The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 2023

Volume 23, Pages 26-33

ICRETS 2023: International Conference on Research in Engineering, Technology and Science

# Data Cleaning in Medical Procurement Database: Performance Comparison of Data Mining Classification Algorithms for Tackling Missing Value

**Amarawan Pentrakan**
Prince of Songkla University

**Arbee L. P. Chen**
Asia University

**Abstract**: Data cleaning is an important process for improving the quality of decision-making information. One of today's popular cleaning tools is data mining techniques. In this paper, we focused on using data mining classification algorithms to resolve missing values in medical purchasing databases. To serve this purpose, the predictive performance of four different classifiers: Decision Tree, Naïve Bayes, K-Nearest Neighbor, and Support Vector Machine (SVM) were compared in this study. We used 2,311 medical data records from procurement database in Thailand between July 2019 and December 2019 in the experimental process. We also discussed the function of feature selection and test options that support analysis to improve model performance. The results showed that the SVM algorithm outperforms with a maximum accuracy of 89.61%. Additionally, we discussed the strengths and weaknesses of these data mining techniques for data cleaning and future research.

**Keywords:** Data mining techniques, Classification algorithms, Medical procurement database, Missing values

## Introduction

The use of medical procurement data in clinical research and management has increased dramatically, but the missing value resources pose data cleaning challenges (Sakly et al., 2022). Although nowadays there are many popular data cleaning tools in various domains, they often process all variables uniformly (Shi et al., 2021), meaning that they might not serve well for medical procurement database because there is specific information to the variables that must be considered. Thus, in this paper, our study proposes a data cleaning tools to correct missing values in the data variables of medical procurement with data mining knowledge widely-used techniques taken into consideration.

Medical procurement data may include a wide range of data, including purchasing products, suppliers, related services, payment information, and the purchase conditions of health facilities (Pentrakan et al., 2023). In the case of medication, it may occur after the product has been selected for inclusion on the health facility list and/or the national reimbursement list. The data collected from real practice is frequently incomplete (Wang et al., 2021), including lots of typos, errors, and missing values. Therefore, the abundance of data resources often raises the challenge of data cleaning. This results in large amounts of time and budget for many analyses.

Data cleaning is a fundamental step of data analysis with the goal of cleaning up raw data (Xu et al., 2015). This is a very important process because all analyses require quality data to find the reliable results. The cleaning process can begin with identifying information that is incomplete or unreasonable, and then improves quality by correcting detected errors and omissions. In practice, it is generally found that data cleaning and preparation take up about 80% of the total data engineering effort (Zhang et al., 2003). This is a crucial research problem for

many analysts and organizations. Consequently, a powerful tool for medical procurement systems that cleans data more accurately and faster, can save time and increase efficiency in data analysis.

With the need for effective cleaning of medical data, many studies have been conducted on the general framework to only intend to assess the quality of the data, but there are no data cleaning procedures (Shi et al., 2021). While other domains have been extensively investigated in data cleaning, one interesting thing is that the problem of estimating missing values can be processed by using data mining techniques. Data mining is used to extract knowledge from existing data. It can be used to discover things that were previously unknown and retrieve interesting patterns and related relationships in a given dataset, including classification rules or trees, sequence modeling, clustering, regression, dependency, and so forth. This knowledge can be utilized in calculating the estimates for missing values. However, each data mining algorithm may have its own specific performance for each task, and no one can be effective for all data (Mandal & Jana, 2019). The chosen algorithm used for each domain depends on the unique variables in each field and its constraints.

Our study aims to examine the performance of data mining techniques for handling missing values by comparing the performance of four different classification algorithms: Decision Tree, Naïve Bayes, K-Nearest Neighbor, and Support Vector Machine. The study is evaluated in terms of accuracy, error distribution, and the time spent building and testing the model. We also discuss the function of feature selection and the testing options that underpin the analysis in this experimental process. The results show both the pros and cons of those algorithms. The expected outcome is to assist users and decision-makers in selecting the best techniques to resolve missing values in medical procurement databases.

