JSITIK, Vol.1 No.1 Desember 2022
ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v1i1.154

Analisa Splitting Criteria Pada Decision Tree dan Random Forest untuk Klasifikasi Evaluasi Kendaraan

Diterima:
19 November 2022
Revisi:
10 Desember 2022
Terbit:
31 Desember 2022

Arie Nugroho
Universitas Nusantara PGRI Kediri

Abstrak—Klasifikasi adalah salah satu topik dalam data mining. Algoritma atau model yang termasuk dalam klasifikasi antara lain Decision tree, K-NN, Naïve bayes. Decision tree merupakan model yang mudah untuk dipahami karena dapat divisualisasikan. Random Forest adalah salah satu model dalam klasifikasi yang merupakan pengembangan dari decision tree. Pemilihan splitting criteria dalam decision tree dan random forest dapat mempengaruhi hasil akurasi. Dalam artikel ini memaparkan perbandingan splitting criteria dalam model klasifikasi dengan decision tree dan random forest untuk data evaluasi kendaraan. Dengan menggunakan split data dan cross validation serta pengujian dengan confusion matrix, pemilihan splitting criteria memberikan pengaruh pada nilai akurasi dari model yang telah dihasilkan.
Kata Kunci—Splitting Criteria; Decision Tree; Random Forest; Klasifikasi

Abstract— *Classification is one of the topics in data mining. Algorithms or models included in the classification include Decision tree, K-NN, Naïve bayes. The decision tree is an easy model to understand because it can be visualized. Random Forest is one of the models in the classification that is the development of the decision tree. The selection of splitting criteria in decision trees and random forests can affect accuracy results. This article describes the comparison of splitting criteria in classification models with decision trees and random forests for vehicle evaluation data. By using split data and cross validation as well as testing with confusion matrix, the selection of splitting criteria has an influence on the accuracy value of the model that has been produced.*
Keywords— *Splitting Criteria; Decision Tree; Random Forest; Classification*

Penulis Korespondensi:

Arie Nugroho,
Program Studi Sistem Informasi,
Universitas Nusantara PGRI Kediri,
Email: arienugroho@unpkediri.ac.id

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v1i1.154

I. PENDAHULUAN

Transportasi adalah salah hal penting dalam kehidupan manusia. Para konsumen mempertimbangkan banyak hal dalam pemilihan alat transportasi yang sesuai dengan kebutuhannya. Untuk kenyamanan, mobil adalah salah satu pilihan alat transportasi atau kendaraan terbaik. Dalam pemilihan mobil , biasanya konsumen akan mempertimbangkan biaya, keamanan dan fasilitasnya. Mobil dengan harga yang tinggi tentunya akan mempunyai kemewahan dan kenyamanan yang lebih, tapi belum tentu semua konsumen membutuhkan hal tersebut. Perusahaan mobil harus mempertimbangkan tingkat penerimaan mobil oleh para konsumen, sehingga produksi mobil akan lebih efektif, meminimalkan kerugian dan meningkatkan keuntungan. Untuk membantu perusahaan mobil menentukan selera konsumen, dapat dengan cara mempelajari data penerimaan mobil yang telah ada.

Artificial Intelligence akan membantu manusia dalam menentukan keputusan sesuai kebutuhannya. *Data Mining* adalah salah satu bidang dalam *Artificial Intelligence*[1]. Dengan menerapkan *data mining* akan membantu para produsen mobil untuk menentukan kelayakan mobil yang akan digunakan sebagai pertimbangan dalam produksi mobil berikutnya. Dalam *data mining* menggunakan model *machine learning* untuk membuat model atau pola dari data yang akan digunakan[2]. Data yang mempunyai label atau target termasuk dalam klasifikasi. Model yang dihasilkan dari *data training* akan diuji dengan *data testing* untuk menguji akurasinya.

Pada penelitian ini membandingkan 3 (tiga) teknik *splitting criteria* pada 2 (dua) model algoritma *machine learning* untuk proses klasifikasinya. Perbandingan tersebut akan digunakan untuk memperoleh hasil akurasi terbaik pada dataset yang akan digunakan. Data yang digunakan adalah salah satu dataset publik dari *UCI Repository*. Pemilihan dataset publik agar di kemudian hari dapat menjadi bahan riset oleh peneliti lain untuk memberikan kontribusi pada pengetahuan. Model yang akan dibandingkan adalah *decision tree* dan *random forest* dengan pengaturan pemilihan *splitting criteria*. *Decision tree* menghasilkan model berupa pohon (*tree*).

