**Sensitivity Analysis:** Sensitivity analysis (SA) is to study the impacts of one or more input variables on the outputs of a model, that is, the sensitivity of the model to one parameter or a combination of parameters[8]. If a tiny change of an input leads to great changes of the output, the model is highly sensitive to that input. SA can help to identify the decisive input parameter of the model[8]. In our experiments, the proportion of missing data in the datasets is the parameter which affects the results of the classification models. The effect of missing data on the prediction accuracy will be investigated through the tiny changing of the missing rate.

**Design of Experiments:** This paper has selected 2 classifiers to study the influence of missing data to classifiers, that is, Naive Bayesian classifier (NB), K-Nearest Neighbor classifier (KNN) datasets were collected from UCI repository and, are used in the experiments, as shown in Table-1.

Three indexes are used to evaluate the missing influence on classifier: Prediction accuracy (Pa), Prediction profit (Pp)[10] and Prediction losing (Pl), which are defined as follows.

$$\mathrm{Pa} = \frac{\text{number of correctly predicted record}}{\text{number of all records}} \times 100\%$$

$$\mathrm{Pp} = \frac{\mathrm{Pa}-\mathrm{MD}}{\mathrm{MD}} \times 100\%$$

$$\mathrm{Pi} = \frac{\mathrm{AC} - \text{Pa under certain missing rate}}{\mathrm{AC}} \times 100\%$$

AC is the prediction accuracy without missing data. MD is the proportion of the class in the dataset which having the greatest number of records.

Without any prediction model, if all the records are classified into that class, MD will be the prediction accuracy of the dataset. Prediction profit is proposed by Peng Liu, Elia El-Darzi et al.[9] to evaluate the performance of the classifier.

Then, a given percentage, 10%, 20%, … , 80% of missing data is artificially inserted into the training subsets at completely random. Finally, the two classification algorithms mentioned above are applied into the training subset to build up classifiers, and, these classifiers are used to classify instances in testing subset to investigate the classification accuracy. The average prediction accuracy of the two classifiers under different missing percentages is displayed from Figure-1 . Limited by the space, only part of the results is showed in this paper.

## Results and Discussion

The figure interprets that with the increase in missing rate, the prediction accuracies of all the classifiers have an obvious trend of decrease. In general, when the proportion of missing data in the dataset is less than 10%, they have little adverse impact on the classifiers. If the missing rate is between 10% and 20%, the impact should not be neglected. The average Prediction losing rises to 4.63%. However, the adverse impact can be reduced significantly by some simple methods, such as replacing missing data by an approximation.

If the missing rate exceeds 20%, there is an obvious decrease in the prediction accuracy and the missing data should be handled with high cautiousness.

**Table-1:** Average prediction losing of classifier (%)[10].

| Missing data | NB | Knn | Average |
|---|---|---|---|
| 0 | 0 | 0 | 0.00 |
| 10% | 0.08 | 4.29 | 2.19 |
| 20% | 0.59 | 13.22 | 6.91 |
| 30% | 0.47 | 16.45 | 8.46 |
| 50% | 1.24 | 22.45 | 11.85 |
| 70% | 2.7 | 24.95 | 13.83 |
| 80% | 9.73 | 29.03 | 19.38 |

Appropriate methods should be chosen to eliminate the adverse impact of the missing data and optimize the performance of classifiers. In the real world, there are great quantities of missing data in databases and, usually, the proportion of missing data exceeds 20%. However, if the proportion of missing data exceeds 50%, the average prediction losing rises to more than 10%.

Obviously, the loss of prediction accuracy caused by missing data is quite huge. With the increase in the missing rate, the losing of prediction is rising with accelerated paces. That is to say, with the increase of quantity of the missing data, the little raising of the missing rate will result in a larger and larger decrease in the prediction accuracy. The impact of missing data depends on classifiers.

