CLASSIFICATION OF BLOOD DONOR DATA USING C4.5 AND K-NEAREST NEIGHBOR METHOD (CASE STUDY: UTD PMI BALI PROVINCE)

Ni Ketut Melly Astuti[1*]; Nengah Widya Utami[2], I Gede Putu Krisna Juliharta[3]

[1, 3] Program Studi Sistem Informasi; [2]Program Studi Sistem Informasi Akuntansi
STMIK Primakara
www.primakara.ac.id
[1*)] tutmelly117@gmail.com; [2] widya@primakara.ac.id; [3]krisna@primakara.ac.id

(*) Corresponding Author

Abstract— Classification of blood donor data at UTD PMI Bali Province by applying the C4.5 and K-Nearest Neighbor algorithms. The number of blood donor data donors is 34,948, of which 90% of the data, namely 31,454 is used as training data. Meanwhile, 10% of the data, which is 3,494 data, is used as the implementation of data testing using the Orange application. C4.5 obtained an accuracy score of 92.9%, F1 of 92.2%, Precision of 93.1%, Recall of 92.9%, specificity of 68.2%. While K-nearest neighbor obtained an accuracy score of 91%, F1 of 90.1%, Precision of 90.8%, Recall of 91%, specificity of 63%. With the AUC (Area Under Curve) value for the C4.5 algorithm is 0.875 and the K-nearest neighbor is 0.813 Good Classification. The results of the evaluation using the confusion matrix C4.5 obtained an accuracy score of 92.6%, F1 of 95.7%, Precision of 99.4%, Recall of 92.4%, specificity of 96%. While k-nearest neighbor obtained an accuracy score of 90.9%, F1 of 94.6%, Precision of 98.4%, Recall of 91.2%, specificity of 88.4%. Based on the evaluation of the confusion matrix and the ROC Analysis Graph, the C4.5 algorithm obtained higher results than the K-Nearest Neighbor algorithm. Based on the data on the characteristics of blood donors at UTD PMI Bali Province, it shows that the gender is male, Badung area, Age 20 to < 30, the occupation of private employees dominates in blood donation.

Keywords: data mining; classification; blood donation C4.5; K-Nearest Neighbor

Abstrak— Klasifikasi data pendonor darah di UTD PMI Provinsi Bali dengan menerapkan algoritma C4.5 dan *K-Nearest Neighbor*. Jumlah data data pendonor darah sebanyak 34.948, dimana 90% data yaitu 31.454 digunakan sebagai data *training*. Sedangkan 10% data yaitu 3.494 data digunakan sebagai data *testing* implementasi dengan menggunakan Aplikasi orange. C4.5 memperoleh *score* akurasi sebesar 92,9%, F1 sebesar 92,2%, *Precision* sebesar 93,1%, *Recall* sebesar 92,9%, *specificity* sebesar 68,2%. Sedangkan K-*nearest neighbor* memperoleh score akurasi sebesar 91%, F1 sebesar 90,1%, *Precision* sebesar 90,8%, *Recall* sebesar 91%, *specificity* sebesar 63%. Dengan nilai AUC (*Area Under Curve)* algoritma C4.5 sebesar 0.875 dan *K-nearest neighbor* sebesar 0.813 *Good Classification.* Hasil evaluasi menggunakan *confusion matrix* C4.5 memperoleh score akurasi sebesar 92,6%, F1 sebesar 95,7%, *Precision* sebesar 99,4%, *Recall* sebesar 92,4%, *specificity* sebesar 96%. Sedangkan *k-nearest neighbor* memperoleh score akurasi sebesar 90,9%, F1 sebesar 94,6%, *Precision* sebesar 98,4%, *Recall* sebesar 91,2%, *specificity* sebesar 88,4%. Berdasarkan evaluasi *confusion matrix* dan Grafik *ROC Analysis* algoritma C4.5 memperoleh hasil lebih tinggi daripada algoritma *K-Nearest Neighbor*. Berdasarkan data prediksi karakteristik pendonor darah di UTD PMI Provinsi Bali menunjukkan jenis kelamin laki-laki, wilayah Badung, Umur ≥ 20 sampai < 30, pekerjaan pegawai swasta mendominasi dalam donor darah.
Kata kunci*: data mining*, klasifikasi, donor darah*, C4.5, K-Nearest Neighbor*

INTRODUCTION

Today in the world of Health, the number of patients who need blood is increasing. However, the blood supply at the blood donation center is reduced due to a decrease in the number of blood donors (Djuardi, 2020). Blood donation is the process of taking blood owned by a given person and then storing and used at any time for blood transfusions (Indonesia, 2015). The standard blood taken from the donor is 350 milliliters (Firdaus, Latif, & Gata, 2020). Blood donation is one of the noble activities because it can help patients who need blood donors.

