Mantik Journal, 8 (1) (2024), ISSN 2685-4236 (Online)

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik

Comparison of the nearest neighbor algorithm and C4.5 for the retrieval on case-bassed reasoning process (case study: children respiratory disorders)

Tursina[1], Rina Septiriana[2]
[1,2]Department of Informatics/Faculty of Engineering/Study Program Informatics,
Tanjungpura University, Indonesia

A R T I C L E I N F O

Article history:

Received Mar 17, 2024
Revised Mar 21, 2024
Accepted Mar 28, 2024

Keywords:

Cased Based Reasoning;
C4.5;
Nearest Neighbor;
Retrieval.

ABSTRACT

The diagnosis of respiratory problems was usually made through direct consultation with a pediatric respiratory specialist or by studying several previous respiratory disorders cases. These cases were gleaned from prior experiences or the knowledge of subject-matter experts. Case-Based Reasoning (CBR) is the processing of diagnosing a patient based on past cases or expertise. Retrieve, reuse, revise, and retain are some of the steps of case-based reasoning. The retrieval stage of CBR was where the classification method searches for similarity values. Numerous algorithms exist for classification techniques, such as C4.5 and the Nearest Neighbour algorithm. This study compares the similarities between the C4.5 and Nearest Neighbor algorithms. The Nearest Neighbour approach was used to search for similarity, and the results show that 99.33% of the items classified based on learning data were nearest to the object. By contrast, the accuracy value for the C4.5 approach was 100%.

Corresponding Author:

Rina Septiriana,
Department of Informatics, Faculty of Engineering, Study Program Informatics,
Tanjungpura University,
Jl. Prof. Dr. Hadari Nawawi, Kota Pontianak, 78124, Indonesia
Email: rinaseptiriana@informatika.untan.ac.id

1. INTRODUCTION

Diseases associated with respiratory tract abnormalities are among the most common in children. It is important to keep an eye out for respiratory tract issues in children since they can affect the child's growth and development in addition to making the child uncomfortable. The frequency and pattern of breathing are two indicators of respiratory tract diseases (Ngastiyah, 2004).

Typically, diagnosing respiratory illnesses involves speaking with a pediatric respiratory specialist directly or researching a number of prior respiratory disease-related instances. These situations can be discovered through personal experience or through the experience of a specialist in the relevant sector (Panggabean, 2022). Finding the most similar examples helps with solving new case problems. These situations can be discovered through personal experience or through the experience of a specialist in the

relevant sector. Finding the most comparable situations and then adapting them to the supplied case (new case) is how new case problems are solved (Utomo & Nasution, 2016).
In order to assess whether there are any instances that are comparable to the new case or cases already in the case base, CBR compares new cases to cases already in the case base. The most comparable instances will be found and used as answers to new case problems as a consequence of similar calculations made throughout the retrieval process. The system will save the case (retain it) in the knowledge case base if no related cases are discovered (Lamy et al., 2019).

The case retrieval process in a CBR system consists of determining the similarities between the new problem and those stored in the case base and then selecting the closest-matching case to be used as a solution to the problem(Homem et al., 2020)]. It is a significant stage for CBR because a retrieval process determines the similarity of the cases. Therefore, help CBR find a solution or save a new one for that case. The process of classifying a text or document by applying a similarity measure and an accurate classifier is known as classification (Amer & Abdalla, 2020). Classification method can be used to find similarity values at the retrieval step. Nearest Neighbor algorithm and C 4.5 are two of the many algorithms used in the classification approach.

K-NN algorithm, also referred to as the closest neighbor algorithm, is a machine learning algorithm that can be used to resolve classification and regression issues. The category of supervised learning includes this algorithm. K-Nearest Neighbor approach was used (Fatoni & Noviandha, 2018) on diphtheria cases, and the similarity rate was 95.17%. The C4.5 method produces decision trees that are easy to comprehend, have an acceptable level of accuracy, are efficient in dealing with discrete type attributes, and can deal with both discrete and numeric type attributes (Widyastuti et al., 2019). When KNN, C4.5, and Neural Network were evaluated in a different study (Astuti, 2016) for the motor vehicle suitability process, it was found that C4.5 had the best accuracy value (92.8%) when compared to KNN and Neural Network.

Based on this context, a comparison study will be conducted to determine the optimal classification algorithm for diagnosing respiratory tract problems in children by comparing similarity findings in the CBR retrieval process between the Nearest Neighbor algorithm and C 4.5. This research aims to determine which one provides the most accurate similarity findings for the CBR retrieval process.