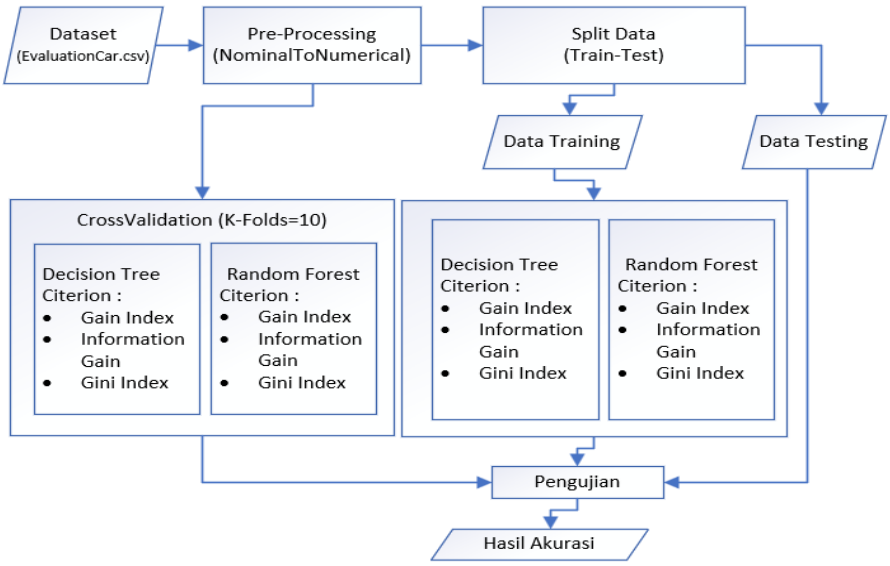
Random forest adalah pengembangan dari *decision tree*. Ada beberapa penelitian terkait yang membahas klasifikasi dengan dataset evaluasi kendaraan ini. Pertama, perbandingan performa dari 3 (tiga) model, yaitu *decision tree, naïve bayes* dan *artificial neural network*(ANN). Artikel tersebut menggunakan konversi atribut dari tipe data nominal ke numerik sebagai proses data cleaning kemudian melakukan proses transformasi menggunakan normalisasi min-max. Hasilnya adalah ANN menghasilkan akurasi tertinggi dibanding metode lain dengan waktu yang lebih lama[3]. Penelitian kedua membahas tentang analisa performa dari banyak model klasifikasi, yaitu *naïve bayes, decision tree* dan *rotation forest*. Pada penelitian ini hasilnya adalah model *rotation forest* memberikan akurasi terbaik[4]. Pada penelitian ini tujuannya adalah mencari

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v1i1.154

akurasi terbaik dari model *decision tree* dan *random forest* dengan pemilihan *splitting criteria* yang ada dalam parameter *decision tree* dan *random forest,* yaitu *information gain, gain ratio* dan *gini index*. Pada *decision tree*, setiap *node* adalah representasi dari atribut, sedangkan class atau label diwakili dengan daun serta pemilihan atribut yang teratas disebut dengan root[5]. *Random forest* adalah salah satu metode atau model *ensemble learning* yang merupakan kombinasi dari beberapa *decision tree* atau pohon yang atribut *root*-nya dipilih secara acak, kemudian untuk hasil klasifikasinya didapat dari pengambilan suara terbanyak (*votting*) dari jumlah pohon yang telah ditetapkan[6]. Random forest juga merupakan salah satu algoritma ensemble learning yang banyak digunakan untuk klasifikasi[7].

II. METODE

Penelitian ini membahas tentang metode klasifikasi pada data mining dengan pemilihan *splitting criteria* pada algoritma *decision tree* dan *random forest*. Setiap model yang akan diuji dengan beberapa kriteria dalam pemilihan *splitting criteria* (*criterion*), yaitu *gain index*, *information gain* dan *gini index*. Ketiga *criterion* tersebut merupakan parameter dalam *decision tree* dan *random forest*[8]. Model akan dibuat dengan 2 (dua) cara pembagian data *training* dan *testing*, yaitu *split data* dan *cross validation*. Alur dalam penelitian ini ditunjukkan pada gambar 1.