Unit Transfusion Darah (UTD) PMI Provinsi Bali is one of the agencies that serve the Bali area blood donor activities that do blood donors voluntarily. Based on the results of interviews conducted with resource persons in the Head of Research and Development and Quality Division UTD PMI Bali Province, namely dr. Ni Putu Chandra

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18 No. 1 March 2022 | DOI: 10.33480/pilar.v18i1.2790

Indriasari. In 2021 the bloodstock will decrease drastically by up to 60%. One of them is caused by blood donor units that are unable to carry out blood donation activities. The strategy used by UTD PMI Bali Province in meeting the hospital's demand for blood needs has not been able to meet the blood needs needed by the hospital for patients. UTD PMI Bali Province has used a blood donor system so that a database for transfusion services has been formed.

A large number of databases can be used as analysis material to find certain knowledge, information, patterns, or trends that are very useful and usually not realized that can be used to make certain decisions, which is often known as data mining. (Pahlevi, Fredlina, & Utami, 2021). In the world of Health, data mining has been used to predict disease and to classify potential donors (Amalia, 2018). C4.5 algorithm and k-nearest neighbor are algorithms that can be used in medical diagnosis (Suyanto, 2017). The advantage of C4.5 is that it converts data into decision trees and produces understandable rules (Haudi, S.Pd., M.M., 2021). Meanwhile, the k-nearest neighbor has the advantage of a fast and effective learning process used with large training data (Kodati & Vivekanandam, 2018).

In previous research, data mining was used to determine the eligibility factor for blood donation by Abdul Latif and Dini Silvi Purnia in 2019 using the C4.5 decision tree algorithm, obtaining an accuracy rate of 97.69% (Latif & Purnia, 2019). In this research, the prediction of potential blood donors uses the naive Bayes algorithm, k-nearest neighbors, and decision tree C4.5. By Hermanto Wahono and Dwiza Riana in 2020 the nave Bayes algorithm accuracy rate of 85.15%, k-nearest neighbors 84.10%, and decision tree C4.5 93.83% (Wahono & Riana, 2020). Based on these two studies, C4.5 and k-nearest neighbors have a very good accuracy rate above 80%. In addition, the research has a similarity in the dataset owned by the author. For the best results from data mining classification, it is better to use two algorithms as a comparison to see which algorithm is more effective and can be used as a benchmark for the level of the best algorithm ability.

The purpose of this study is to classify data on blood donors using the C4.5 and k-nearest neighbor methods to identify potential and non-potential donors. In addition, it aims to determine the level of accuracy of the two algorithms.

MATERIALS AND METHODS

The research method is carried out with the stages of the Knowledge Discovery in Data (KDD) method. The research flow is the research stage to explain the steps used in carrying out the research. The data source of this research is primary data which is directly obtained from the informant.

Identification of problems

At this stage, the author conducted interviews with the Head of Research and Development and Quality at UTD PMI Bali Province. Regarding the problems that occurred at UTD PMI Bali Province and what solutions can be given in solving these problems. The author has explained this in the introduction to this research.

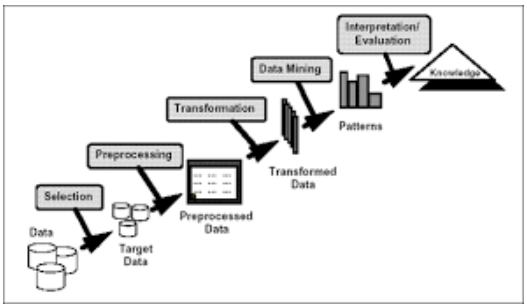
Study of literature

The literature study stage carried out to support problem-solving was found in the research. The author searches for the theoretical basis that supports and will be used in research that comes from several theses, books, scientific journals, and scientific articles.