**Description of Algorithms and Tools**

Several open-source data mining software are available for free on the Web. In this experimental study, we applied the Waikato Environment for Knowledge Analysis (WEKA) tool version 3.9.3 for processing classification models. WEKA has become widely used as a toolkit for data mining tasks and was originally established by the University of Waikato (New Zealand). This software contains a large function of data mining techniques and can be accessed through standard terminal applications. It contains several techniques for pre-processing, classification, clustering, association rules, visualization, and regression (Holmes et al., 1994). It also supported versions for Windows, Linux, and MAC operating systems. WEKA, therefore, is currently popular with academics and widely applied for teaching, research, and industrial applications (Hussain et al., 2018). Additionally, we also use Tableau software for data visualization (Chabot et al., 2003). This tool can provide an accessible way to understand patterns and results (Batt et al., 2020).

In this article, we determine four classifier learning algorithms that are implemented in WEKA (Sahoo & Kumar, 2012). Three algorithms consist of Decision Tree (DT), Naïve Bayes (NB), and Support Vector Machine (SVM) techniques, which are parametric classifiers based on statistical probability distributions. The other one, the K-Nearest Neighbor (KNN), is a nonparametric classifier based on the probability density function. A brief description of each algorithm and how it works in WEKA is detailed as follows: (1) DT is one of the data mining techniques that classifies recursive training data instances by deep-first or broad-first greedy methods until all data instances belong to a specific class. Of them, there are many advanced DT algorithms such as, CART, ID3, C4.5, C5.0, etc. This paper uses the C4.5 algorithm developed by Ross Quinlan, which is called J48 in WEKA. It can handle both categorical and numerical entries, and can generate thresholds to use for continuous-label classification (Patil et al., 2009); (2) NB works in conjunction with Bayes theorem-based statistical methods (Bouckaert, 2008). The probabilities for each class are calculated from the given data and are considered independent of each other. This may be called conditionally independent. With this classification, it is possible to predict the probability of group membership and can remove irrelevant data; and (3) KNN is an instance learning method also known as lazy learning. It stores all available data and classifies the new data points based on their similarity (Steinbach & Tan, 2009). This means that when new information appears, the dataset can be easily classified into categories. Therefore, it runs very quickly during training, but it takes more computational time to classify new instances that come into the model (Zhang, 2010). Additionally, it is susceptible to noise in the dataset and memory limitations; and (4) the SVM technique is a classifier method that performs classification tasks by constructing hyperplanes. It was generally developed for binary classification problems, but now extensions to this technique have been made to support multi-class classification (Rajvanshi & Chowdhary, 2017). In WEKA, this technique can be selected at the SMO function, which stands for Sequential Minimal Optimization. It is the specific efficient optimization algorithm used inside the SVM implementation and works by finding a line that best separates the data into the two groups. This is done by an optimization process that considers only those data instances in the training data set that are closest to the line

that best separates the classes (Pentrakan, 2021). These instances are known as support vectors. This calculates the maximal margin that can reduce the generalization error.

In addition, this study also examines two test option techniques to test whether there is an improvement in the accuracy measure when using suitable test options. First is the Cross-Validation (k-fold) option (Browne, 2000). The performance measure designed by k-fold cross-validation is the average of the values computed in the loop. The training set is split into k smaller groups, and then using k-1of the folds trained model as training data. Then, the resulting model is validated in the remaining part of the data. This is used as a test set to estimate a performance measure, such as accuracy. The second is the Percentage Split option. The dataset is randomly split into two disjoint sets (Kabakchieva, 2013). The first set is called the training set that works as knowledge form. This extracted knowledge is used to test against the second set which is called the test set.

## Method

### Dataset and Process

The study dataset contains medical procurement data in Thailand from July 2019 to December 2019. We have listed all relevant features in Table 1. A total of 2,749 transaction records were associated with 2,311 complete records and 438 incomplete records. If the analyst excludes those records from the analysis, they may lose important values and cause significant errors in calculating the average price of each medical product. Therefore, in this study, we aimed to investigate the effectiveness of data mining techniques for handling missing values of the units per pack (SIZE feature) for this sample dataset.

To serve this purpose, we used 2,311 complete records without missing values to examine the model of each classifier. We hypothesized that six features that exhibit in this dataset (the segmented purchasers, the group of generic product and trade product names, procurement methods, suppliers, and total purchase budgets) could be used in developing the model to estimate the missing values of pack sizes for medical products effectively.