Gambar 1. Alur Penelitian

Pada gambar 1 ditunjukkan alur penelitian dimulai dengan langkah pertama adalah membaca dataset dengan tipe *file comma separated value* (CSV) kemudian melakukan *pre-processing* yaitu

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v1i1.154

dengan mengubah atribut nominal menjadi numerik. Hal ini dilakukan untuk mempernudah proses klasifikasi[9]. Berikutnya adalah *split dataset* untuk *training* dan *testing* serta *cross validation*. Dalam setiap model akan diatur dengan *parameter* yang berbeda-beda, yaitu dengan pilihan tiga *criterion*. Berikutnya adalah menampilkan akurasi dengan confusion matriks. Langkah pertama adalah memilih dataset yang akan digunakan. Penelitian ini menggunakan salah satu dataset publik *dari UC Irvine Machine Learning repository*, yaitu evaluasi mobil. Dataset ini mempunyai 1728 record dan 7 atribut yang terdiri dari 6 fitur dan 1 label serta tidak mempunyai *missing values*[10]. Atribut-atributnya adalah *safety* (tingkat keamanan), *door* (jumlah pintu), *maint* (perawatan), *person* (kapasitas penumpang), *Lug-boot* (ukuran bagasi) dan *buying* (harga mobil). Label atau targetnya adalah *class* yang pilihan nilainya adalah *unacc* (tidak diterima), *acc* (diterima), *good* (baik), *v-good* (sangat baik). Semua nilai dari atributnya adalah nominal kecuali jumlah pintu dan kapasitas penumpang yang bertipe numerik. Jumlah record untuk label *unacc* ada 1210 record, *acc* sebanyak 384 record, *good* sebanyak 69 reecord dan *v-good* sebanyak 65 record. Deskripsi nilai dari atribut-atributnya ditunjukkan pada tabel 1.

Tabel 1.Nilai dari atribut dan label

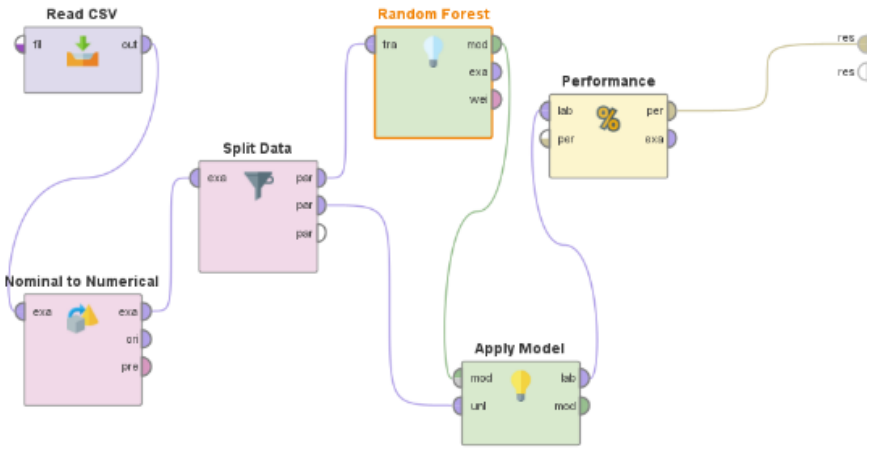
Kategori	Nama	Nilai
Atribut	Safety	low,med,high
	Lug-boot	small,med,big
	Door	2,3,5more
	Person	2,4,more
	Maint	low,med,high,v-high
	Buying	low,med,high,v-high
Label	Class	unacc,acc,good,v-good

Langkah kedua adalah *split* data *training* dan *testing* yang dimulai dari 90 % :10 % s/d 70%:30%[11]. Sebagai pembanding juga dilakukan *cross validation* dengan 10 *k-folds*, yaitu dataset dibagi menjadi 10 bagian[12], yaitu 10 % data pertama dijadikan *training*, sisanya dijadikan *testing*, kemudian 10 % data kedua dijadikan *training* dan sisanya sebagai data *testing*, dan seterusnya[13]. Dengan 10 *k-folds* tersebut dilakukan sebanyak 10 kali[14]. Langkah ketiga adalah membuat model dari *decision tree* dan *random forest*. *Random forest* dan normalisasi atribut dengan mengubah tipe data nominal menjadi numerik dapat meningkatkan akurasi pada klasifikasi[15]. Setiap model yang dibuat menggunakan pemilihan atribut dengan 3 (tiga) pilihan *criterion*. Langkah keempat adalah menampilkan akurasi dari setiap percobaan yang telah dilakukan kemudian dianalisa untuk menjelaskan kontribusi penelitian.

JSITIK, Vol.1 No.1 Desember 2022
ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v1i1.154

III. HASIL DAN PEMBAHASAN

Hasil dari penelitian ini adalah nilai akurasi dari setiap model *decision tree* dan *random forest* dengan 3 (tiga) parameter *criterion* dan 2 (dua) teknik pengujian yaitu *split data* dan *cross validation*. Pembagian data pada proses *splitting* menjadi *training* dan *testing*. *Splitting* data *training* dimulai dari 90 sd 70 %, sedangkan *splitting* data *testing* dimulai dari 10 sd 30 %. Proses pembuatan model *random forest* dengan *split data* ditunjukkan dengan gambar 2.