2. RESEARCH METHOD

2.1 Research Stage

A classification approach from data mining is used in the research stages of the algorithm comparison process for the retrieval step of Cased Based Reasoning. As a result, the stages completed comprise preparation data, model development and validation processes, as well as evaluation, the outcomes of which will be compared later. Figure 1 depicts all of these steps.

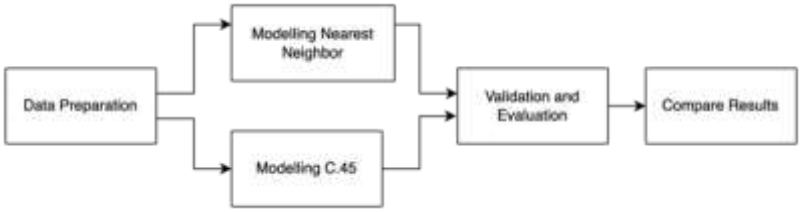

Figure 1 Research Stage of Comparison Between KNN and C4.5

2.2 Preparation Data

The conversion of raw data from various sources into standard formats is an essential step in the data analysis process. The following methods were followed (Brownlee, 2020): a) Data collection: Information on children's respiratory diseases is

requested. This data set includes symptom information, disease terminology, and medical records assembled into a group of cases; b) Validation and cleanup Data is standardized in stages by removing duplicates and errors, filling in blanks where possible, and placing it in a consistent format; c) Transforming and Enriching Data: Existing data is converted to xls format, and each row is manually evaluated to determine whether any information is missing. However, the data is acceptable and thorough enough to be handled immediately after the preceding processed data cleansing and validation.

2.3 Modelling Nearest Neighbor and C4.5

Modeling is creating a machine learning model that can recognize specific patterns in data to make predictions. In this case, the model built using the trained data will be viewed for its predictions using the similarity approach, namely the nearest neighbor algorithm and the C4.5 algorithm.

a. Nearest Neighbor

The Nearest Neighbor method uses weighted comparisons of numerous existing attributes to determine how close new and old cases are to one another. It is intended to use solutions from past patients to identify a remedy for a new patient. To determine which patient cases to employ, the distance between new patient cases and all existing patient cases is determined. The best closeness answer from geriatric patient instances would be taken and used to new patient cases. Figure 1 provides a representation of case proximity.

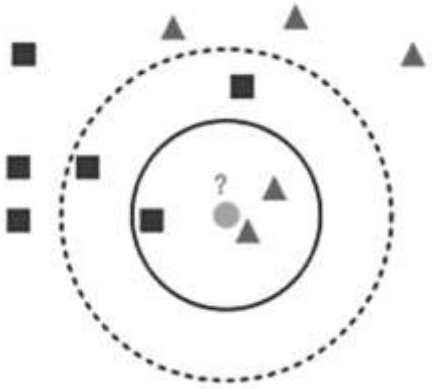

Figure 2 Illustration of Adjacency Between Cases](Ramadhani et al., 2017)

The following equation can be used to determine how similar two examples are (Lubis et al., 2020):

$$similarity(T,S) = \frac{\sum_{i=1}^{n} f(T_i,S_i)x}{w_i} \quad (1)$$

Description:

T: new case
S: cases in storage
n : the total number of features in each case
i : unique property ranging from 1 to n
f : Case T and Case S have similarities in the function of attribute i.
w : weight given to attribute

Based on the formula (1), the ability of the closest neighbor algorithm to identify similarities is assessed using the case data that already exists without engaging in a training phase. By testing the results using the available data and then obtaining the accuracy and RMSE values as the closest neighbor algorithm performance scale, the

results of the highest rating for the similarity value, which is believed to be the greatest are reviewed.

Nearest Neighbors is a supervised learning strategy for classification primarily relies on the closest k neighbors. To determine the k-NN, it employs Euclidean distance-based metrics to figure out the shortest distance between the query instance and the training samples (Tchomté et al., 2020). The nearest Neighbor pseudocode is depicted in Figure 3 (Goel & Thareja, 2017).

1: The data is loaded.
2: The value *k* is initialized.
3: For every point in the training data
4: DistanceVector <- Calculate the distance between testing data and each row of training data. Here, Euclidean distance or other metrics such as Chebyshev or cosine can be used.
5: vectorSorted <- Sorting DistanceVector in ascending order based on distance values;
6: topK <- Get top k rows from vectorSorted;
7: frequentClass <- Get the most frequent class of these rows;
8: Return frequentClass

Figure 3 Pseudocode of Nearest Neighbor

b. C4.5

Classification version 4.5, often known as C4.5 and a development of ID3, is one of the tree classification methods that is frequently employed. According to numerous studies, the C4.5 algorithm is simple to understand and has a respectably high level of accuracy. In addition, C4.5 resolves discrete and numerical properties well (Han et al., 2022). The C4.5 method loads all training data samples into memory before reading them from storage and using them to build the tree.

