DOI: 10.24996/ijs.2023.64.5.32

ISSN: 0067-2904

Application of Data Mining and Imputation Algorithms for Missing Value Handling: A Study Case Car Evaluation Dataset

Wahyu Widyananda*, Muhammad Fauzan Edy Purnomo, Muhammad Aswin, Panca Mudjirahardjo, Sholeh Hadi Pramono

Electrical Engineering Department, Brawijaya University, East Java, Indonesia

Received: 21/12/2021 Accepted: 5/2/2022 Published: 30/5/2023

Abstract

Data mining is a data analysis process using software to find certain patterns or rules in a large amount of data, which is expected to provide knowledge to support decisions. However, missing value in data mining often leads to a loss of information. The purpose of this study is to improve the performance of data classification with missing values, precisely and accurately. The test method is carried out using the Car Evaluation dataset from the UCI Machine Learning Repository. RStudio and RapidMiner tools were used for testing the algorithm. This study will result in a data analysis of the tested parameters to measure the performance of the algorithm. Using test variations: performance at C5.0, C4.5, and k-NN at 0% missing rate, performance at C5.0, C4.5, and k-NN at 5–50% missing rate, performance at C5.0 + k-NNI, C4.5 + k-NNI, and k-NN + k-NNI classifier at 5–50% missing rate, and performance at C5.0 + CMI, C4.5 + CMI, and k-NN + CMI classifier at 5–50% missing rate, The results show that C5.0 with k-NNI produces better classification accuracy than other tested imputation and classification algorithms. For example, with 35% of the dataset missing, this method obtains 93.40% validation accuracy and 92% test accuracy. C5.0 with k-NNI also offers fast processing times compared with other methods.
Keywords: C5.0, k-NNI, Data Mining, Missing Value Handling, R Studio, Rapid Miner.

1. Introduction

When consumers consider buying a car, several factors can influence their decision. Safety, cost, and luxury are important factors that must be considered when buying a car [1]. Assessing the cost and quality of a new product in the marketing stage of development allows a more accurate prediction of consumer acceptance of the product or service [2]. Collecting data on car purchases regarding these factors is needed to evaluate cars based on consumer interests, which results in a car evaluation dataset. Data mining algorithms have the ability to analyze data in various research fields [3] [4] [5] [6], and classification is one of the main roles in data mining that can be applied to car evaluation datasets to create predictive models based on consumer interest. Several studies have been carried out on car evaluation to make prediction models by applying classification algorithms [7] [8] [9]. Datasets with missing values are a common problem in data mining, which can lead to a loss of information and result in poor predictive models [10]. A decision tree is one of the classification algorithms that can handle missing values during the classification process; besides that, there is also a data imputation technique. This is one of the techniques used to handle missing values in a data set.

*Email: wahyu.widyananda111@gmail.com

Various studies have been conducted to solve the problem of missing values in classification using the imputation method. Testing Decision Tree C4.5 without adding more imputation methods resulted in better prediction accuracy than adding listwise deletion and mean imputation methods to data with missing values [11]. Decision Tree C5.0 is an improvement over the Decision Tree C4.5 algorithm. The Class Mean Imputation method with the Decision Tree C5.0 algorithm performed better than Mean Imputation on a dataset with many categorical variables [12]. k-Nearest Neighbor Imputation (k-NNI) is an imputation method based on the k-Nearest Neighbor classification algorithm. k-NNI can improve Decision Tree C4.5 performance on small software projects [13] and student records [14].

Although there have been many studies related to missing value datasets using classification and imputation methods, there are no studies on the presence of missing values in car evaluation. In fact, datasets with missing values can lead to a loss of information and result in poor predictive models. Poor predictive models can lead to wrong decisions. The k-NNI approach combined with Decision Tree C5.0 can be used to improve the prediction accuracy of the Car Evaluation dataset when missing values are present.

In contrast to previous studies, this study aims to improve the classification performance of the Car Evaluation dataset with missing values' presence using Decision Tree C5.0 and k-Nearest Neighbor Imputation (k-NNI). A comparison of other classification algorithms such as Decision Tree C4.5 and k-NN, combined with Class Mean Imputation and k-NNI, was also carried out. Parameters used to measure classification performance include performance accuracy and average processing time. This research was conducted to overcome the problem of data classification with missing values to make decisions quickly, precisely, and accurately on classification data.