Gambar 2. Proses Model Random Forest dengan split data

Pada proses pembuatan model dengan *random forest* dan *split data* , diawali dengan *read* csv , yaitu membaca dataset. Berikutnya dilakukan normalisasi atribut dari nominal ke numerik untuk memudahkan dalam klasifikasi. Proses berikutnya adalah *split* data *training* dan *testing*, kemudian dibuat model *random forest* dan memilih atribut untuk testing. Setelah model digunakan dengan data *testing*, berikutnya adalah menampilkan akurasi. Hal yang sama juga diberlakukan dengan algoritma *decsion tree*. Pada proses pembuatan model dengan *random forest* dan *decision tree* menggunakan *cross validation* , diawali dengan read csv , yaitu membaca dataset.

Berikutnya dilakukan *cross validation* dengan 10 *k-folds*, yang di dalamnya dibuat model *random forest* dan *decision tree*. Setelah model digunakan, berikutnya adalah menampilkan akurasi. Pada model *decision tree* dengan *criterion gain ratio* dan *gini index* memperoleh akurasi tertinggi sebesar 93,60 % dengan 90 % data training dan 10 % data testing. Pada criterion information gain memperoleh akurasi tertinggi yaitu 94,19 % dengan ukuran data training dan testing yang sama. Pada model random forest dengan criterion gain ratio, information gain dan gini index memperoleh akurasi tertinggi sebesar 96,51 % dengan 90 % data training dan 10 % data testing, hasil pengujian ditunjukkan pada tabel 2.

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v1i1.154

Tabel 2. Pengujian dengan Split data

Model	Criterion	Training(%)	Testing(%)	Akurasi(%)
Decision Tree	Gain ratio	90	10	93,60
		80	20	91,62
		70	30	88,03
	Information gain	90	10	94,19
		80	20	90,17
		70	30	88,22
	Gini Index	90	10	93,60
		80	20	91,62
		70	30	88,03
Random Forest	Gain ratio	90	10	96,51
		80	20	94,22
		70	30	94,59
	Information gain	90	10	96,51
		80	20	93,93
		70	30	93,63
	Gini Index	90	10	96,51
		80	20	93,35
		70	30	93,63

Pengujian dengan *split data* yang ditampilkan pada tabel 2 menunjukkan pemilihan *criterion* pada model *decision tree* dan *random forest* relatif ada perbedaan nilai akurasi. Pengujian dengan *cross validation* dari *decision tree* memperoleh hasil akurasi tertinggi dengan *criterion information gain* sebesar 93,06 % dan untuk *random forest* memperoleh akurasi tertinggi dengan *criterion gain ratio* sebesar 94,56 %, hasil pengujian ditunjukkan pada tabel 3.

Tabel 3. Pengujian dengan Cross Validation

Model	Criterion	Akurasi(%)
Decision Tree	Gain ratio	92,94
	Information gain	93,06
	Gini Index	92,77
Random Forest	Gain ratio	94,56
	Information gain	94,33
	Gini Index	94,21

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v1i1.154

Pengujian dengan *cross validation* ditampilkan pada tabel 3 menunjukkan pemilihan criterion pada model *decision tree* ada perbedaan nilai akurasi, sedangkan untuk model *random forest* relatif tidak ada perbedaan.

IV. KESIMPULAN

Pengujian dengan *split data* pada kedua model dan *criterion* yang dipilih pada dataset evaluasi mobil memperoleh hasil akurasi tertinggi pada 90 % data *training* dan 10 % data *testing* . Hasil akurasi cenderung menurun ketika prosentase data *training* berkurang. Hal ini membuktikan bahwa semakin besar data *training* , akurasi akan semakin tinggi, karena semakin banyak data *training* yang digunakan , model akan semakin banyak mengenali pola dari data. Dibandingkan dengan *split data* , pengujian dengan *cross validation* , akurasi menurun dan membutuhkan waktu yang lebih lama. Hal ini disebabkan dalam 10 *k-fold cross validation* semua data digunakan sebagai data training dan testing, data dibagi menjadi 10 bagian , kemudian setiap bagiannya digunakan sebagai training dan testing secara bergantian, hal ini yang membuat mengapa cross validation membutuhkan waktu yang lebih lama , tapi dengan *cross validation* pengujian model menjadi lebih adil (*fair*). Penelitian berikutnya dapat ditambahkan dengan pengaturan parameter yang lain dan menambahkan dataset sebagai pembanding.