Table 1. Features and the definitions of a dataset

| Features | Definition | Distinct values | Total records | Missing values (%) |
|---|---|---|---|---|
| DEPT | Departments who purchase the medical product | 394 | 2,747 | 2 (0.07%) |
| GEN | Generic product name (e.g., Parecoxib 40 mg injection) | 4 | 2,746 | 3 (0.11%) |
| TRAD | Trade product name or brand (e.g., DYNASTAT® ) | 36 | 2,746 | 3 (0.11%) |
| METH | Procurement method (e.g., bidding method) | 2 | 2,749 | 0 (0%) |
| COM | Company or supplier who sells the medical product | 38 | 2,749 | 0 (0%) |
| PRICE | The purchasing price per pack (Thai Baht) | 572 | 2,749 | 0 (0%) |
| SIZE* | The number of units per pack (e.g., 30 tablets per box) | 12 | 2,319 | 430(15.6%) |

*Output variable corresponding to the class labels of pack size for medical products.

We used six purchasing properties as input features: DEPT, GEN, TRAD, METH, COM, and PRICE according to the different descriptions and values shown in Table 1: DEPT is the name of 394 purchasing departments that make the decision to buy medical products for hospitals; GEN is the name of a generic product with four different names according to the anatomical chemotherapy (ATC) classification code; and TRAD is the name of a trade product with 36 different names. METH is the procurement method that includes two approaches used in the procurement system; COM is the winning company or supplier associated with 38 different companies that is recorded to sell medical products to hospitals; and PRICE is the total purchase budget consisting of 572 various labels. These six features were used to develop a model to estimate the missing value of package sizes (SIZE) for medical products.

In the pre-processing step, we have discretized the continuous values of the SIZE feature into the class label of nominal values. The price values of PRICE feature were also discretized into class labels of price ranges for medical products. To remove irrelevant and redundant features, all features were examined using the feature selection method. This was intended to maximize classification accuracy. To do this process, we used the function of wrapper subset evaluator (Karegowda et al., 2010) implemented in the WEKA. The features selected depend on the classifier that builds the model. It worked to find the smallest subset of the attribute by selecting important features for the underlying clusters based on the criteria of the algorithm. Then, in this step, we examined two test options (cross-validation and the percentage split) to determine the best options for classification algorithms.

### Model Evaluation Metrics

After the model was developed and validated by test options, we can obtain feedback from metrics that can explain the performance of the model. Generally, the performance comparison of the four different algorithms was measured in terms of measuring accuracy and error distribution (Galdi & Tagliaferri, 2018). Accordingly, this study used several relevant metrics, including accuracy rate, F-measure, precision, recall (sensitivity), mean absolute error (MAE), and root mean squared error (RMSE). We also used the kappa statistic (McHugh, 2012) and an area under the receiver operating characteristic curve (Hamel, 2009). In addition, the time spent building and testing the model were represented in the results of our study.

## Results and Discussion

In this section, we discuss the empirical analysis of the study. First, we represent the descriptive statistics of variables to give information about relevant variables. Then, we discuss the results of the feature and test option analyses followed by the performance analysis.

### Descriptive Statistics

In our analysis, we used the complete input data of 2,311 records to develop the model. The sample dataset consisted of four generic products frequently purchased from July 2019 to December 2019. All medical products listed in the purchasing dataset were defined according to the lowest level of the Anatomical Chemotherapy (ATC) classification system (Skrbo, 2004), as shown in Table 2.