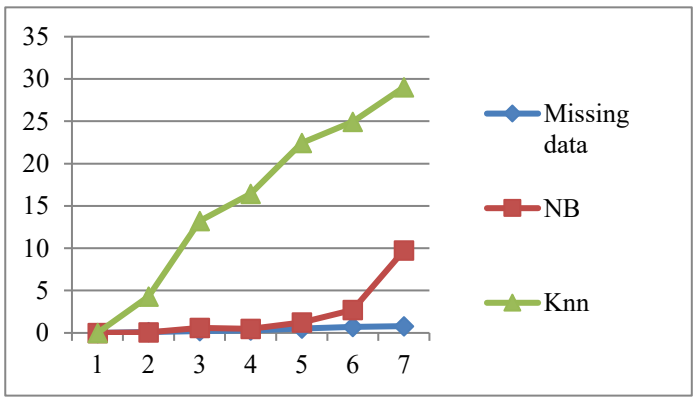


**Figure-1:** Average losing rate of classifiers.

Among the classifiers, Naive Bayesian classifier is the least sensitive to missing data, And K-Nearest Neighbor susceptible to missing data most, with the increasing in the missing rate, the prediction accuracy of Naive Bayesian classifier is almost the same as that with no missing data. Only when more than 70% data are missing, the prediction accuracy drops obviously. the prediction losing of Naive classifier is always the lowest and that of the KNN is always the highest.

Furthermore, the pace of KNN is also very fast. When missing rate exceeds 10%, there is an obvious and sharp increase in prediction losing. The prediction profit of Naive Bayesian classifier drops with the lowest and smoothest pace. And the K-Nearest Neighbor has the sharpest trend. In summary, Naive Bayesian classifier is the least sensitive to missing data. Even if training subset has a lot of missing data, Naive Bayesian classifier can still make full use of the existing data and operate effectively.

Among all the classifiers, though the classification accuracy (with no data missing) of Naive Bayesian classifier isn't the highest, it is the most adaptive to missing data. The Naive Bayesian classifier is recommended for the datasets with great quantities of missing data. While, selecting Naive Bayesian classifier to deal with missing data can also get a better solution.

## Conclusion

Missing data may reduce the accuracy of prediction models. This paper mainly studies the impact of missing data to classification algorithms. The sensitivity of classifiers to missing data is analyzed. The results showed that, with the increasing of the missing rate, the classification accuracies of all the classification algorithms have an obvious trend of decrease. If the proportion of missing data exceeds 20%, there is an obvious decrease in the accuracy of prediction. Methods for missing data treatment should be chosen cautiously to eliminate the negative impact on the classification accuracy and optimize the performance of classifiers. Among the two classifiers, the Naive Bayesian classifier is the least sensitive to missing data and K-Nearest Neighbor is the most sensitive. In conclusion, for datasets suffered with missing data, we prefer to use Naive Bayesian classifier to deal with missing.

## References

1. Lin, W. C., & Tsai, C. F. (2020). Missing value imputation: A review and analysis of the literature (2006–2017). *Artificial Intelligence Review,* 53(2), 1487–1509.
2. Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big data*, 8(1), 140.
3. Hasan, M. K., Alam, M. A., Roy, S., Dutta, A., Jawad, M. T., & Das, S. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked*, 27, 100799.
4. Liu, M., Li, S., Yuan, H., Ong, M. E. H., Ning, Y., Xie, F., Saffari, S. E., Shang, Y., Volovici, V., Chakraborty, B., &

Liu, N. (2023). Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artificial Intelligence in Medicine,* 142, 102587.

**5.** Rahman, M. G., Islam, M. Z., Islam, M. M., & Uddin, J. (2021). A systematic review of machine learning-based missing value imputation techniques. *Data Technologies and Applications, 55*(4), 558–585.

**6.** Twala, B. E., Cartwright, M., & Shepperd, M. (2011). Missing data imputation using the EM algorithm and machine learning techniques. *Applied Artificial Intelligence,* 25(5), 373–393.

**7.** Acuña, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. *Classification, Clustering, and Data Mining Applications*, 639–647.

**8.** García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. *Neural Computing and Applications,* 19(2), 263–282.

**9.** García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks,* 24(1), 121–129.

**10.** Liu, W., Luo, L., & Zhou, L. (2023). Online missing value imputation for high-dimensional mixed-type data via generalized factor models. *Computational Statistics & Data Analysis,* 187, 107822.

**11.** Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley.

**12.** Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

**13.** Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.

**14.** Van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman & Hall/CRC.

**15.** Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics,* 28(1), 112–118.