The research specifications are designated to describe the proposed approach and its limitations due to the fact that it is both a phase and a package of the main design. They are summarized as follows:

- The dataset used in this study is a real-world car evaluation dataset from the UCI Machine Learning Repository. This dataset has 1728 samples with six categorical type attributes, including purchase price, maintenance price, number of doors, person capacity, luggage boot size, and estimated security.
- Missing values in the dataset are artificially generated using the MCAR mechanism with ratios of 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, and 50%.
- R Studio 1.3.1073 is used for the data imputation process, while RapidMiner 9.7.002 is used for the data classification and analysis process.
- The proposed approach uses 3 data mining classification algorithms (Decision Tree C5.0, Decision Tree C4.5, and k-NN) combined with 2 data imputation algorithms (k-NNI and CMI) to offer better decision support results.

Literature Review

A. Data Grouping

The grouping of data in data mining is divided into two categories: classification and clustering. Classification is the grouping of data that requires training data (supervised). While clustering is data grouping without the need for training data (unsupervised), [15] explained that clustering is the process of dividing unlabeled data into groups of data that have similarities. Each data group (cluster) consists of objects that have similarities with each other, and each cluster has dissimilarities with other clusters. Clustering is commonly used in multivariate data analysis.

B. Data Mining

Data mining is a field of several scientific fields that combines techniques from machine learning, pattern recognition, statistics, databases, and visualization for handling problems of retrieving information from large databases [16]. Broadly speaking, data mining can be grouped into two main categories, namely [17]:

a) descriptive mining, which is a process to find important characteristics of data in a database. Data mining techniques included in descriptive mining are clustering, association, and sequential mining.

b) predictive, which is the process of finding patterns from the data by using several other variables in the future. One of the techniques contained in predictive mining is classification.

C. **Classification**

Classification is the process of finding a model or function that explains or distinguishes a concept or data class, with the aim of being able to estimate the class of an object whose label is unknown [18]. Classification is a learning function that maps (classifies) an element (item) of data into one of several predefined classes. The input data for classification is a collection of records. Each record is known as an instance, defined by a tuple (x,y), where x is a set of attributes and y is a specific attribute, which is expressed as a class label (also known as a category or target attribute).

Some of the classification techniques used are decision trees, k-nearest neighbors, neural networks, support vector machines, and naive Bayes classifiers. Each technique uses a learning algorithm to identify the model that provides the most suitable relationship between the attribute set and the class label of the input data. K-nearest neighbor is a distance-based classification technique [19]. It searches the pattern space based on the training samples, which are close to unknown samples. This technique produces a higher level of accuracy than the naive Bayes technique in classifying Parkinson's disease [20] and is better than the naive Bayes technique and SVM for classifying news [21]. Decision Tree uses several algorithms to generate decision model patterns, including ID3, C4.5, and C5.0. The ID3 algorithm uses information-gain calculations in the formation of a decision tree and can only classify categorical data. The C4.5 algorithm is a development of the ID3 algorithm. Unlike the previous algorithm, C4.5 uses a gain ratio calculation in the formation of a decision tree, can classify continuous and categorical data types, and can handle training data sets with missing values [22]. The C5.0 algorithm is a development of the C4.5 algorithm. The C5.0 algorithm again uses information gain calculations to form a decision tree and can also handle data with missing values. The C5.0 algorithm has a much lower error rate for the prediction case [23].

D. Missing data imputation

Missing data is a condition where some features are lost in the dataset. Missing data can be caused by system errors, such as no response from sensors or input receiving devices. It can also be caused by human errors such as incomplete data entry in the database or respondents' misunderstandings in filling out questionnaires in large-scale surveys so that they pass through the form provided. Existing methods in data mining can only process data that has complete features, so special handling is needed for this problem.

There are 3 methods used for handling missing data, namely: case deletion, parameter estimation, and imputation techniques [24]. Case deletion is the easiest method, namely deleting data that contains missing. The weakness of this method is that it is possible to delete important information when missing data is deleted. The imputation technique is a method of handling missing data that is more widely studied. Data imputation is estimating the value of missing data by getting a pattern from data that has complete features. Some popular imputation methods are mean, median, mode, clustering, k-NN imputation, and class mean imputation [25]. In K-NN imputation, filling in missing values is done by taking into account the vector distance between attributes [26]. In class mean imputation, the missing value will be replaced by the mean value of all available values in a related group or class [12]. There are 3 data omission mechanisms, including Missing Completely at Random (MCAR), i.e., if the distribution of missing data on an attribute does not depend on the observed data or missing data. This method will use a complete dataset and then generate missing data randomly based on certain proportions. The advantage of this method is that it makes it easier for researchers to make computational estimates from the proposed model. Another mechanism is "Missing at Random" (MAR), i.e., if the distribution of missing data on an attribute depends on the observed data but does not depend on the missing data. The last one is Not Missing at Random (NMAR), if the distribution of missing data on an attribute depends on the missing data [27].