Method of collecting data

After conducting a literature study, the authors collected data by conducting interviews with the Head of Research and Development & Quality Headquarters at UTD PMI Bali Province. Interviews were conducted to find out information about the process of donating blood, as well as the criteria used as requirements for blood donation. The data needed in this study is a dataset of blood donors obtained from UTD PMI Bali Province.

Knowledge Discovery in Data



Sumber:(Jejaring, 2019)

Gambar 1 konsep tahapan KDD

Data Selection

At the selection stage, the data obtained from the PMI database is analyzed to determine what parameters are needed based on the needs analysis obtained. The parameters used as the PMI dataset parameters are age, gender, blood pressure, hemoglobin level, body weight, temperature, number of donors, occupation, address, and class category.

Data Preprocessing

At this stage, the data will be cleaned. Quality data greatly affects the quality of the

decisions that will be obtained. Unqualified data is data with missing attribute values, errors, data duplication, missing values, and inconsistent data in filling out the attributes. (Utami, Sukajaya, Made Candiasa, & Dewi, 2019). The data obtained in this study were 38, 413 records. From data cleaning, there were 3,464 deleted data. Because some attributes are irrelevant, there is missing value data, and the address written on the data comes from outside Bali. Furthermore, the researchers deleted the data to obtain good classification results. So that the total number of data in the dataset is 34, 948 records will be processed in the orange application for classification and evaluation of the method used. The distribution of data used in the classification process is 90:10.

Transformation

At this stage, the transformation process aims to improve accuracy and efficiency. Data is converted or combined into a format suitable for processing in data mining. The transformations carried out in the study were changing the address in the form of a street into a district which was done manually, changing the hemoglobin format from hourly format to general, changing the blood pressure format from numbers to nominal which was done manually in Microsoft excel.

For testing the C4.5 algorithm has the advantage of processing data with nominal, orthogonal, and continuous values (Latif & Purnia, 2019). So that the values of each attribute contained in the dataset do not need to be transformed. However, for testing the k-nearest neighbor algorithm, the data used is numerical, the data must be transformed first. Because the author's research uses orange data mining tools, the transformation process can be done automatically by orange data mining tools.

Data Mining

The next stage is a data processing to look for interesting patterns or information in the PMI dataset using the C4.5 algorithm and k-nearest neighbor which is implemented in the orange data mining tools.

1. C4.5 is an algorithm used to form a decision tree.
2. K-Nearest Neighbor (K-NN) is an algorithm for classifying objects based on the learning data that is closest to the object.

Interpretation / Evaluation

At this stage, the testing phase is carried out using the PMI dataset by looking at the accuracy results in the classification process on the PMI dataset using the C4.5 and k-nearest neighbor method algorithms. And evaluated using the confusion matrix method and the ROC. curve
At this stage, explain the results of the research briefly and clearly regarding the results of the classification of blood donor data and the level of ability using the C4.5 and k-nearest neighbor methods.

RESULTS AND DISCUSSION

This study aims to classify blood donors who have the potential to donate blood back to the blood donor unit and see the level of ability of the C4.5 and k-nearest neighbor algorithms using the PMI dataset. Then analyze the ability of the two algorithms by comparing the two algorithms in each dataset. So that the most suitable dataset and algorithm are obtained to classify blood donors who will donate blood again or not.

The application of data mining for data classification of blood donors with the C4.5 and k-nearest neighbor algorithms is carried out using the Orange data mining tools. Orange data mining is an open-source machine learning technology and data visualization (Utami & Paramitha, 2021). The classification model implemented using Orange data mining is as shown in Figure 2.

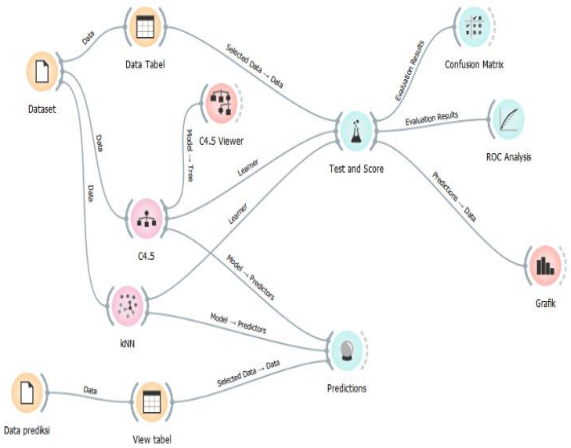


Figure 2. Classification Model Design

Figure 2 above shows the design of the classification model on the Orange data mining tools by involving the widget file, data table, C4.5, Viewer C4.5, k-NN, test & score, confusion matrix, ROC analysis, prediction, graph.