After generating the training data, the first step is to choose properties that can be calculated using the entropy principle. A collection of things' impurity is expressed as entropy. The formula for calculating entropy is as follows (Sembiring et al., 2018).

$$\textbf{Entropy}\ (\mathbf{S}) = \sum_{\mathbf{i=1}}^{\mathbf{n}} - \mathbf{p_i}\ \mathbf{log_2}\ \mathbf{p_i} \quad (2)$$

Description:

S = Set of Cases

n = Number of S Partitions

pi = Probability is computed by dividing the total number of cases by the number of classes.

The C4.5 algorithm uses Information Gain to pick the attributes after calculating the entropy value. The gain can be determined by applying the following formula:

$$\textbf{Gain}\ (\mathbf{S}, \mathbf{A}) = \textbf{Entropy}\ (\mathbf{S}) - \sum_{\mathbf{i=1}}^{\mathbf{n}} * \textbf{Entropy}\ (\mathbf{Si}) \quad (3)$$

Description :

S: Set of cases

A: Attributes

N: Number of attributes

|Si|: Number of i partition

|S|: number of cases in S

C4.5 approach is one approach for converting significant information into a decision tree that represents the rules. The objective of building a decision tree in the C4.5 method is to make existing problems easier to solve. There are phases to transforming the C4.5 algorithm into rules. The generated decision tree will serve as the

classification framework, and the classification outcomes will be assessed and validated to produce accuracy and rmse values (Kurniawan et al., 2020).

The steps for creating a decision tree using the C4.5 algorithm are as follows (Rismayanti, 2017)(Yuliansyah et al., 2021): (a) Data training preparation. (b) Counting roots of the tree. The specified attribute will be used as the root. The first root will be determined by calculating the gain value of each feature. Calculate the entropy value of the part before calculating the gain value. (c) Using the following formula, compute the Gain value, (d) The decision tree results are obtained from calculating a leaf or node, and each leaf node marks the class label.

2.4 Confusion Matrix

A confusion matrix of size n x n linked to a classifier shows the predicted and actual categorization, where n is the number of possible classes. Table 1 displays a confusion matrix, with the following entries (Zeng, 2020): Actual negatives that are accurately identified as negatives is defined as True Negative. Actual positives that are wrongly labelled as negatives is defined as False Negative. Actual positives that are correctly identified as positive is defined as True Positive. Actual negatives that are wrongly labelled as positives is defined as False Positive.

Table 1 Predicted Matrix

	Predicted Negative	Predicted Positive
Actual Negative	True Negative	False Positive
Actual Positive	False Negative	True Positif

This matrix can be used to calculate the prediction accuracy and classification error as follows:
Description:
TP = True Positive
TN = True Negative
NB = Total Negative Predicted
NG = Total Positive Predicted

2.5 RMSE

The root mean square error (RMSE) is the residuals' standard deviation (prediction errors that illustrate the standard deviation of prediction for correct and incorrect test dataset estimates (Shadman Roodposhti et al., 2019). Better accuracy in forecasting continuous survival time is shown by a lower RMSE (Altuhaifa et al., 2023). Formula of RMSE as follows (Zhang et al., 2020):
Description: n = Number of samples Y0 = Actual Value Yp = Predicted Value

The most prevalent difficulty with the usage of this statistic is its susceptibility to outliers. In fact, the normal distribution, which underpins the use of the RMSE, describes the existence of outliers and their chance of occurrence quite well.

2.6 Compare Result

The test results are derived from the consequences of performance calculations for each method utilizing various scenarios, such as using overall data without splitting it into sample and test spaces, then dividing it into 70:30, and ultimately dividing it into 60:40. The accuracy and RMSE findings will compare the outcome of the previously given scenario.

3. RESULTS AND DISCUSSIONS

3.1 Data Preparation

One of the crucial steps in carrying out a similarity matching process using Classification approach is getting the data ready so that the data that will be examined for comparison can yield accurate findings. The outcomes of the data distribution process, which can be seen in Figure 4, were acquired after carrying out the data preparation process in system design.

Figure 4 Distribution Data

Figure 4 shows that there are 120 case data, which are divided into 12 different types of diseases. The distribution is highest for conditions P12, P8, and P9, where there are 15 cases of illness, and lowest for diseases P1, P10, P2, P3, P4, and P5, where there are up to 12 cases. Table 5 contains information about the names of diseases.