3. Research Methodology

The stages of the research methodology are: (1) processing the dataset and deleting data from the dataset with a predetermined ratio; (2) dataset imputation using the k-NNI method; (3) Studying the training data using the decision tree method; and (4) Testing the resulting model with the test set that has been prepared. The research methodology for completing the classification of the Car Evaluation dataset with missing values is shown in Figure 1.

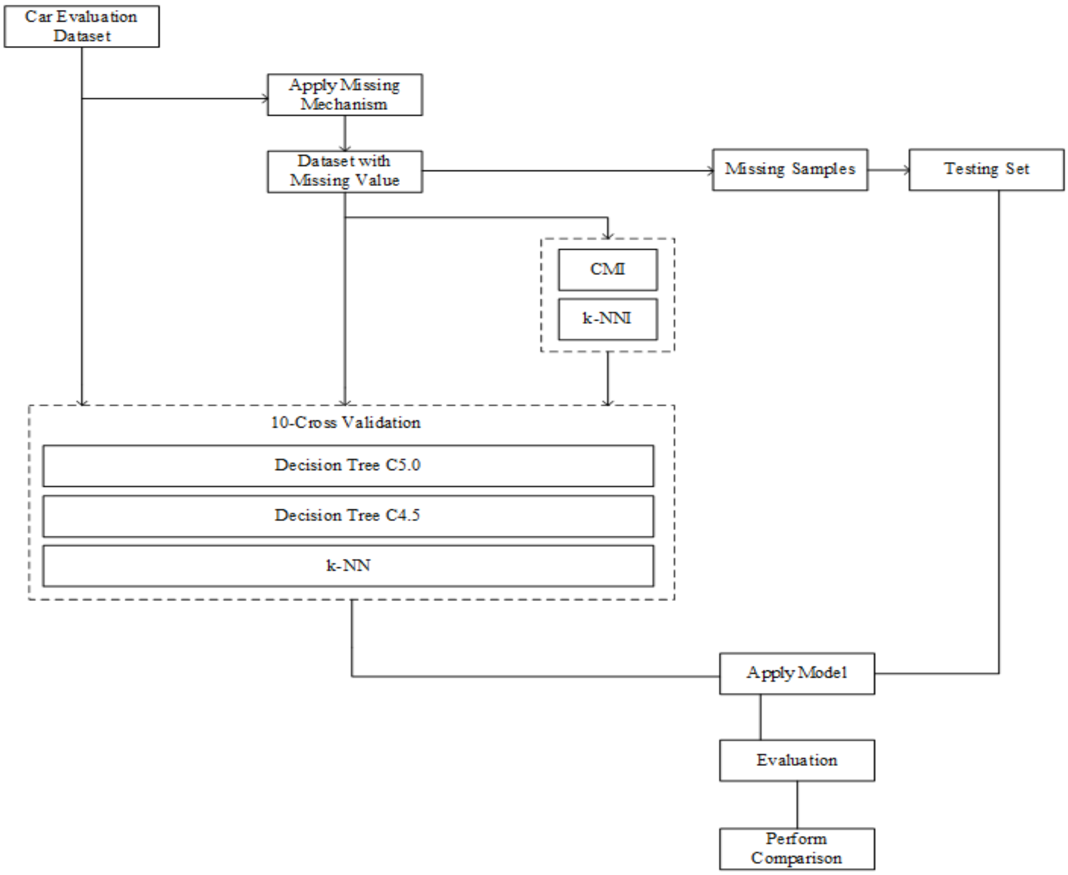

Figure 1: Research Methodology Design

3.1 Training Dataset

The dataset used for testing the algorithm uses the Car Evaluation dataset from the UCI Machine Learning Repository. The Car Evaluation dataset has a multiclass type and categorical attribute characteristic values. Therefore, it is good to use for hypothesis testing, as shown in Table 1.

Table 1: Characteristics of Research Dataset

No	Dataset	Total Instances	Total Attribute	Total Class	Type Attribute
1	Car Evaluation	1728	6	3	Categorical

The training set was obtained by taking 1728 data samples from the Car Evaluation dataset. The Car Evaluation Dataset contains information on several car evaluation parameters with six characteristic attributes, including Purchase Price, Maintenance Price, Number of Doors, Person Capacity, Boot Baggage Size and Safety Estimate. In addition, the dataset produces four types of classes, namely Unacceptable, Acceptable, Good and Very Good. The class distribution of the research dataset can be seen in Table 2.

