International Journal of Informatics and Applied Mathematics
e-ISSN:2667-6990 Vol. 5, No. 2, 1-11

Hybrid Analytic Method for Missing Data Imputation in Medical Big Data

Karima BENHAMZA[1], Nadjette BENHAMIDA[1,2],
Mohamed Ilyes BOURAHDOUN[3], and Bilel BOUDJAHEM[3]

[1] LabSTIC Laboratory, Computer Science Department, University of 8 Mai 1945, Guelma, Algeria Benhamza.karima@univ-guelma.dz
[2] LaMOS Research Unit, Faculty of Exact Sciences, University of Bejaia, Algeria Benhamida.Nadjette@univ-guelma.dz
[3] Department of Computer Science, University 8 Mai 1945, Guelma, Algeria bourahdounmohamedilyes@outlook.fr boudjehemb@gmail.com

Abstract. Compared to other traditional datasets, medical data has several hidden challenges. In fact, the possibility of missing values for certain attributes presents a great dispute for data mining researchers to make correct medical decisions. In this paper, a hybrid scheme combining the k-means method and regression analysis is proposed. A combination of these two analytical methods allows to find the best distributional model of numerical data in space and helps to predict missing data. Applied to medical data (diabetes dataset), the proposed model predicts the values with a minor error rate, which is considered very satisfactory.

Keywords: Medical Big data · Missing data · imputation · K-means · Regression.

1 Introduction

Medical big data refers to human health data that has become easier to collect with the advancement of information and communication technology. Big data in healthcare is used to enhance the quality of human life by predicting diseases and epidemics and preventing deaths.

There are many frameworks to handle big data: The Hadoop framework is an open source batch processing system that may be used to process big datasets using distributed memory. Hadoop consists of two parts: the Hadoop Distributed File System (HDFS) for storing data and MapReduce for processing it. Apache Storm Big Data is another framework that can be used with any programming language and is designed to handle large streams efficiently.

Apache Flink is an open source big data platform for streaming large volumes of data. It has accurate and powerful data streaming applications. Nevertheless, unstructured and structured data can be processed more efficiently using the Spark framework. Indeed, the execution time of data processing is faster than the Hadoop MapReduce platform.

Therefore, the results of data processing tasks in Spark are much faster than in Hadoop. At last, there is no single system that best suits all needs but Spark seems to be the best for batch processing, while Storm appears to be most appropriate for streaming [1,2,3,4].

On the other hand, Missing data in medical big data is a common and important problem that both data analysts and medical researchers are aware of. When more than 10 percent of the data is missing, the statistical analysis degrades and the results are skewed, compromising the ability to draw reliable conclusions[5].

Missing values can be caused by a number of circumstances, including data corruption due to improper maintenance or users intentionally omitting information [6]. Understanding the different types and reasons for missing data can aid in selecting the best analysis technique for dealing with missing data and, therefore, the potential bias that missing data can present [7].

Rubin categorizes missing data into three mechanisms [8,9]: data is Missing Completely At Random (MCAR) when the probability of an instance having a missing value for a variable depends neither on the known value nor on the missing data.

Data is Missing At Random (MAR), when the probability of a case's variable missing value may depend on a known value rather than the value of the missing data itself; when the probability of an instance having a variable's missing value may depend on that variable's value, the data is Missing Not At Random (MNAR).

One of the easiest ways to deal with missing data is to remove any information that contains missing data. This approach works better when there are few missing data that could introduce bias into the dataset. Furthermore, missing values in the dataset can be replaced by the mean value for numerical attributes and the mode for nominal attributes: this is the most typical method of dealing with missing data. Nonetheless, various techniques have been proposed to

replace missing values with predictive methods. This process is often referred to as "Missing data Imputation" [10,11,12,13].

This work is part of this research. A hybrid imputation strategy for dealing with missing data in medical big data is developed. The rest of the paper is organized as follows: we present related works in the second section. In the section 3, the proposed model is described. The implementation and results are exposed in Section 4. Finally, the conclusion is presented in the last section.

2 Related Works

Imputation techniques inspired by machine learning are based on predictive models that estimate missing values from the complete information available in the datasets. Well-known learning algorithms such as K Nearest Neighbor (KNN), Multilayer Perceptron (MLP), and Decision Tree (DT) algorithms have been widely used as imputation methods in various problem domains and emerging disciplines such as medical big data [14].