Table 2 Respiratory Disorder

Diseases code	Diseases name
P1	Salesma
P2	Faringitis
P3	Influenza
P4	Mild croup syndrome
P5	Moderate croup syndrome
P6	Severe croup syndrome
P7	Acute otitis media
P8	Bronchitis
P9	Lungs tuberculosis
P10	Mild atshma
P11	Moderate atshma
P12	Severe atshma

3.2 Modelling

The results of the data preparation stage are followed by a modelling process using the nearest-neighbor and C4.5 algorithms, which includes an evaluation and validation process for the accuracy of the similarity values using a classification approach.

a. Nearest Neighbor Algorithm

Based on the figure 3, there are several stages that wil be passed by Nearest Neigbor algortithm. In this research, that stages will be done by Rapid Miner tools. RapidMiner is a machine learning, predictive analytics, and business data analysis software developed by the company. It is used in commerce and business applications, in addition to research, education, and training, as well as quick prototyping and application development. It helps with data preparation, visualization of results, validation, and optimization, among other things. RapidMiner is built on an open-source platform, with the RapidMiner Basic Edition available for free under the GNU General Public License (Arunadevi et al., 2018). Design of RapidMiner for Nearest Neighbor Algorithm is illustrated on figure 5.

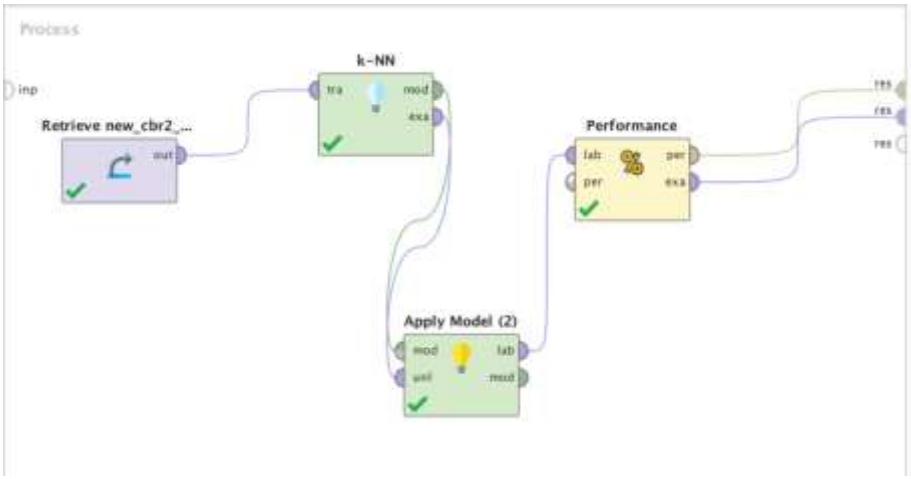

Figure 5 Classification Nearest Neighbor in RapidMiner tool

Figure 5 depicts the flow of received data, which is subsequently categorized using Nearest Neighbor and used to demonstrate the outcomes of evaluation and validation using performance metrics. The procedure depicted in Figure 5 is carried out without the use of a data split or data sharing mechanism. Figure 5 depicts the obtained results in the form of a confusion matrix.

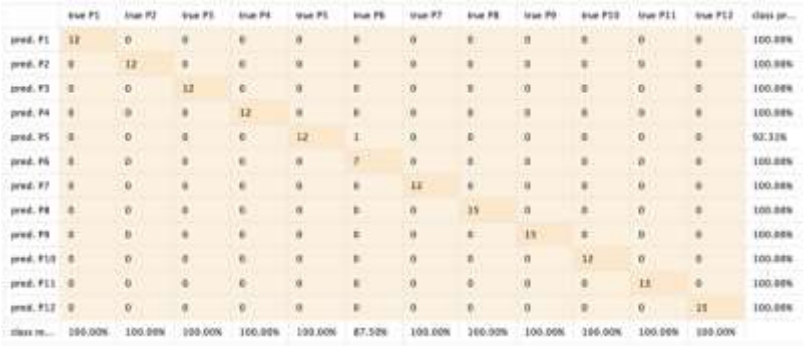

Figure 6 Confusion Matrix Nearest Neighbor without Split Data

According to Figure 5, there is a prediction error in P6 and P5, where there is one instance that is between the two disorders. As a result, P5 has an accuracy value of 92.31% and P6 has an accuracy value of 85.5%. Following that, the Nearest Neighbor algorithm was used in the modelling stage, with a data split of 70:30 and 60:40, where 70% and 60% of the data were used as sample data in classification evaluation, and the remaining 30% and 40% of the data were used as validation to see how the Nearest Neighbor algorithm performed. Result of that process is shown on the figure 7.

