JNATIA Volume 1, Nomor 1, November 2022
Jurnal Nasional Teknologi Informasi dan Aplikasinya

Implementasi Algoritma *Random Forest* Dalam Menentukan Kualitas Susu Sapi

I Putu Ryan Paramaditya[a1], Cokorda Pramartha[b2]

[a]Informatics Department, Udayana University
[b]Net-Centric Computing Laboratory, Udayana University
[1]ryanparamaditya@gmail.com
[2]cokorda@unud.ac.id

Abstract

Milk is one of the food ingredients as a source of animal protein that can meet the needs and improve the nutrition of the people of Indonesia, especially protein, carbohydrates, fats and minerals of high nutritional value, namely calcium and phosphorus which can help the healing process of various diseases. Use the classification method performed with the Random Forest algorithm on the data content of cow's milk content where the results of the experiments that have been carried out, the accuracy rate of classification with the Random Forest algorithm is 98%. In addition, in testing precision, recall, and f1-score generated from each variable has the same value of 98%. then the highest level of accuracy of the response variable is in the "medium" category.

Keywords: *Classification, Random Forest, Accuracy, Confusion Matrix, ROC Curve*

1. Pendahuluan

Susu merupakan salah satu bahan pangan sebagai sumber protein hewani yang dapat memenuhi kebutuhan dan meningkatkan gizi masyarakat Indonesia, termasuk protein, karbohidrat, dan mineral, khususnya kalsium dan fosfor, yang dapat membantu proses pemulihan dari berbagai penyakit.[1] Sapi, kerbau, kuda, dan kambing hanyalah beberapa contoh hewan yang diperah untuk menghasilkan susu, elemen makanan dengan nilai gizi yang tinggi. Dalam mengetahui dan menentukan kualitas dari susu tersebut diperlukan pengujian pada kadar yang dikandung pada susu dari kadar lemak, protein, nilai *pH* hingga suhu yang mempengaruhinya.

Istilah *data mining* sering diartikan sebagai proses untuk menemukan pola-pola yang berwawasan, menarik, dan baru, serta model deskriptif, mudah dipahami dan prediktif dari data skala besar. *Data mining* adalah proses yang mencari tautan dan pola tersembunyi dalam data menggunakan berbagai metodologi dan alat untuk analisis data. Pendekatan dasarnya ialah untuk meringkas data dan ekstrasi informasi berguna yang sebelumnya tidak diketahui. Klasifikasi dan regresi merupakan contoh tugas prediktif, sedangkan *clustering* dan asosiasi adalah contoh tugas deskriptif.[2] Maka dari itu dalam mengetahui proses pola dalam menentukan kualitas susu sapi berdasarkan pada kadar setiap kandungan susu, dilakukan data mining dengan metode klasifikasi.

Menggunakan metode klasifikasi dilakukan, karena pada data tersebut memiliki *class* yang dapat digunakan dalam menentukan pola dan prediksi yang sesuai dengan kategori (*class*) pada data tersebut.[3] Penggunaan algoritma dalam melakukan klasifikasi salah satunya adalah *random forest. Random Forest* merupakan perpanjangan dari pendekatan Pohon Klasifikasi dan Regresi (*CART*) yang menggabungkan agregasi *bootstrap* dan pemilihan fitur acak. Manfaat dari pendekatan ini termasuk peningkatan akurasi, kemampuan untuk menangani sejumlah besar data secara efisien, dan tidak ada pemangkasan variabel seperti pada algoritma pohon klasifikasi tunggal. *Random Forest* menciptakan nilai signifikansi dari faktor prediktor dalam mengkategorikan variabel respons selain memberikan hasil prediksi yang sangat akurat.[4]

Dalam penelitiannya, Widya Apriliah dkk. membandingkan kinerja klasifikasi dengan algoritma *support vector machine, random forest,* dan *naive bayes*. Temuan yang dikumpulkan menunjukkan bahwa *Random Forest* melampaui algoritma lainnya dalam memprediksi kemungkinan diabetes pada tahap awal, dengan skor akurasi terbesar 97,88%.[5] Penggunaan algoritma *random forest* dalam memprediksi harga ponsel yang dilakukan oleh Vanissa Wanika

Siburian, dkk. Dimana tingkat akurasi prediksi yang menggunakan pendekatan *Random Forest* adalah 81% ketika klasifikasi dalam penelitian ini mencakup tujuh variabel prediksi dan satu variabel respon.[6] Maka pada paper ini digunakan algoritma *Random Forest* dalam melakukan metode klasifikasi pada data kualitas kadar pada kandungan susu sapi. Hasil yang diharapkan dengan akurasi yang tinggi dari penelitian yang sudah dilakukan sebelumnya.