The KNN imputation method is the most used method [15,16,17]. Missing values are imputed using k-nearest neighbors values. Thus, the nearest neighbors are identified by minimizing a distance from which a replacement value is estimated to replace the missing attribute value.

There are four types of distance metrics: Euclidean Distance; represents the shortest distance to measure the similarity between observations. Manhattan Distance also is the sum of absolute differences between points across all the dimensions. Minkowski Distance is the generalized form of Euclidean and Manhattan Distance. Finally, Hamming Distance measures the similarity between two strings of the same length. An important parameter for the KNN method is the value of K, which is sensitive to outliers. By using the MLP model, each incomplete sample value is predicted using only the complete cases. The network must be trained on one variable each time. Various hybrid MLP models have also been proposed [18,19].

DT algorithms have also been used to impute values in databases. Missing values imputation using this method is done by building decision trees to observe the missing values of each variable, and then filling the missing values of each variable by using its corresponding tree [20,21].

Clustering imputation, such as k-means clustering, has been used widely for the treatment of missing data in the literature [22,23]. However, it has been reported that clustering methods alone are not sufficient to address this problem [24].

Significant progress has been in the development of analytic tools to estimate causal effects in the context of missing data [25]. Increased use of approaches such inverse probability weighting [26], multiple imputation [27], and likelihood-based analysis [8] greatly improved rigor over previously dominant methods. When data are missing, every statistical analysis relies on improvable assumptions about the unobserved data and the reasons they are missing [28]. Therefore, missing data imputation remains an active research area.

3 Proposed Model

In this study, a hybrid analytic scheme combining the k-means approach with regression analysis is proposed. Regression analysis is a technique for predicting causal relationships between variables that are used in predictive modeling.This method compares the impact of variables assessed on different scales and evaluates the link between dependent and independent variables.However, the regression method is extremely sensitive to outliers, which can skew the regression line and, ultimately, the predicted result.