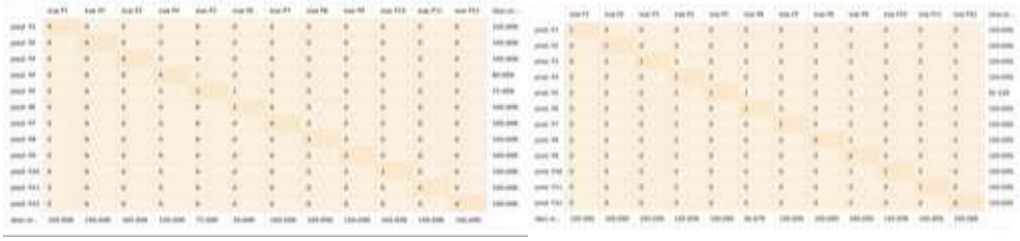

Figure 7 Confusion Matrix of Nearest with Data Split 70:30 (left) and 60:40 (right)

Figure 7 shows Nearest Neighbor's performance results, employing a data split of 60% for sample data and the rest for validation. The validation findings with the lowest accuracy were forecasts for P5, and the correct value for P6 was 66.67%. Figure 7 shows performance results of Nearest Neighbor, employing a data split of 60% for sample data and the rest for validation.

b. C4.5 Algorithm

The C4.5 technique performs the class classification process using a decision tree, allowing the similarity matching process to be performed using a classification approach utilizing a decision tree. The performance testing procedure employs both a split data process and one that does not.

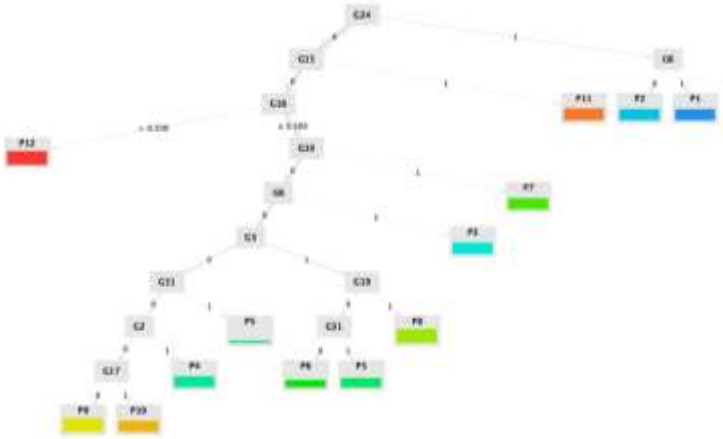

Figure 8 Decision Tree of Respiratory Disorder Data

Figure 8 depicts the decision tree generated by the C4.5 algorithm. According to Figure 8, the selected root is G24 or symptom 24. Based on all available case data, each symptom is used as a branch. Meanwhile, the illness classification serves as a leaf to demonstrate the function of each classification. The decision tree was tested without using split data, and the results are shown in Figure 9.

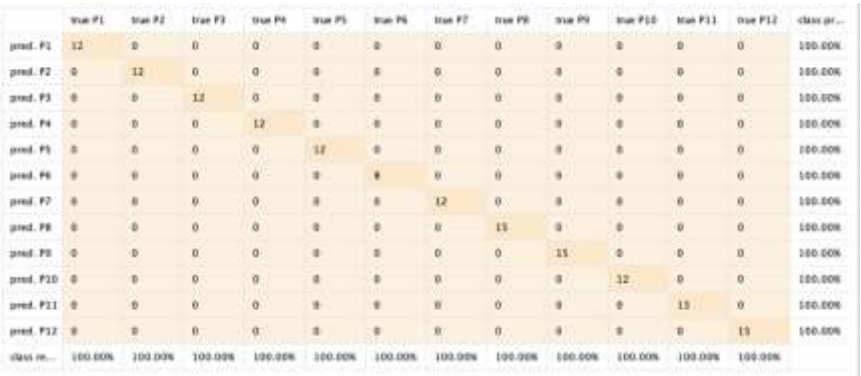

Figure 9 Confusion Matrix of C4.5 Without Split Data

Figure 9 depicts the results of a search utilizing the C4.5 decision tree without previously performing a data split process. Based on the results, it can be seen that the performance is operating extremely well, as seen by the accuracy value of 100%, which means that the similarity value can be searched accurately in all circumstances.