After the implementation of the blood donor data on the Orange data mining tools, then the results of the test & score on the C4.5 and K-nearest neighbor algorithms on the Orange data mining tools will be obtained.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18 No. 1 March 2022 | DOI: 10.33480/pilar.v18i1.2790

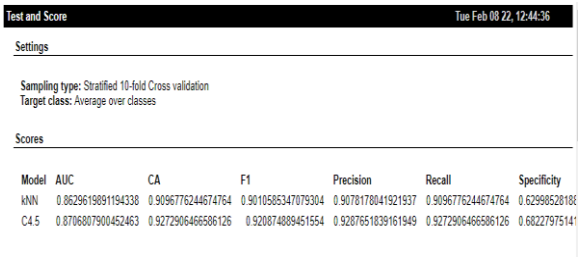


Figure 3. Test Results & Score

Test & score is the process of testing the algorithm on the data. The results show a table with different classifier performance measures, such as classification accuracy and area under the curve. Below is a table 1 comparison of the performance of the two algorithms.

Table 1. Comparison of Test & Score

Model	*Score*					
	AUC	CA	F1	*Precision*	*Recall*	*specificity*
C4.5	*87,5%*	*92,9%*	*92,2%*	*93,1%*	*92,9%*	*68,2%*
k-NN	*81,3%*	*91%*	*90,1%*	*90,8%*	*91%*	*63%*

Based on table 1, the comparison of the data used in the data mining process is 90:10. By using training data, 90% of the 34,948 data are 31,454 blood donor data. Data testing 10%, namely 3,494 blood donor data obtained the highest score on both algorithms.

The performance of the AUC accuracy level can be divided into several groups, namely (Gorunescu, 2011):

a. 0.90 – 1.00 = *Excellent Classification*
b. 0.80 – 0.90 = *Good Classification*
c. 0.70 – 0.80 = *Fair Classification*
d. 0.60 – 0.70 = *Poor Classification*
e. 0.50 – 0.60 = *Failure Classification*

Based on the classification results, it can be concluded that the C4.5 and K-NN algorithms are accurate for predicting the classification of blood donor data because the AUC value is included in the Performance Good Classification (0.80 – 0.90). The decision tree generated by the C4.5 algorithm on the implementation of the orange data mining tools is depicted in Figure 4 below.

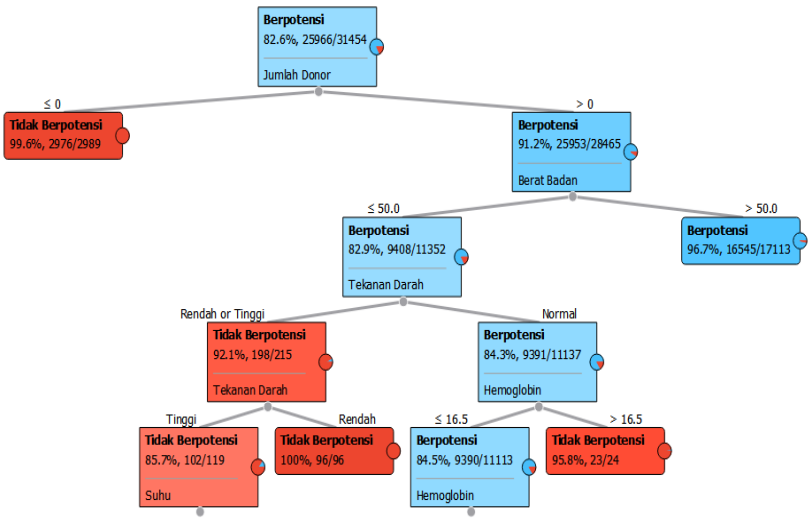


Figure 4 Decision Tree Display from Classification Results C4.5

Evaluation and Validation of Classification Results

After the data is processed, the level of ability can be tested to see the performance of the C4.5 and KNN algorithm methods. Ability level testing is carried out using a confusion matrix and ROC/AUC (Area Under Cover) curve.