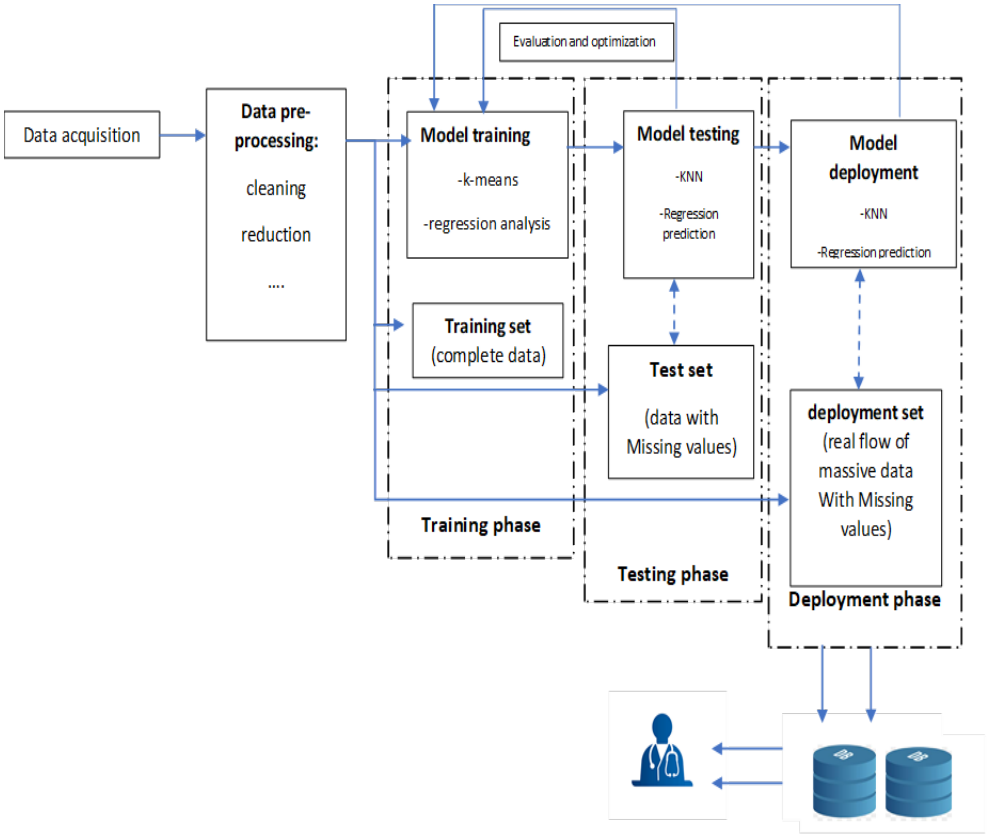

Fig. 1: Proposed Model Process

The k-means approach is initially employed to correct this deficiency. Indeed, the principle that k-means approaches allow data to be structured into suitable groupings to deal with data heterogeneity. This combination facilitates the selection of the appropriate collection of variables to structure the predictive models.

After, data pre-processing phase, the training process and testing process are the two main phases in the proposed model. In the training process, initially, the complete medical data is partitioned using K-means clustering. Then and for each partition, the linear regression analysis is performed. In the testing process, each test data with missing values is imputed with values generated by linear regression prediction (Fig.1).Each phase of model is detailed bellow.

3.1 Pre-processing Phase

In this phase, erroneous data (outliers) are identified and removed from the dataset. Records having corrupted or erroneous values, as well as data missing a high number of columns (sup 80%), are eliminated from raw data. Duplicate data is also identified and deleted. Typically, the MCAR and MAR missing data types (previously defined) are detected in the medical datasets.

3.2 Training phase

Initially, in the training process, the Complete medical data is partitioned using K-means clustering. The linear regression model is then applied to each partition. Clustering and regression analysis are performed on the dataset without missing values in order to uncover the correct patterns. K-means is a data partitioning technique that divides the data into k clusters so that intra-cluster similarity is high but inter-cluster similarity is low. The mean value of the instances in a cluster is used to determine cluster similarity.
The steps of the K-means algorithm are as follow [29].

Let X the collection of n data instances $X = \{x1, x2, x3, ..., xn\}$.
1) Choose "k" cluster centers at random as $v1, v2, ..., vk$.
2) Determine each data instance's Euclidean distance from the cluster center.
3) Assign a data instance to the cluster center with the shortest distance between it and all other cluster centers.
4) Recalculate, using (eq. 1), the new cluster center vi [$1 \leq$ i $\leq$ k] of the i^{th} cluster :

$$v_i = \frac{1}{c_i} \sum_{j=1}^{c_i} x_j \tag{1}$$

where c_i is the number of data points in the i^{th} cluster;
and x_j is the data point in the i^{th} cluster, 1 $\leq$j $\leq c_i$.
5) Recalculate each instance's distance from the newly discovered cluster center.
6) Stop if no instances were reassigned; otherwise,continue from step 3.

After the clustering step, regression analysis is used to predict the distribution of data points for each cluster. Regression is a technique for determining a linear relationship between the independent variables and the dependent variables

being studied (one or more predictors). The basic model is:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m + \varepsilon \tag{2}$$

To determine formula matrix :

$$\beta = (X^T X)^{-1} X^T y \tag{3}$$

Where :

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_m \end{pmatrix} \; X = \begin{pmatrix} 1 & x_{11} & ... & x_{1m} \\ 1 & x_{21} & ... & x_{2m} \\ ... & ... & ... & ... \\ 1 & x_{n1} & ... & x_{nm} \end{pmatrix} \; y = \begin{pmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{pmatrix}$$

Variables with missing data are used as dependent variables. Cases with complete predictor data are used to generate regression equations (eq.2and eq.3). These equations are then used to predict missing values for incomplete cases. In an iterative process, the values of the missing variables are interpolated, and then all cases are used to predict the dependent variable.

3.3 Testing phase

Each new instance is assigned to the appropriate cluster using the KNN algorithm in this step. The values generated by the linear regression models are then imputed to datasets containing missing values. There are two steps in the KNN classification [30],[31]: The first step is to figure out who the closest neighbors are. The class is determined in the second phase by using these neighbors. The resultant cluster C_i with specific regression model is used to generate missing data. This phase is proceeded as follows:

Given a data instance y,
1) Compute the distance $D(y, Ci)$, $i = 1, 2, ..k$; between the data instance y and each cluster center Ci by using Euclidean distance as:

$$D(y, x_{ij}) = \sqrt{(y - x_{ij})^T (y - x_{ij})} \tag{4}$$

2) Choose the cluster with the nearest center C_i to y, $C_i = min\{D(y, C_i)\}$; $i = 1, 2, .., k$.
3) Use the corresponding cluster C_i with specific regression model to generate missing data.