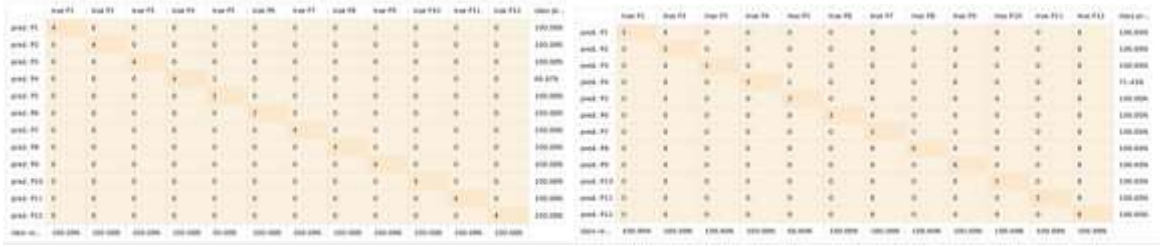

Figure 10 Confusion Matrix of Nearest with Data Split 70:30 (left) and 60:40 (right)

According to Figure 10, utilizing 70% of the data as a sample resulted in a validation of roughly 66.67% for P4 and approximately 50% for P5, which is reasonable. Furthermore, the results of C4.5 using 60% as a data sample yielded a P4 validity of 71.4% and a P5 validity of 60%. There appears to be more than 70:30 data.

3.3 Evaluation and Validation

Based on the previous process, a judgement was made using a confusion matrix, providing a performance vector for Nearest Neighbor shown in Figure 11 and a performance vector for C4.5 shown in Figure 12.

PerformanceVector

PerformanceVector:
accuracy: 99.33%
ConfusionMatrix:

True:	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
P1:	12	0	0	0	0	0	0	0	0	0	0	0
P2:	0	12	0	0	0	0	0	0	0	0	0	0
P3:	0	0	12	0	0	0	0	0	0	0	0	0
P4:	0	0	0	12	0	0	0	0	0	0	0	0
P5:	0	0	0	0	12	1	0	0	0	0	0	0
P6:	0	0	0	0	0	7	0	0	0	0	0	0
P7:	0	0	0	0	0	0	12	0	0	0	0	0
P8:	0	0	0	0	0	0	0	15	0	0	0	0
P9:	0	0	0	0	0	0	0	0	15	0	0	0
P10:	0	0	0	0	0	0	0	0	0	12	0	0
P11:	0	0	0	0	0	0	0	0	0	0	13	0
P12:	0	0	0	0	0	0	0	0	0	0	0	15

root_mean_squared_error: 0.113 +/- 0.000

Figure 11 Performance Vector of Nearest Neighbor

Performance Vector on the accuracy of Nearest Neighbor is less than the accuracy of C4.5. The algorithm of C4.5 results in a perfect score of 100% with 0 RMSE value. Besides that, the C4.5 Algorithm just got a small error on Predict P6 and actual P5, as shown in Figure 11. Figure 11 and Figure 12 are examples of performance vectors of both algorithms without data split. The result of the performance vector with split data at 60:40 and 70:30 is shown in Table 3.

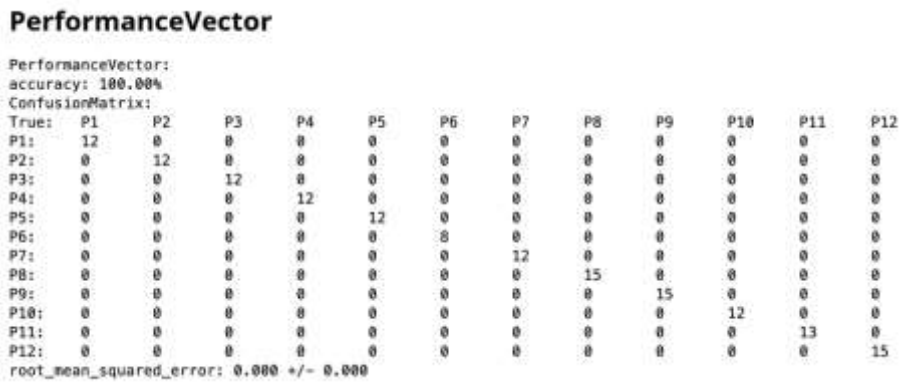

Figure 12 Performance Vector of C4.5

3.4 Comparing Results

Examine the outcomes of Nearest Neighbour and C4.5. The final stages were completed in order to obtain an accurate value. When evaluating and comparing different categorization models or machine learning techniques, comparing algorithms is extremely useful. The Confusion Matrix is the foundation of many metrics because it contains all of the information about the algorithm and classification rule performance (Grandini et al., 2020).