Table 2: Class Distribution of Research Dataset

No	Attributes	Total Samples	Total Samples (%)
1	Unacc	1210	70.02
2	Acc	384	22.22
3	Good	69	3.99
4	Very good	65	3.76

3.2 Missing Completely at Random Mechanism

Missing Completely at Random (MCAR) is a missing value mechanism where data loss occurs randomly. MCAR is most often encountered in actual cases and significantly affects the performance of the classification results. Therefore it is used in this study. The following is the MCAR mechanism:

Algorithm: Initialize MCAR
Input: ‘data’ as data input, ‘mp’ as percentage of missing value
Output: ‘data’ with missing value

```
BEGIN
   Set x as instance numbers in data, set y as attribute number in data
   Set counter = 0, mv = x* y * mp
   While counter < mv
      Data [random (0, x), random (0, y)] = null
END
```

3.3 k-NN Imputation

k-NN Imputation (k-NNI) is an imputation method based on the k-NN classification algorithm to impute the missing value based on several values that are close to the missing value. k-NNI can handle missing values by determining the nearest neighbor symbolized by k, then calculating the smallest distance from each neighbor that does not have a missing value. The distance between the missing value and its neighbors can be calculated using the Euclidean distance formula. The steps for entering missing values with the k-NNI method are as follows:

o Determine the value of k.

o Find data with missing values in the data set.
o Calculate the nan-Euclidean distance from the initial observation data and other observation data, using the formula:

$$d(x_a, x_b) = \sqrt{w_{ab} * \sum_{j=1}^{m}(x_{aj} - x_{bj})^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

o Choose k observational data with the smallest value
o Select data on attributes related to missing values from selected observation data.
o Fill in the missing values with approximate values from the selected data.

3.5 Decision Tree C5.0

A decision tree is a classification method that studies data in a tree-shaped pattern to produce decisions. C5.0 is the algorithm used in the decision tree method to classify and develop the previous algorithms, namely C4.5 and ID3. The C5.0 algorithm makes a decision tree model pattern based on the entropy value and information acquisition.
Entropy (S) is a value that expresses the uncertainty or impurity in a random data set from a data set expressed in bits. The entropy value is needed to calculate the information gain. Information gain is a measure of the effectiveness of an attribute in classifying data. Information gain is used to determine the order of attributes used to form a classification model pattern.
The following is the formula used to get the entropy and information gain values:

$$Entropy\ (S) = -\sum_{i=1}^{n} P_i\ (P_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

$$Gain(S,T) = Entropy(S) - \sum_{j=1}^{n}\left(p_j\ x\ Entropy(p_j)\right) \quad \ldots\ldots\ldots\ldots \quad (3)$$

The decision concept of the C5.0 decision tree method is as follows:
• First, calculate the total entropy value of the dataset using equation 2.
• Calculate the entropy value and information gain for each attribute criterion using equations 2 and 3.
• Finally, determine the root node based on the largest information gain value using equation 3.
Define an internal node to generate a leaf node based on the entropy value and information gain. The process stops when all attributes have been used. Formation of rules based on the formed classification model and pattern

3.6 Validation

In this study, 10-fold cross-validation was used for the validation process. Cross-validation (CV) is an analytical method that can be used to evaluate the performance of a classifier, where the dataset is divided into two subsets, namely learning data and test data. The selection of the type of CV is based on the size of the dataset. The way k-fold cross-validation works is by dividing the dataset into two groups, namely training data and test data, then carrying out the testing process with k repetitions. The test results are then averaged to produce an accuracy value. A value of k that is too small can cause the accuracy value to be low. This is because the value of k is small, so it is easily affected by noise. A value of k that is too large will result in an ineffective process. This is because a large value of k takes a long time to test. The 10-fold cross-validation method has become the standard method for learning and testing data [28]. The following is an illustrative example of 10-fold cross-validation.