Confusion Matrix Algoritma C4.5

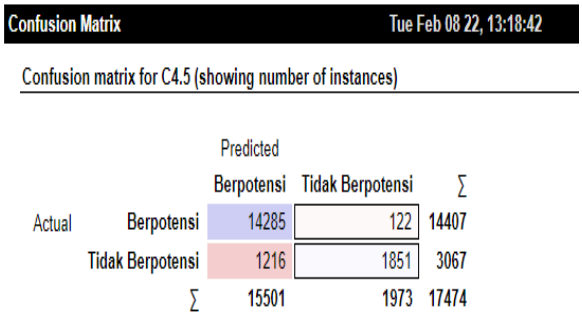


Figure 5 Confusion Matrix C4.5

The manual calculation of the value in Figure 5 is as follows:

$$Accuracy = \frac{25.830 + 3.318}{31.454} = 0{,}926 = 92{,}6\%$$

$$Precision = \frac{25.830}{136 + 25.830} = 0{,}994 = 99{,}4\%$$

$$Recall = \frac{25.830}{2.110 + 25.830} = 0{,}924 = 92{,}4\%$$

$$F_1 = \frac{2\ (0{,}924x\ 0{,}994)}{0{,}924+0{,}994} = 0{,}957 = 95{,}7\%$$

$$specificity = \frac{3.318}{136 + 3.318} = 0{,}960 = 96\%$$

Based on Figure 5 of 31,454 donor data that will be classified using the C4.5 algorithm. A total of 5,488 data on blood donors in the non-potential category and 25,966 data on blood donors in the potential category. Data for donors in the category of potential donors that match the classification is 25,830 records. Meanwhile, the data for donors in the potential category included in the non-potential category were 136 records. There are 2,110 records in the non-potential category that are included in the potential classification. Meanwhile, the data for the non-potential category according to the classification is 3,378 records.

Confusion Matrix Algoritma *K-nearest neighbor*

Confusion Matrix — Tue Feb 08 22, 13:19:38

Confusion matrix for kNN (showing number of instances)

		Predicted: Berpotensi	Predicted: Tidak Berpotensi	∑
Actual	Berpotensi	14165	242	14407
	Tidak Berpotensi	1657	1410	3067
	∑	15822	1652	17474

Figure 6 Confusion matrix k-NN

The manual calculation of the value in Figure 6 is as follows:

$$Accuracy = \frac{25.567 + 3.046}{31.454} = 0{,}909 = 90{,}9\%$$

$$Precision = \frac{25.567}{399 + 25.567} = 0{,}984 = 98{,}4\%$$

$$Recall = \frac{25.567}{2.442 + 25.567} = 0{,}912 = 91{,}2\%$$

$$F_1 = \frac{2\ (0{,}912\ x\ 0{,}984)}{0{,}912 + 0{,}984} = 0{,}946 = 94{,}6\%$$

$$Specificity = \frac{3.046}{399 + 3.046} = 0{,}884 = 88{,}4\%$$

Based on Figure 6 of 31,454 donor data that will be classified using the k-NN algorithm. A total of 5,488 data on blood donors in the non-potential category and 25,830 data on blood donors in the potential category. Data for donors in the category of potential donors that match the classification is 25,567 records. Meanwhile, the data for donors in the potential category included in the non-potential category were 399 records. There are 2,442 records in the non-potential category that are included in the potential classification. Meanwhile, the data for the non-potential category according to the classification is 3,046 records.

Table 2 Comparison of Confusion Matrix Results

Model	Confusion Matrix Evaluation				
	Accuracy	*Precision*	*Recall*	F1	*specificity*
C4.5	92,6 %	99,4%	92,4%	95,7%	96%
kNN	90,9%	98,4%	91,2%	94,6%	88,4%

Based on table 2, the evaluation of the C4.5 and K-nearest neighbor algorithms using the confusion matrix algorithm C4.5 shows better results.