Table 3 Scenario of Validation and Testing

	WITHOUT DATA SPLIT		DATA SPLIT (70:30)		DATA SPLIT (60.40)	
	ACCURATION(%)	RMSE	ACCURATION(%)	RMSE	ACCURATION(%)	RMSE
NEAREST NEIGHBOR	99,33	0,113	95,65	0,341	98,36	0,288
C4.5	100	0	95,65	0,209	96,72	0,181

According to Table 3, model C4.5 has the best performance, with the highest accuracies without using split data of 100% and the slightest error of 0%. The nearest Neighbor less than C4.5 without split data condition is 99,33%, with a modest error value of 0,113 for the experiment with the C4.5 approach.

On the other hand, Nearest Neighbor and C4.5 were used by Fatoni and Noviandha (2018) and Astuti (2016). In their Fatoni & Noviandha (2018) research, the Nearest Neighbor on diphtheria cases had a similarity rate of 95.17%. And then, on Astuti (2016), The C4.5 had the best accuracy value at (92.8%) (Astuti, 2016; Fatoni & Noviandha, 2018). However, when the result was compared with this, it showed a better result for Nearest Neighbor and C4.5. Althought C4.5 gives the best result out of them.

Furthermore, based on the study's results, other analyses are being conducted to compare the impact of the study using the Nearest Neighbor and C4.5. The following are the analysis results: a) The Nearest Neighbor Algorithm with C4.5 has the highest value on conditions without data split, where the data splitting can be classified as a data learning process. As a result, validating data samples or using training does not appear to be necessary in this case, making both algorithms more suitable for use as a method for retrieval in case-based reasoning that does not require previous training. b) C4.5 produces outstanding results for instances involving respiratory disorder, with a score of over 100% and an error score of zero, making it more acceptable for use in Case-Based Reasoning retrieval procedures. c) The 70:30 split data setup produced the most significant performance results for Nearest Neighbor and C4.5. An identical score of 95,65% was attained. However, the error rate for the Nearest Neighbor was more significant, at 0,132. d) On Nearest Neighbor, the accuracy value for the 60:40 data split condition is 98.36%, but for C4.5, it is 96,72%. As shown, Nearest Neighbor, when

trained, gets an average score higher than C4.5. However, the model utilizing C4.5 has a reduced average error value. Therefore, both approaches are pretty good to use, although C4.5 has a lower potential to make errors than Nearest Neighbor.

4. CONCLUSION

Both the Nearest Neighbour and C4.5 algorithms produce high classification performance values, making them suitable for use in the retrieval method of Case-Based Reasoning. Subsequently, the outcomes of contrasting the Nearest Neighbour and C4.5 approaches revealed that, in contrast to the Nearest Neighbour technique, C4.5 had the lowest average error value and the highest accuracy value—100%. This means that the C4.5 technique is better than Nearest Neighbour when it comes to use in the retrieval phase of Case-Based Reasoning.

Additionally, there are some limitations in this research, especially on the data that was used, which was tabular data that has been completed, making it easier to do a retrieval process and search for similarities between them. So, in future work, the data that was used can be formatted variously. Moreover, another algorithm can be combined to get better results for the Nearest Neighbor Algorithm.

ACKNOWLEDGEMENTS

This study and the research behind it would not have been feasible without the extraordinary assistance of Teknik Faculty Universitas Tanjungpura and our colleagues at Informatika. During Fiscal Year 2023, DIPA research activities at Teknik Faculty Universitas Tanjungpura aided this study.

REFERENCES

Altuhaifa, F. A., Win, K. T., & Su, G. (2023). Predicting lung cancer survival based on clinical data using machine learning: A review. *Computers in Biology and Medicine*, *165*, 107338. https://doi.org/10.1016/j.compbiomed.2023.107338

Amer, A. A., & Abdalla, H. I. (2020). A set theory based similarity measure for text clustering and classification. *Journal of Big Data*, *7*(1), 74. https://doi.org/10.1186/s40537-020-00344-3

Arunadevi, J., Ramya, S., & Raja, M. R. (2018). *A study of classification algorithms using Rapidminer*. https://www.researchgate.net/publication/325718529

Astuti, P. (2016). Komparasi Algoritma C 4.5 , KNN, dan Neural Network dalam proses kelayakan pnenerimaan kridit Kendaraan Bermotor. *Faktor Exacta*.

Brownlee, J. (2020). *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery.