Table 2. Descriptive statistics of medicine used in this study

| ATC code of medical products | Package size | Number of records | Number of trade products | Number of purchasing departments | Number of selling companies | Range of purchasing budgets |
|---|---|---|---|---|---|---|
| M01AH04 | 1 unit/box | 258 | 2 | 47 | 4 | 1,939 – 5,816,400 |
| | 5 units/box | 416 | 2 | 79 | 12 | 969 – 930,643 |
| | 10 units/box | 9 | 2 | 5 | 3 | 969 – 963,942 |
| | 50 units/box | 5 | 2 | 1 | 1 | 19,388 – 198,388 |
| | 100 units/box | 3 | 1 | 1 | 1 | 19,388 – 19,388 |
| A02BA02 | 1 unit/box | 335 | 17 | 107 | 9 | 200 – 97,200 |
| | 10 units/box | 9 | 4 | 9 | 4 | 2,000 – 20,000 |
| | 50 units/box | 8 | 7 | 1 | 6 | 792 – 40,000 |
| | 100 units/box | 348 | 16 | 125 | 10 | 190 – 126,720 |
| | 250 units/box | 6 | 4 | 4 | 5 | 160 – 10,000 |
| | 500 units/box | 145 | 3 | 86 | 3 | 400 – 99,000 |
| | 1000 units/box | 30 | 3 | 18 | 4 | 600 – 478,932 |
| C10AA07 | 1 unit/box | 129 | 6 | 23 | 6 | 6,420 – 832,032 |
| | 28 units/box | 328 | 7 | 49 | 11 | 706 – 499,647 |
| | 30 units/box | 32 | 1 | 9 | 3 | 9,540 – 256,800 |
| | 56 units/box | 7 | 3 | 6 | 3 | 23,946 – 360,323 |
| | 84 units/box | 4 | 1 | 2 | 1 | 12,947 – 25,894 |
| C10AA01 | 1 unit/box | 53 | 7 | 29 | 5 | 4,000 – 1,144,044 |
| | 100 units/box | 81 | 4 | 36 | 2 | 2,000 – 500,000 |
| | 250 units/box | 6 | 5 | 5 | 3 | 12,000 – 497,250 |
| | 500 units/box | 2 | 2 | 2 | 2 | 7,200 – 25,145 |
| | 1000 units/box | 97 | 4 | 47 | 3 | 4,200 – 288,000 |

### Feature Analysis

The alternative of selecting features depends on both the algorithm used and the type of data given. In this study, we used the wrapper subset estimator to select relevant features based on the algorithm applied in the model. As shown in Table 3, the results show that the features of generic products (GEN) and the purchase budgets (PRICE) were selected for all algorithms. The feature of company supplier (COM) was further selected in the Decision Tree and Support Vector Machine applications. Two additional features, the departmental group

(DEPT) and trade product (TRAD), showed important contributions to the Naïve Bayes algorithms while the feature of the procurement method (METH) was not selected for the entire algorithm's operation.

Table 3. Relevant features selected for four different algorithms

| Algorithms | Selected features | Number of features |
|---|---|---|
| Naïve Bayes (NB) | GEN, COM, PRICE, DEPT, TRAD | 5 |
| Decision Tree (DT) | GEN, COM, PRICE | 3 |
| Support Vector Machine (SVM) | GEN, COM, PRICE | 3 |
| K-Nearest Neighbor (KNN) | GEN, PRICE | 2 |

The optimum split of the test, validation, and train set depends upon the features, algorithms, and dimension of the given data. Therefore, after selecting the relevant features, we then examined two widely-used test option techniques: cross-validation (k-fold) and percentage split options for algorithms.

As shown in Figure 1, the results showed different trend patterns between those two options. First, when performing cross-validation, we tried different number of folds settings. The results presented differences in performance. The increased number of folds smoothly increased accuracy, while SVMs with 10 folds (k=10) provided excellent performance. Second, increasing the percent splitting of the data for training greatly improves accuracy. We found that SVMs with 90% splitting for training provided the highest accuracy and were more accurate than those using cross-validation techniques. That means, if we used 10% data for testing and the remaining 90% data for training in this research, the SVM technique could have approximately 89.61% of the best properly classified instances, followed by KNN, DT, and NB with 88.74%, 86.15%, and 80.09, respectively.