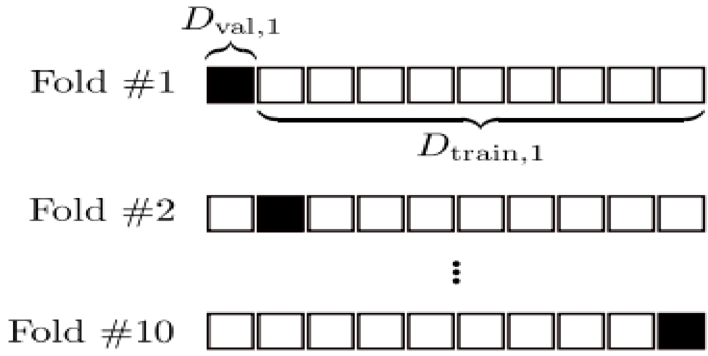

Figure 2: Illustration of 10-cross validation method

3.7 Testing Dataset

The model generated from the validation process is undergoing testing using the test set that has been prepared. The test set is obtained by taking 100 data samples from the complete dataset before any loss of data occurs. We selected 100 samples from the complete dataset based on the samples having missing values in the missing dataset. Files provided for testing are predicted with one of the predefined labels: Unacceptable, Acceptable, Good, and Very Good using the C5.0 Decision Tree classifier. There are four test scenarios:

1. Performance on C5.0, C4.5, and k-NN at 0% missing rate
2. Performance on C5.0, C4.5, and k-NN at 5-50% missing rate
3. Performance at C5.0 + k-NNI, C4.5 + k-NNI, and k-NN + k-NNI classifier at 5-50% missing rate
4. Performance on C5.0 + CMI, C4.5 + CMI, and k-NN + CMI classifier at 5-50% missing rate.

4. Results and Discussion

In this study, the Rstudio tool is used to implement the missing mechanism and imputation process, and the RapidMiner tool is used to implement the classification algorithm. The VIM package on RStudio was developed to explore and analyze the structure of missing values in a dataset, relate missing values to several imputation methods, and verify the imputation process using visualization tools [29]. In this study, we use Rstudio to generate the MCAR mechanism in the dataset and implement the k-NNI algorithm using the k-NNI method provided by the VIM package. The results of the analysis of the performance of the C5.0 and k-NN algorithms on the Car Evaluation dataset processed in the validation process with or without the imputation method are shown in Figure 5.

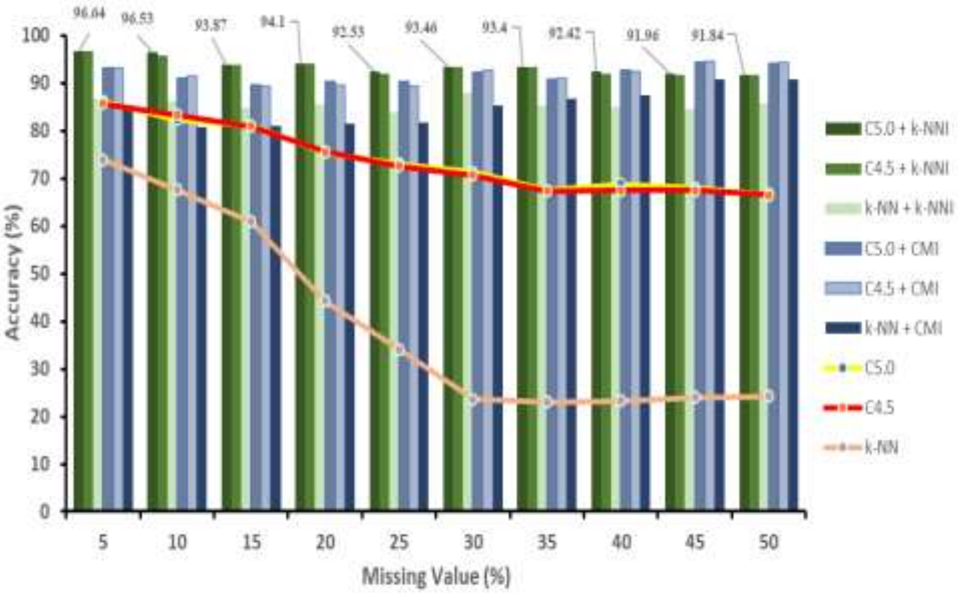

Figure 5: Algorithm performance result on validation with and without imputation method

Experiments show varying results for different imputation and classification methods. We found that C5.0+k-NNI gave better prediction accuracy than other test methods in the validation and testing process, as shown in Figures 5 and 6. C5.0+k-NNI resulted in an average prediction accuracy of over 94% in the dataset under validation and a mean prediction accuracy of 95% in the test for the missing 5-35%. However, when the data set has 40% or more missing values, C5.0+CMI and C4.5+CMI perform better than C5.0+k-NNI. The resulting accuracy for C5.0+CMI and C4.5 CMI is around 92.42-94.5%, while C5.0+k-NNI only produces a prediction accuracy of around 91.84-92.42% when 40-50% of the dataset is missing, probably due to more flaws in the data set. There are fewer neighbors for the k-NNI method to generate the imputed value, and the CMI method can produce the imputed value better because it is based on the distribution of known values of the same class, thus having a sample.