Fatoni, C. S., & Noviandha, F. D. (2018). Case Based Reasoning Diagnosis Penyakit Difteri dengan Algoritma K-Nearest Neighbor. *Creative Information Technology Journal*, *4*(3), 220. https://doi.org/10.24076/citec.2017v4i3.112

Goel, P., & Thareja, R. (2017). Analysis of Various Data Mining Techniques using Novel Ratings Prediction. *IARS International Research Journal*, *7*(2). https://doi.org/10.51611/iars.irj.v7i2.2017.75

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*.

Han, J., Pei, J., & Tong, H. (2022). *Data mining: concepts and techniques* (4th ed.). Morgan Kaufmann.

Homem, T. P. D., Santos, P. E., Reali Costa, A. H., da Costa Bianchi, R. A., & Lopez de Mantaras, R. (2020). Qualitative case-based reasoning and learning. *Artificial Intelligence*, *283*, 103258. https://doi.org/10.1016/j.artint.2020.103258

Kurniawan, D., Anggrawan, A., & Hairani, H. (2020). Graduation Prediction System On Students Using C4.5 Algorithm. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *19*(2), 358–365. https://doi.org/10.30812/matrik.v19i2.685

Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J., & Séroussi, B. (2019). Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine*, *94*, 42–53. https://doi.org/10.1016/j.artmed.2019.01.001

Lubis, A. R., Lubis, M., & Khowarizmi, A.-. (2020). Optimization of distance formula in K-Nearest Neighbor method. *Bulletin of Electrical Engineering and Informatics*, *9*(1), 326–338. https://doi.org/10.11591/eei.v9i1.1464

Ngastiyah. (2004). *Asuhan Keperawatan Penyakit Dalam Edisi I*. EGC.

Panggabean, D. M. (2022). Peran dan Fungsi Keluarga Dalam Perawatan Paliatif.

Ramadhani, R., Helilintar, R., & Rochana, S. (2017). *Data Mining K-Nearest Neigbor*. akultas Teknik Universitas Nusantara PGRI Kediri.

Rismayanti, R. (2017). IMPLEMENTASI ALGORITMA C4.5 UNTUK MENENTUKAN PENERIMA BEASISWA DI STT HARAPAN MEDAN. *JURNAL MEDIA INFOTAMA*, *12*(2). https://doi.org/10.37676/jmi.v12i2.413

Sembiring, M. A. A. S., Sibuea, M. F. L., & Sapta, A. (2018). Analisa Kinerja Algoritma C.45 Dalam Memprediksi Hasil Belajar. *Journal Of Science And Social Research*, *1*(1).

Shadman Roodposhti, M., Aryal, J., Lucieer, A., & Bryan, B. (2019). Uncertainty Assessment of Hyperspectral Image Classification: Deep Learning vs. Random Forest. *Entropy*, *21*(1), 78. https://doi.org/10.3390/e21010078

Tchomté, N., Asghar, S., Javaid, N., Dayang, P., Danga, D., & Oyono, D. (2020). A Case Based Reasoning Coupling Multi-Criteria Decision Making with Learning and Optimization Intelligences: Application to Energy Consumption. *EAI Endorsed Transactions on Smart Cities*, *4*(9), 162292. https://doi.org/10.4108/eai.26-6-2018.162292

Utomo, D. P., & Nasution, S. D. (2016). Sistem Pakar Mendeteksi Kerusakan Toner dengan Menggunakan Metode Cased Based- Reasoning. *Jurnal Riset Komputer (JURIKOM)*, *3*(5).

Widyastuti, M., Fepdiani Simanjuntak, A. G., Hartama, D., Windarto, A. P., & Wanto, A. (2019). Classification Model C.45 on Determining the Quality of Custumer Service in Bank BTN Pematangsiantar Branch. *Journal of Physics: Conference Series*, *1255*(1), 012002. https://doi.org/10.1088/1742-6596/1255/1/012002

Yuliansyah, H., Imaniati, R. A. P., Wirasto, A., & Wibowo, M. (2021). Predicting Students Graduate on Time Using C4.5 Algorithm. *Journal of Information Systems Engineering and Business Intelligence*, *7*(1), 67. https://doi.org/10.20473/jisebi.7.1.67-73

Zeng, G. (2020). On the confusion matrix in credit scoring and its analytical properties. *Communications in Statistics - Theory and Methods*, *49*(9), 2080–2093. https://doi.org/10.1080/03610926.2019.1568485

Zhang, Z., Yang, W., & Wushour, S. (2020). Traffic Accident Prediction Based on LSTM-GBRT Model. *Journal of Control Science and Engineering*, *2020*, 1–10. https://doi.org/10.1155/2020/4206919