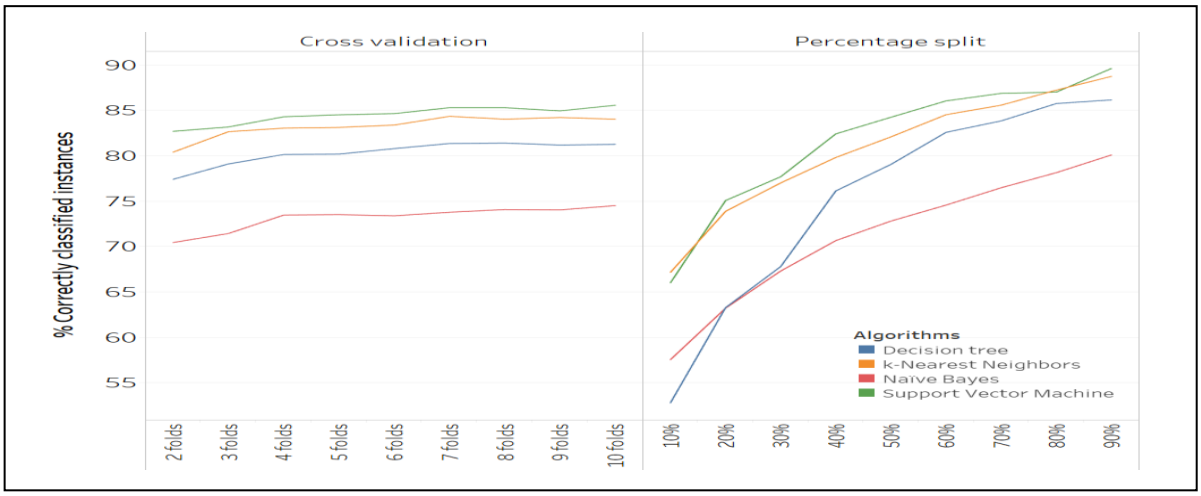


Figure 1. Comparing the percentage of correctly classified instances between two different test options

**Performance Evaluation**

After analyzing the features, those features selected for each algorithm were used to develop a model for classifying the package size of medical product corresponding to the twelve class labels, described in Table 1. The results of model testing show that all algorithms provide good model performance in our metrics. The evaluation results obtained after testing four algorithms using the 90 percent split test were shown in Figure 1. The proposed model from SVM outperformed the prediction accuracy and has significant agreement for the kappa statistical coefficient. In addition, the model had great precision, sensitivity (recall), and F-measurement. However, after we calculated the time spent building and testing the model, the SVM took much longer than other algorithms while NB and KNN provided very high speeds in less than 0.01 seconds in modeling.

In our study, there was a limitation in our proposed model. The results of model performance were based on the specific characteristics of medical products procured in the given time period. Different types of products and study periods could yield different findings. Therefore, future studies may use more types of products to further examine these classification algorithms.

Table 4. Results of the accuracy measures tested in four classification algorithms

| Parameters | DT | NB | KNN | SVM |
|---|---|---|---|---|
| Accuracy | 0.862 | 0.801 | 0.887 | 0.896 |
| Kappa statistic | 0.814 | 0.727 | 0.849 | 0.861 |
| Precision | 0.907 | 0.876 | 0.952 | 0.952 |
| Recall | 0.951 | 0.961 | 0.971 | 0.971 |
| F-Measure | 0.929 | 0.917 | 0.962 | 0.962 |
| ROC area | 0.945 | 0.972 | 0.995 | 0.981 |
| MAE | 0.012 | 0.015 | 0.009 | 0.048 |
| RMSE | 0.080 | 0.088 | 0.068 | 0.153 |
| Time to build model (seconds) | 0.10 | <0.01 | <0.01 | 14.65 |
| Time to test model (seconds) | 0.01 | 0.02 | 0.02 | 0.88 |

## Conclusion

Tackling missing data in medical procurement databases is very important because it can often bias analyzes (Groenwold & Dekkers, 2020). Medical purchase data also needs a powerful tool to solve this problem in order to improve analytics and results. As an example of drug procurement data available in the Thai government's procurement system, we found many missing values for package sizes that cause big problems for analysts. In this paper, we therefore examined four models developed from data mining classification techniques to predict missing values for these package sizes. To do so, we used the complete medical data samples as inputs for model training and used six relevant purchasing features as input features (Pentrakan et al., 2022). We also tried to improve the prediction accuracy of model by using these algorithms with the wrapper feature selection function. The results in Table 4 show that the data mining classification algorithm can provide good performance in predicting the missing values of a given data set. As summarized in Figure 2, our study supports that Support Vector Machine (SVM) with 90% splitting of data for training can stand out for being more accurate than the others (Mustaqeem et al., 2018), although this algorithm takes more time for modeling.