3.4 Deployment phase

The model is incorporated into the existing medical environment in this step, allowing practical imputation decisions to be made based on real data. It is the machine learning life cycle's final, dynamic, and adaptable step.

4 Results and discussion

The proposed model was implemented with the following development tools: Python V.3.2 and with lib.: pyspark, pandas, pyplot, and Apache Spark V.2.4.6.
It was tested with the medical Dataset (Diabetes.csv) available on (https://www.kaggle.com/).

The description of the execution steps of the proposed model on Spark is shown in figure 2 and described in Table 1.

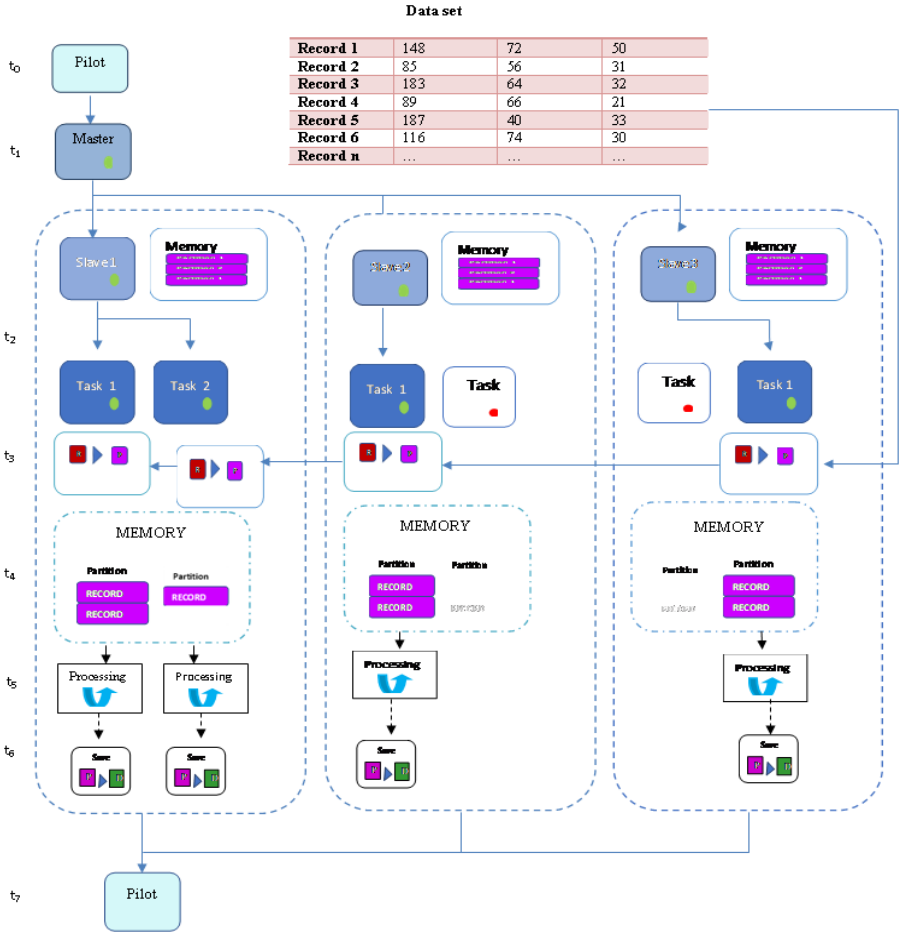

Fig. 2: Execution steps on Spark environment

Table 1: Execution steps of the proposed model on Spark

Steps	**Description**
$Step_0$	Start Execution.
$Step_1$	Connection of Pilot to the Master. Creation of a Spark session.
$Step_2$	Spark runs distributed loading of datasets through the different nodes in the cluster. Slaves create tasks to read the file. Each slave has access to the memory of node and assigns a memory partition to the task. Tasks are created according to the available resources. The Master can create multiple tasks and assign a memory partition to the task.
$Step_3$	The record is copied from the datasets to the partition during the reading process (Record R to Partition P).
$Step_4$	Each task continues by reading a portion of the data set. As the task reads rows, it stores them in the read rows partition (a dedicated partition).
$Step_5$	Once the data is loaded, Spark proceeds to the record processing by using partitioning codes (k-means, regression, and KNN).
$Step_6$	Save results.
$Step_7$	Send results to users.