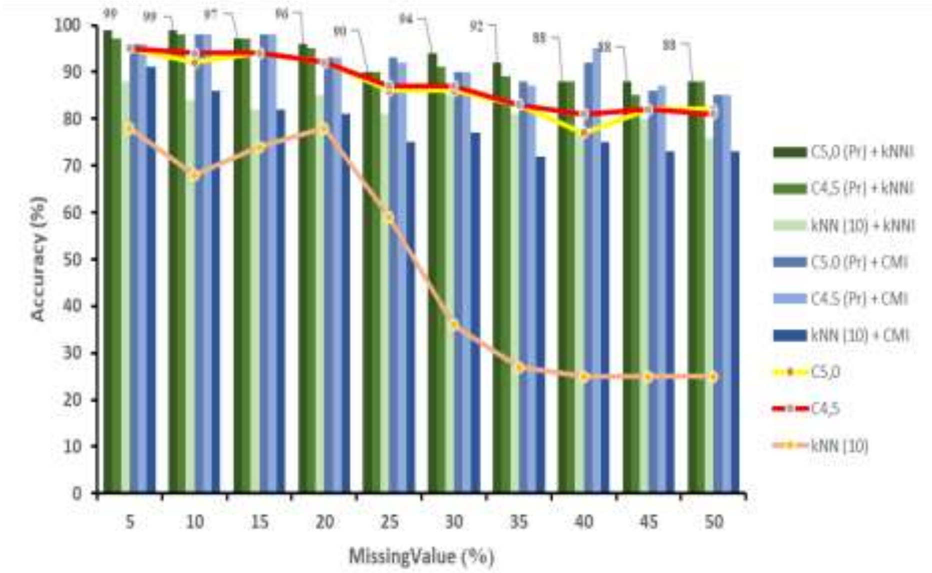

Figure 6: Algorithm performance result on testing with and without imputation method.

Based on the results in Figures 5 and 6, the C4.5 model without k-NNI and the CMI model produce slightly better accuracy than the C5.0 and k-NN models in the test. Therefore, the C5.0 model with k-NNI produces higher accuracy than competing models in both validation and testing. A decision tree is the most suitable type of dataset for car evaluation databases than k-NN, Random Forest, Naive Bayes, and Rule Induction, whose decision trees have an accuracy of 91.1%. The means imputation group performs better on the training data, and the drop test set accuracy supports the imputed means, which relate to the new data much better. Therefore, C5.0 with k-NNI provides a better correct classification rate than the other tested classification algorithms, although C4.5 still slightly outperforms C5.0 and k-NNI before k-NNI was implemented.

Changes in the accuracy of each classification and imputation method form a straight-line equation with an R2 of 0.89 to 0.95. This shows that the greater the missing value, the greater the accuracy will be. To determine the effect of missing values on the processing time of an algorithm, using an Intel Core i7 CPU running at 2.8 GHz and 16 GB of RAM, the average data processing time for each missing value was obtained as shown in Figure 7.

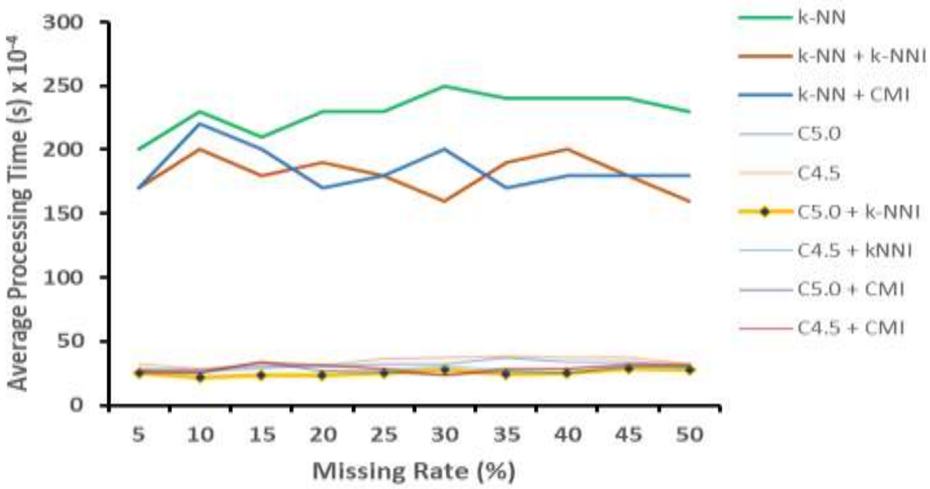

Figure 7: Average Computational Time with and without Imputation Method.