ROC Analysis Chart

ROC Analysis merupakan kurva yang menggambarkan probabilitas dengan variabel *sensitivitas* dan *specificity* dengan nilai batas antara 0 hingga 1 (Orangedatamining.com, 2021). Semakin dekat kurva mengikuti batas kiri dan kemudian batas atas ruang ROC, semakin akurat pengklasifikasinya. *ROC Analysis* menunjukkan performa kedua algoritma dalam mengukur atau menganalisa data. Gambar 7 Grafik *ROC Analysis* untuk mengetahui kinerja algoritma klasifikasi.

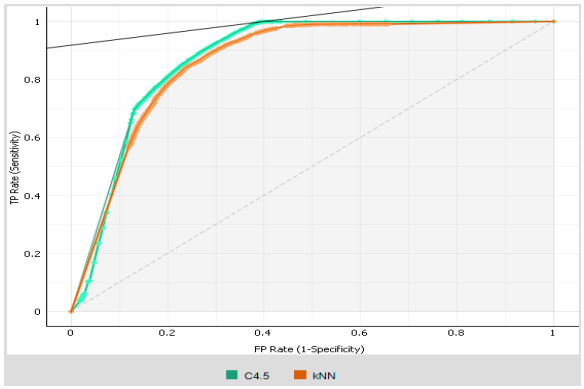


Figure 7 Potential ROC Analysis

Based on Figure 7 above, the ROC Analysis graph shows the potential category. Where the curve with the color Tosca is the C4.5 algorithm and the curve with the orange color is the k-nearest neighbor

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18 No. 1 March 2022 | DOI: 10.33480/pilar.v18i1.2790

algorithm. Based on the graph above, it can be concluded that the performance of the C4.5 curve is better than the performance of the K-nearest neighbor curve. In Figure 7 above, it can be seen that the model's performance is very good where the intersection of the line between the TP Rate (Sensitivity) and FP Rate (1-Specificity) is close to a value of 1, so the model classifies data that has the potential to show good classification. (Orangedatamining.com, 2021).

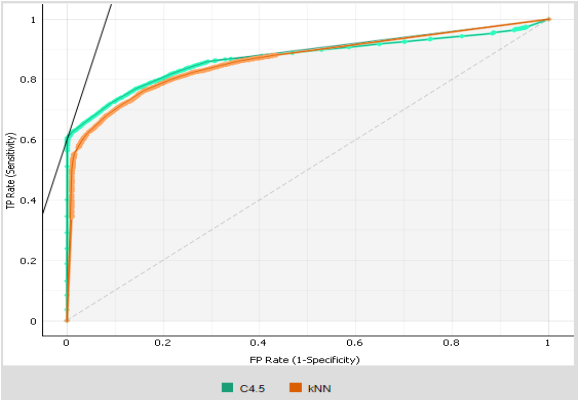


Figure 8. ROC Analysis No Potential

Based on Figure 8 above, it shows the ROC Analysis chart for the non-potential category. Where the curve with the color Tosca is the C4.5 algorithm and the curve with the orange color is the k-nearest neighbor algorithm. Based on the graph above, it can be concluded that the performance of the C4.5 curve is better than the performance of the K-nearest neighbor curve. In Figure 7 above, it can be seen that the model's performance is very good, where the intersection of the line between the TP Rate (Sensitivity) and FP Rate (1-Specificity) is close to a value of 1, so the model classifying data does not have the potential to show good classification. (Orangedatamining.com, 2021).

CONCLUSION

The results of the evaluation using the confusion matrix C4.5 obtained an accuracy score of 92.6%, F1 of 95.7%, Precision of 99.4%, Recall of 92.4%, specificity of 96%. While k-nearest neighbor obtained an accuracy score of 90.9%, F1 of 94.6%, Precision of 98.4%, Recall of 91.2%, specificity of 88.4%. With the AUC (Area Under Curve) algorithm C4.5 of 0.875 and K-nearest neighbor of 0.813 Good Classification The evaluation results using the ROC analysis graph, the performance of the C4.5 curve is better than the performance of the K-nearest neighbor curve. The results of the evaluation using the confusion matrix C4.5 obtained an accuracy score of 92.6%, F1 of 95.7%, Precision of 99.4%, Recall of 92.4%, specificity of 96%. While k-nearest neighbor obtained an accuracy score of 90.9%, F1 of 94.6%, Precision of 98.4%, Recall of 91.2%, specificity of 88.4%. Based on the evaluation of the confusion matrix, the C4.5 algorithm obtained higher results than the K-nearest neighbor algorithm

Based on the predictive data on the characteristics of blood donors at UTD PMI Bali Province, it shows that male gender, donors having their address in South Denpasar, donors who have private employees, and donors aged 20 to <30 dominate the participation in donating blood.