DAFTAR PUSTAKA

[1] Q. V. Pham, D. C. Nguyen, T. Huynh-The, W. J. Hwang, and P. N. Pathirana, “Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts,” IEEE Access, vol. 8, pp. 130820–130839, 2020, doi: 10.1109/ACCESS.2020.3009328.

[2] R. Agrawal, Fundamentals of Machine Learning. New York: Manning Publications Co, 2018. doi: 10.1201/9780429330131-1.

[3] 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, pp. 78–82, 2014, doi: 10.14445/22312803/ijctt-v13p117.

[4] M. Das and R. Dash, “Performance Analysis of Classification Techniques for Car Data Set Analysis,” Proceedings of the 2020 IEEE International Conference on Communication and Signal Processing, ICCSP 2020, pp. 549–553, 2020, doi: 10.1109/ICCSP48568.2020.9182332.

[5] S. Shumaly, P. Neysaryan, and Y. Guo, “Handling Class Imbalance in Customer Churn

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v1i1.154

Prediction in Telecom Sector Using Sampling Techniques, Bagging and Boosting Trees," 2020 10h International Conference on Computer and Knowledge Engineering, ICCKE 2020, pp. 82–87, 2020, doi: 10.1109/ICCKE50421.2020.9303698.

[6] P. Vats and K. Samdani, "Study on machine learning techniques in financial markets," 2019 IEEE International Conference on System, Computation, Automation and Networking, ICSCAN 2019, pp. 1–5, 2019, doi: 10.1109/ICSCAN.2019.8878741.

[7] B. Dai, R. C. Chen, S. Z. Zhu, and W. W. Zhang, "Using random forest algorithm for breast cancer diagnosis," Proceedings - 2018 International Symposium on Computer, Consumer and Control, IS3C 2018, pp. 449–452, 2019, doi: 10.1109/IS3C.2018.00119.

[8] Q. Wu, Y. Ye, H. Zhang, M. K. Ng, and S. S. Ho, "ForesTexter: An efficient random forest algorithm for imbalanced text categorization," Knowledge-Based Systems, vol. 67, pp. 105–116, 2014, doi: 10.1016/j.knosys.2014.06.004.

[9] A. Nugroho, A. Z. Fanani, and G. F. Shidik, "Evaluation of Feature Selection Using Wrapper for Numeric Dataset with Random Forest Algorithm," Proceedings - 2021 International Seminar on Application for Technology of Information and Communication: IT Opportunities and Creativities for Digital Innovation and Communication within Global Pandemic, iSemantic 2021, pp. 179–183, 2021, doi: 10.1109/iSemantic52711.2021.9573249.

[10] Y. Hao and F. Liu, "Application of Fuzzy Equivalence Relation Kernel Clustering Algorithm to Car Evaluation," Proceedings of 2018 IEEE International Conference of Safety Produce Informatization, IICSPI 2018, pp. 591–594, 2019, doi: 10.1109/IICSPI.2018.8690512.

[11] K. M. Kahloot and P. Ekler, "Algorithmic Splitting: A Method for Dataset Preparation," IEEE Access, vol. 9, pp. 125229–125237, 2021, doi: 10.1109/ACCESS.2021.3110745.

[12] E. H. Rachmawanto, D. R. Ignatius Moses Setiadi, N. Rijati, A. Susanto, I. U. Wahyu Mulyono, and H. Rahmalan, "Attribute Selection Analysis for the Random Forest Classification in Unbalanced Diabetes Dataset," in 2021 International Seminar on Application for Technology of Information and Communication (iSemantic), 2021, pp. 82–86. doi: 10.1109/iSemantic52711.2021.9573181.

[13] M. Liang, Z. Chang, Z. Wan, Y. Gan, E. Schlangen, and B. Šavija, "Interpretable Ensemble-Machine-Learning models for predicting creep behavior of concrete," Cement and Concrete Composites, vol. 125, no. October 2021, 2022, doi: 10.1016/j.cemconcomp.2021.104295.

[14] T. Gunasegaran and Y. N. Cheah, "Evolutionary cross validation," ICIT 2017 - 8th International Conference on Information Technology, Proceedings, pp. 89–95, 2017, doi: 10.1109/ICITECH.2017.8079960.

ISSN: 2986-0458 (Print) / 2986-044X (Online)
DOI: https://doi.org/10.53624/jsitik.v1i1.154

[15] A. Nugroho and A. Husin, "Analisis Performa Random Forest Menggunakan Normalisasi Atribut," SISTEMASI: Jurnal Sistem Informasi, vol. 11, no. 1, pp. 186–196, 2022, doi: https://doi.org/10.32520/stmsi.v11i1.1681.