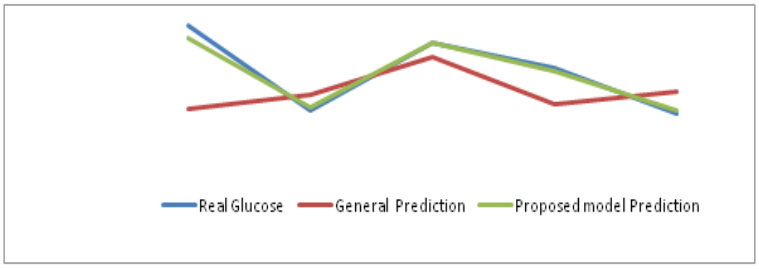

Fig. 3: Comparative curve

The analytic model offers a more exact prediction of missing data than the general prediction (General regression prediction applied to all the dataset). After partitioning with the K-means method and the application of predictive regression in each cluster, the result of the prediction of missing values is very closer to the exact real values (Fig.3). This strongly supports the proposed model.

In statistics, the Mean Squared Error (MSE)[32] is a risk function that measures the quality of an estimator. MSE is defined as an average of the square of the difference between actual and estimated values. It is derived from the square of the Euclidean distance, with lower values indicating a better fit.

In this work, MSE is used to check how close the estimated results of the proposed model are to actual values. Indeed, the lower the MSE, the closer is forecast to the actual. Figure 4 shows the MSE curves of the comparative Datasets for the imputation of missing glucose data and confirms the efficiency of the proposed model.

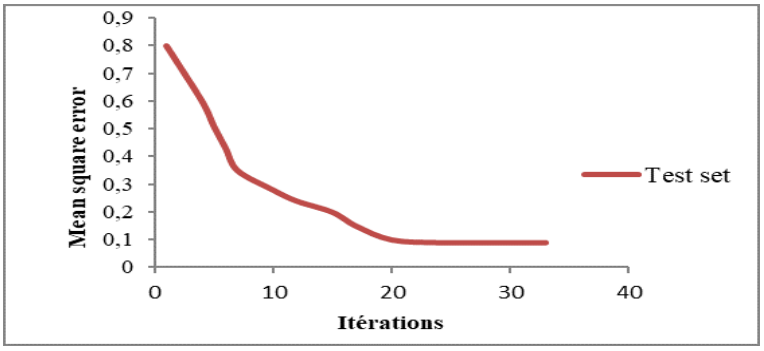

Fig. 4: Convergence of MSE curve

5 Conclusion

Missing data can cause all kinds of problems. Indeed, they reduce statistical analysis and lead to biased parameter estimates and therefore invalid conclusions. In the proposed model, k-means is first performed on the complete data set (training set). Then, and for each cluster, a regression analysis is applied. Then, each instance of the test datasets is assigned to its cluster using the KNN method. Missing data are finally imputed using the regression model of each cluster. The imputed dataset is merged with the dataset to readapt iteratively the model. The results are good and show the high performance of the proposed analytic method.

References

1. Nazari, Elham, Mohammad Hasan Shahriari, and Hamed Tabesh. "BigData analysis in healthcare: apache hadoop, apache spark and apache flink." Frontiers in Health Informatics 8.1 (2019): 14
2. Palanisamy, Venketesh, and Ramkumar Thirunavukarasu. "Implications of big data analytics in developing healthcare frameworksA review." Journal of King Saud University-Computer and Information Sciences 31.4 (2019): 415-425.
3. Kumar, Sunil, and Maninder Singh. "Big data analytics for healthcare industry: impact, applications, and tools." Big data mining and analytics 2.1 (2018): 48-57.
4. Bahri, Safa, et al. "Big data for healthcare: a survey." IEEE access 7 (2018): 7397-7408.
5. Bennett, Derrick A. "How can I deal with missing data in my study?." Australian and New Zealand journal of public health 25.5 (2001): 464-469.
6. Graham, John W. "Missing data: Analysis and design". Springer Science and Business Media, 2012.
7. Mack, Christina, Zhaohui Su, and Daniel Westreich. "Managing missing data in patient registries: addendum to registries for evaluating patient outcomes: a users guide." (2018).
8. Rubin, Donald B. Inference and missing data. Biometrika 63.3 (1976): 581-592.