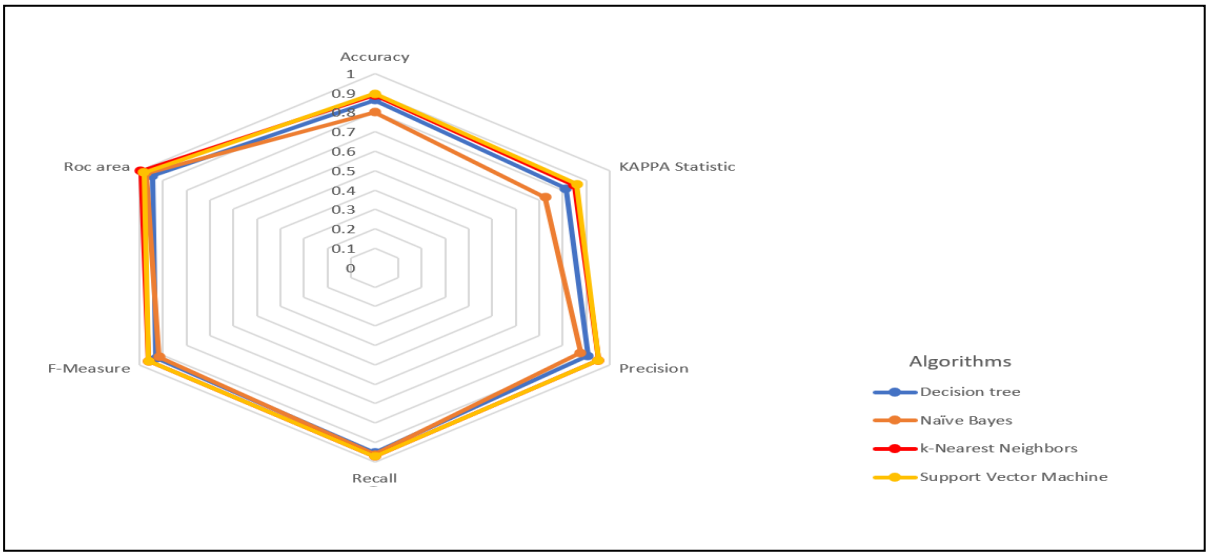


Figure 2. Comparing the accuracy measures of four classification algorithms

Our study application allows medical analysts to resolve the problem of missing values and perform better analyses. This can save a lot of time cleaning up and make their data more reliable. Although some methods have been developed to estimate missing values in many sectors, there is limited evidence and application for the medical procurement system. Therefore, the findings of our study might be useful in applying them to this system and other countries that are facing the challenge of missing medical supply information.

## Scientific Ethics Declaration

The authors declare that the scientific ethical and legal responsibility of this article published in EPSTEM journal belongs to the authors.

## Acknowledgements or Notes

* This article was presented as a poster presentation at the International Conference on Research in Engineering, Technology and Science ( www.icrets.net ) held in Budapest/Hungary on July 06-09, 2023.

*We also thank the Comptroller General's Department of Thailand for providing access to the procurement data in pharmaceuticals. We thank the Thai Health Information Standards Development Center (THIS) in Thailand for providing the updated data of Thai Medicine Terminology (TMT).

## References

Batt, S., Grealis, T., Harmon, O., & Tomolonis, P. (2020). Learning Tableau: A data visualization tool. *The Journal of Economic Education*, *51*(3-4), 317-328.

Bouckaert, R. R. (2008). Bayesian network classifiers in weka for version 3-5-7. *Artificial Intelligence Tools*, *11*(3), 369-387.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*(1), 108-132.

Chabot, C., Stolte, C., & Hanrahan, P. (2003). Tableau software. *Tableau Software*, *6*.

Galdi, P., & Tagliaferri, R. (2018). Data mining: accuracy and error measures for classification and prediction. *Encyclopedia of Bioinformatics and Computational Biology*, *1*, 431-436.

Hamel, L. (2009). Model assessment with ROC curves. In *Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 1316-1323). IGI Global.

Holmes, G., Donkin, A., & Witten, I. H. (1994, 29 Nov.-2 Dec. 1994). WEKA: a machine learning workbench. *Proceedings of ANZIIS '94 - Australian New Zealnd Intelligent Information Systems Conference*.