REFERENCE

Amalia, H. (2018). Perbandingan Metode Data Mining SVM Dan NN Untuk Klasifikasi Penyakit Ginjal Kronis. *Pilar Nusa Mandiri: Journal of Computing and Information System*, *14*(1), 1-6. Retrieved from http://ejournal.nusamandiri.ac.id/index.php/pilar/article/view/80

Djuardi, A. M. P. (2020). Donor Darah Saat Pandemi COVID-19. *Jurnal Medika Hutama*, *2*(1), 298–303. Retrieved from http://www.jurnalmedikahutama.com/index.php/JMH/article/view/74

Firdaus, M. R., Latif, A., & Gata, W. (2020). Klasifikasi Kelayakan Calon Pendonor Darah Menggunakan Neural Network. *Sistemasi*, *9*(2), 362. https://doi.org/10.32520/stmsi.v9i2.840

Gorunescu, F. (2011). Data Mining: Concepts, Models and Techniques. In *Data mining - Concepts, Models and Technique*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-19721-5

Haudi, S.Pd., M.M., D. B. . (2021). *Teknik Pengambilan Keputusan* (C. Hadion Wijoyo, S.E., S.H., S.Sos.M.H., M.M., AK., Ed.). Sumatra Barat: CV Insan Cendekia Mandiri.

Indonesia, M. K. R. (2015). *Peraturan Menteri Kesehatan RI No 91 Tahun 2015 Tentang Standar Pelayanan Transfusi Darah*. *2009*, 1–27.

Jejaring. (2019). Tahap-Tahapan Knowladge Discovery In Database (KDD). Retrieved from jejaring.web.id website: https://www.jejaring.web.id/tahap-tahapan-knowladge-discovery-in-database-kdd/

Kodati, S., & Vivekanandam, R. (2018). Analysis of Heart Disease using in Data Mining Tools Orange and Weka Sri Satya Sai University Analysis of Heart Disease using in Data Mining Tools Orange and Weka. *Global Journal of Computer Science and Technology*, *18*(1).

Latif, A., & Purnia, D. S. (2019). Implementasi Data

Mining Untuk Mengetahui Faktor Kelayakan Donor Darah UTD Kota Tasikmalaya Menggunakan Algoritma C4.5. *INTI Nusa Mandiri*, *14*(1), 145–150.

Orangedatamining.com. (2021). ROC Analysis. Retrieved from Orange website: https://orangedatamining.com/widget-catalog/evaluate/rocanalysis/

Pahlevi, R., Fredlina, K. Q., & Utami, N. W. (2021). Penerapan Algoritma Id3 Dan Svm Pada Klasifikasi Penyakit Diabetes Melitus Tipe 2. *Prosiding Snast*, *2*, 64–75.

Suyanto. (2017). *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung, Indonesia: Informatika.

Utami, N. W., & Paramitha, A. A. I. I. (2021). Penerapan Data Mining Untuk Mengetahui Pola Pemilihan Program Studi Di Stmik Primakara Menggunakan Algoritma K-Means Clustering. *Jurnal Teknologi Informasi Dan Komputer*, *7*(4), 456–463.

Utami, N. W., Sukajaya, I. N., Made Candiasa, I., & Dewi, E. G. A. (2019). The implementation of data mining to show UKT (students' tuition) using fuzzy C-means algorithm: (Case study: Universitas Pendidikan Ganesha). *2019 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2019*, 101–106. https://doi.org/10.1109/ICACSIS47736.2019.8979933

Wahono, H., & Riana, D. (2020). Prediksi Calon Pendonor Darah Potensial Dengan Algoritma Naïve Bayes, K-Nearest Neighbors dan Decision Tree C4.5. *JURIKOM (Jurnal Riset Komputer)*, *7*(1), 7. https://doi.org/10.30865/jurikom.v7i1.1953

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18 No. 1 March 2022 | DOI: 10.33480/pilar.v18i1.2790