Figure 7 shows that, in terms of processing time, the C5.0+k-NNI offers fast processing compared with other methods. k-NNI and CMI can increase the processing time of C5.0 and C4.5. Those imputation methods can also increase the processing time of k-NN classifiers. However, the k-NN classifier's processing time is still the slowest compared to other methods for both complete and incomplete data.

The limitation in this study is that the mechanism of data loss is MCAR (missing completely at random), using the missing value with a ratio of 0-50%, and using the Rstudio and RapidMiner data science tools. It is advisable to carry out further research using data loss mechanisms and other data science tools.

5. Conclusions

The overall result of testing the algorithm on the Car Evaluation dataset at a 0–50% missing rate is that C5.0 with k-NNI provides better prediction accuracy than the other tested classification algorithms, even though C4.5 still slightly outperforms C5.0 and k-NNI before k-NNI is applied. In terms of processing time, the C5.0 with k-NNI has better performance than the C4.5 and k-NN algorithms, both with and without the imputation method.

Acknowledgement

The authors would like to thank the Department of Electrical Engineering at Brawijaya University for their assistance with this research.

References

[1] J. Awwalu, A. Ghazvini, and A. Abu Bakar, "Performance Comparison of Data Mining Algorithms: A Case Study on Car Evaluation Dataset," *International Journal of Computer Trends and Technology*, vol. 13, no. 2, 2014, [Online]. Available: http://www.ijcttjournal.org

[2] S. T. Luo, C. T. Su, and W. C. Lee, "Constructing intelligent model for acceptability evaluation of a product," *Expert Systems with Applications*, vol. 38, no. 11, pp. 13702–13710, Oct. 2011, doi: 10.1016/j.eswa.2011.04.162.

[3] T. K. Mustafa and M. S. Abd, "Proposed Approach for Analysing General Hygiene Information Using Various Data Mining Algorithms," *Iraqi Journal of Science*, vol. 58, no. 1B, pp. 337–344, Jan. 2022.

[4] Odule, T. J. Adesina, A. O. Abdullah, K. K. Adebisi, Ogunyinka, and I. Peter, "Using affiliation rules-based data mining technique in referral system," *Iraqi Journal of Science*, vol. 61, no. 11, pp. 3095–3103, 2020, doi: 10.24996/ijs.2020.61.11.30.

[5] M. A. Ali and T. K. Mustafa, "An application of data mining algorithms for analyzing psychological researches," *Iraqi Journal of Science*, vol. 62, no. 10, pp. 3705–3718, Oct. 2021, doi: 10.24996/ijs.2021.62.10.29.

[6] B. H. Hadi and T. K. Mustafa, "Inspecting Hybrid Data Mining Approaches in Decision Support Systems for Humanities Texts Criticism," *Iraqi Journal of Science*, vol. 62, no. 11, pp. 4101–4109, Nov. 2021, doi: 10.24996/ijs.2021.62.11.30.

[7] P. Jain and S. Kr Vishwakarma, "A Case Study on Car Evaluation and Prediction: Comparative Analysis using Data Mining Models," *International Journal of Computer Applications*, Vol. 172, No. 9, pp. 21-25, August 2017.

[8] A. Baykal, "PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHMS OF SEVERAL DATA MINING SOFTWARES," *Middle East Journal of Science*, vol. 4, no. 2, pp. 104–112, Dec. 2018, doi: 10.23884/mejs.2018.4.2.06.

[9] O. M. Al-Mubayyed, B. S. Abu-Nasser, and S. S. Abu-Naser, "Predicting Overall Car Performance Using Artificial Neural Network," 2019. [Online]. Available: http://www.ijeais.org/ijaar

[10] R. R, S. B, and V. S. A. Sofia, "Data Mining Issues and Challenges: A Review," *IJARCCE*, vol. 7, no. 11, pp. 118–121, Nov. 2018, doi: 10.17148/ijarcce.2018.71125.

[11] S. Gavankar and S. Sawarkar, "Decision tree: Review of techniques for missing values at training, testing and compatibility," in *Proceedings - AIMS 2015, 3rd International Conference on Artificial Intelligence, Modelling and Simulation*, Oct. 2016, pp. 122–126. doi: 10.1109/AIMS.2015.29.