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447-459.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, *13*(1), 61-72.

Karegowda, A. G., Jayaram, M., & Manjunath, A. (2010). Feature subset selection problem using wrapper approach in supervised learning. *International Journal of Computer Applications*, *1*(7), 13-17.

Mandal, L., & Jana, N. D. (2019, 13-15 Dec. 2019). A comparative study of naive bayes and k-NN algorithm for multi-class drug molecule classification. *2019 IEEE 16th India Council International Conference*.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276-282.

Mustaqeem, A., Anwar, S. M., & Majid, M. (2018). Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants. *Computational and Mathematical Methods in Medicine*. Article ID 7310496. https://doi.org/10.1155/2018/7310496.

Patil, B. M., Toshniwal, D., & Joshi, R. C. (2009, 6-7 March 2009). Predicting burn patient survivability using decision tree in WEKA environment. *2009 IEEE International Advance Computing Conference*.

Pentrakan, A., Lin, K. H., Sriphon, T., Wang, J. Y., & Wong, W. K. (2022). The impact of pharmaceutical electronic bidding procurement on prices of medicines: A systematic review. *Indian Journal of Pharmaceutical Sciences*, 86-98.

Pentrakan, A., Wang, J.-Y., & Wong, W.-K. (2023). The impact of centralized electronic bidding system on procurement prices for generic medicines: A case study from Thailand. *Songklanakarin Journal of Science and Technology*, *44*, 1532-1538.

Pentrakan, A., Yang, C.-C., & Wong, W.-K. (2021). How well does a sequential minimal optimization model perform in predicting medicine prices for procurement system? *International Journal of Environmental Research and Public Health*, *18*(11), 5523.

Rajvanshi, N., & Chowdhary, K. (2017). Comparison of SVM and naïve Bayes text classification algorithms using WEKA. *International Journal of Engineering Research and*, *6*(09), 141-143.

Sakly, H., Said, M., Seekins, J., & Tagina, M. (2022). Big data and artificial intelligence for e-health. In N. Rezaei (Ed.), *Multidisciplinarity and interdisciplinarity in health* (pp. 525-544). Springer International Publishing.

Shi, X., Prins, C., Van Pottelbergh, G., Mamouris, P., Vaes, B., & De Moor, B. (2021). An automated data cleaning method for electronic health records by incorporating clinical knowledge. *BMC Medical Informatics and Decision Making*, *21*(1), 267.

Skrbo, A., Begović, B., & Skrbo, S. (2004). Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes. *Medicinski Arhiv*, *58*(1 Suppl 2), 138-141.

Steinbach, M., & Tan, P.-N. (2009). kNN: k-nearest neighbors. In *The top ten algorithms in data mining* (pp. 165-176). Chapman and Hall/CRC.

Wang, J., Yang, Y., Xu, L., Shen, Y., Wen, X., Mao, L., Wang, Q., Cui, D., & Mao, Z. (2021). The impact of national centralized drug procurement policy on the use of policy-related original and generic drugs in public medical institutions in China: A difference-in-difference analysis based on national database. *MedRxiv*, 2021-06.

Xu, S., Lu, B., Baldea, M., Edgar, T. F., Wojsznis, W., Blevins, T., & Nixon, M. (2015). Data cleaning in the process industries. *Reviews in Chemical Engineering*, *31*(5), 453-490.

Zhang, M.-L. (2010). A k-nearest neighbor based multi-instance multi-label learning algorithm. *22nd IEEE international conference on tools with artificial intelligence.*

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5-6), 375-381.

## Author Information

**Amarawan Pentrakan**
Prince of Songkla University, Thailand
15 Karnjanavanich Rd., Hat Yai, Songkhla City, Thailand 90110, Thailand
Contact e-mail: *amarawan.p@psu.ac.th*

**Arbee L.P. Chen**
Asia University, Taiwan
No. 500, Liufeng Rd, Wufeng District, Taichung City, Taiwan 413, Taiwan

**To cite this article:**

Pentrakan, A. & Chen, A.L.P. (2023). Data cleaning in medical procurement database: Performance comparison of data mining classification algorithms for tackling missing value. *The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 23*, 26-33.