9. Little, Roderick JA, and Donald B. Rubin. Statistical analysis with missing data. Vol. 793. John Wiley and Sons, 2019.
10. Ludbrook J. Outlying observations and missing values: how should they be handled? Clin Exp Pharmacol Physiol. 2008;35(56):6708.
11. Zhang Z. Missing values in big data research: some basic skills. Ann Transl Med. 2015;3(21):323.
12. Langkamp DL, Lehman A, Lemeshow S. Techniques for handling missing data in secondary analyses of large surveys. Acad Pediatr. 2010;10(3):20510.
13. Donders ART, Van Der Heijden GJ, Stijnen T, Moons KG. A gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59(10):108791
14. Jerez, Jos M., et al. "Missing data imputation using statistical and machine learning methods in a real breast cancer problem." Artificial intelligence in medicine 50.2 (2010): 105-115.
15. Hruschka ER, Hruschka ER, Ebecken NFF. Towards efficient imputation by nearest-neighbors: a clustering-based approach. In: AI 2004: advances in artificial intelligence, vol. 3339 of lecture notes in computer science. Springer Berlin/Heidelberg; 2005. p. 51325
16. Zhang, Shichao. "Nearest neighbor selection for iteratively kNN imputation." Journal of Systems and Software 85.11 (2012): 2541-2552.
17. Pujianto, Utomo, Aji Prasetya Wibawa, and Muhammad Iqbal Akbar. "K-nearest neighbor (k-NN) based missing data imputation." 2019 5th International Conference on Science in Information Technology (ICSITech). IEEE, 2019.
18. Silva-Ramrez, Esther-Lydia, Rafael Pino-Mejas, and Manuel Lpez-Coello. "Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns." Applied Soft Computing 29 (2015): 65-74.
19. Purwar, Archana, and Sandeep Kumar Singh. "Hybrid prediction model with missing value imputation for medical data." Expert Systems with Applications 42.13 (2015): 5621-5631.
20. Twala B. An empirical comparison of techniques for handling incomplete data using decision trees. Appl Artif Intell. 2009;23(5):373405.
21. Gimpy D, Rajan Vohra M. Estimation of missing values using decision tree approach. Int J Comput Sci Inf Technol. 2014;5(4):521620.
22. Zhang, Shichao, et al. "Missing value imputation based on data clustering." Transactions on computational science I. Springer, Berlin, Heidelberg, 2008. 128-138.
23. Zhang, Zhaoyang, Hua Fang, and Honggang Wang. "Multiple imputation based clustering validation (miv) for big longitudinal trial data with missing values in ehealth." Journal of medical systems 40.6 (2016): 1-9.
24. Emmanuel, Tlamelo, et al. "A survey on missing data in machine learning." Journal of Big Data 8.1 (2021): 1-37.
25. Enders CK. Applied missing data analysis. New York: The Guilford Press; 2010.
26. Carpenter JR, Kenward MG, Vansteelandt S. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2006;169(3):571584.
27. Beale EM, Little RJ. Missing values in multivariate analysis. Journal of the Royal Statistical Society Series B (Methodological) 1975:129145.
28. Carpenter JR, Kenward MG. Missing data in randomised controlled trials a practical guide, 2007.
29. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., Wu, A. Y. "An efficient k-means clustering algorithm: Analysis and implementation."

IEEE transactions on pattern analysis and machine intelligence 24.7 (2002): 881-892.
30. Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." IEEE transactions on information theory 13.1 (1967): 21-27.
31. Gou, Jianping, et al. "A generalized mean distance-based k-nearest neighbor classifier." Expert Systems with Applications 115 (2019): 356-372.
32. Allen, David M. "Mean square error of prediction as a criterion for selecting variables." Technometrics 13.3 (1971): 469-475.