[12] F. U. F. Khan, K. U. Z. Khan, and S. K. Singh, "Is Group Means Imputation Any Better Than Mean Imputation: A Study Using C5.0 Classifier," in *Journal of Physics: Conference Series*, Jul. 2018, vol. 1060, no. 1. doi: 10.1088/1742-6596/1060/1/012014.

[13] Q. Song, M. Shepperd, X. Chen, and J. Liu, "Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation," *Journal of Systems and Software*, vol. 81, no. 12, pp. 2361–2370, Dec. 2008, doi: 10.1016/j.jss.2008.05.008.

[14] M. Rajan and V. Gimpy, "Missing Value Imputation in Multi Attribute Data Set." [Online]. Available: www.ijcsit.com

[15] U. R. Yelipe, S. Porika, and M. Golla, "An efficient approach for imputation and classification of medical data values using class-based clustering of medical records," *Computers and Electrical Engineering*, vol. 66, pp. 487–504, Feb. 2018, doi: 10.1016/j.compeleceng.2017.11.030.

[16] Y. Ariyanto, B. Harijanto, and A. N. Asri, “Analyzing Student’s Learning Interests in the Implementation of Blended Learning Using Data Mining,” *International journal of online and biomedical engineering*, vol. 16, no. 11, pp. 153–160, 2020, doi: 10.3991/ijoe.v16i11.16453.
[17] F. Siraj and M. Ali, “Mining Enrollment Data Using Descriptive and Predictive Approaches,” in *Knowledge-Oriented Applications in Data Mining, InTech*, 2011. doi: 10.5772/14210.
[18] A. P. Wibawa et al., “Naïve Bayes Classifier for Journal Quartile Classification,” *International Journal of Recent Contributions from Engineering, Science & IT* (iJES), vol. 7, no. 2, p. 91, Jun. 2019, doi: 10.3991/ijes.v7i2.10659.
[19] C. Shah and A. G. Jivani, "Comparison of data mining classification algorithms for breast cancer prediction," 2013 *Fourth International Conference on Computing, Communications and Networking Technologies* (ICCCNT), Tiruchengode, India, 2013, pp. 1-4, doi: 10.1109/ICCCNT.2013.6726477.
[20] A. Elen and E. Avuclu, “A comparison of classification methods for diagnosis of Parkinson’s,” *International Journal of Intelligent Systems and Applications in Engineering*, vol. 8, no. 4, pp. 164–170, 2020, doi: 10.18201/ijisae.2020466309.
[21] Y. Muliono and F. Tanzil, “A Comparison of Text Classification Methods k-NN, Naïve Bayes, and Support Vector Machine for News Classification,” *Jl. Kh. Syahdan*, vol. 03, no. 02, 2018.
[22] S. Kumar and H. Sharma, “A Survey on Decision Tree Algorithms of Classification in Data Mining,” 2016. [Online]. Available: www.ijsr.net
[23] P. Patidar and A. Tiwari, “Handling Missing Value in Decision Tree Algorithm,” 2013.
[24] H. Kang, “The prevention and handling of the missing data,” *Korean Journal of Anesthesiology*, vol. 64, no. 5. pp. 402–406, May 2013. doi: 10.4097/kjae.2013.64.5.402.
[25] Z. Zhang, “Missing data imputation: Focusing on single imputation,” *Annals of Translational Medicine*, vol. 4, no. 1, Jan. 2016, doi: 10.3978/j.issn.2305-5839.2015.12.38.
[26] L. Muflikhah, N. Hidayat, and D. J. Hariyanto, “Prediction of hypertension drug therapy response using K-NN imputation and SVM algorithm,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 1, pp. 460–467, Jul. 2019, doi: 10.11591/ijeecs.v15.i1.pp460-467.
[27] Y. Dong and C.-Y. J. Peng, “Principled missing data methods for researchers,” *SpringerPlus*, vol. 2, No. 1, pp. 222-230, 14 May. 2013, doi:10.1186/2193-1801-2-222 . [Online]. Available: http://www.springerplus.com/content/2/1/222
[28] D. Berrar, “Cross-validation,” in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1–3, Elsevier, 2018, pp. 542–545. doi: 10.1016/B978-0-12-809633-8.20349-X.
[29] A. Kowarik and M. Templ, “Imputation with the R package VIM,” *Journal of Statistical Software*, vol. 74, no. 7, pp. 1–16, Oct. 2016. doi: 10.18637/jss.v074.i07.