2. Metode Penelitian

2.1. Dataset dan Analisis

Metode penelitian yang diterapkan dengan pendekatan kuantitatif. Dimana pada penelitian ini melakukan klasifikasi menggunakan algoritma *Random Forest* dengan data yang memiliki tiga kategori. Pada penelitian tersebut menggunakan data sekunder yang meklasifikasikan kadar pada kandungan susu sapi berdasarkan pada kualitasnya. Data yang didapatkan pada situs kaggle https://www.kaggle.com/datasets/yrohit199/milk-quality dengan data yang digunakan sebanyak 1059 data.

	pH	Temprature	Taste	Odor	Fat	Turbidity	Colour	Grade
0	6.6	35	1	0	1	0	254	high
1	6.6	36	0	1	0	1	253	high
2	8.5	70	1	1	1	1	246	low
3	9.5	34	1	1	0	1	255	low
4	6.6	37	0	0	0	0	255	medium
...	...	...	...	...	...	...	...	...
1054	6.7	45	1	1	0	0	247	medium
1055	6.7	38	1	0	1	0	255	high
1056	3.0	40	1	1	1	1	255	low
1057	6.8	43	1	0	1	0	250	high
1058	8.6	55	0	1	1	1	255	low

1059 rows × 8 columns

Gambar 1. Milk Quality Database

Terdapat tujuh variabel prediksi pada data tersebut ialah *"pH" "Temprature", "Taste", "Odor", "Fat ", "Turbidity",* dan *"Color"*. Semua variabel tersebut memiliki perbedaan nilai. Terdapat nilai *count* (jumlah keseluruhan data dari variabel), *mean* (nilai rata-rata keseluruhan per variabel), *std* (persebaran keseluruhan data dari variabel), *min* (nilai terkecil dari variabel), 25% dari nilai dari per variabel, 50% dari nilai dari per variabel, 75% dari nilai dari per variabel, *max* (nilai terbesar dari variabel. Berikut nilai variabel prediksi pada gambar 2.

	pH	Temprature	Taste	Odor	Fat	Turbidity	Colour
count	1059.000000	1059.000000	1059.000000	1059.000000	1059.000000	1059.000000	1059.000000
mean	6.630123	44.226629	0.546742	0.432483	0.671388	0.491029	251.840415
std	1.399679	10.098364	0.498046	0.495655	0.469930	0.500156	4.307424
min	3.000000	34.000000	0.000000	0.000000	0.000000	0.000000	240.000000
25%	6.500000	38.000000	0.000000	0.000000	0.000000	0.000000	250.000000
50%	6.700000	41.000000	1.000000	0.000000	1.000000	0.000000	255.000000
75%	6.800000	45.000000	1.000000	1.000000	1.000000	1.000000	255.000000
max	9.500000	90.000000	1.000000	1.000000	1.000000	1.000000	255.000000

Gambar 2. Perbedaan Nilai Variabel Prediksi

Pada variabel *"Grade"* sebagai variabel respon yang memiliki tiga kategori, yakni *"high", "low", "medium"*. Kemudian ditransformasi menjadi variabel numerik. Seperti pada tabel 1 di bawah ini.

Tabel 1. Perbedaan Nilai Variabel Prediksi

Kategori	Variabel
high	0
low	1
medium	2

Dalam melakukan analisis klasifikasi data tersebut menggunakan *Google Collab* dengan bahasa pemrograman *Python 3.x* dengan format *ipynb*. Tahapan analisis yang dilakukan yakni:

1. Melakukan eksplorasi data untuk mendapatkan gambaran umum pada data tersebut.
2. Membagi data tersebut menjadi data latih dan data uji.
3. Melakukan transformasi data yang menggunakan data tipe nominal menjadi tipe numerik
4. Melakukan klasifikasi *random forest* pada data latih.
5. Melakukan prediksi kelas berdasarkan kategori variabel respon dengan data uji.
6. Mengevaluasi model klasifikasi dengan menghitung nilai akurasi.

2.2. *Random Forest*

Metode *Random Forest* menggunakan banyak pohon keputusan sebagai bagian dari teknik pembelajaran mesin. Salah satu algoritma pembelajaran mesin terbesar yang digunakan dalam beberapa disiplin ilmu, teknik ini baru-baru ini menunjukkan kemanjurannya dalam masalah regresi dan klasifikasi.[5] Pendekatan ini membagi jaringan menjadi *root node. internal node*, dan *leaf node* dengan memilih kualitas dan data secara acak sesuai dengan hukum yang relevan. Simpul atas pohon keputusan, juga dikenal sebagai akar pohon keputusan, dikenal sebagai simpul akar. *Internal node* adalah node percabangan yang hanya memiliki satu input dan satu atau lebih output. Simpul terakhir, yang dikenal sebagai simpul daun atau simpul terminal, hanya memiliki satu masukan dan tidak ada keluaran. Nilai *entropy* pertama-tama dihitung oleh pohon keputusan untuk menentukan tingkat ketidakmurnian atribut dan pentingnya perolehan informasi. Rumus persamaan 1 digunakan untuk menghitung *entropy*, sedangkan rumus persamaan 2 digunakan untuk menghitung perolehan informasi.[6][7]

$$\boldsymbol{Entropy}\ (\boldsymbol{Y}) = -\sum \mathbf{p}\ (\mathbf{c}|\mathbf{Y})\ \mathbf{i}\ \mathbf{log2}\ \mathbf{p}\ (\mathbf{c}|\mathbf{Y}) \quad (1)$$

Dimana Y merepresentasikan himpunan kasus dan p(c|Y) merepresentasikan proporsi nilai Y yang termasuk dalam kelas c.

$$\boldsymbol{Information\ Gain}\ (\boldsymbol{Y}, \boldsymbol{a}) = \boldsymbol{Entropy}\ (\boldsymbol{Y}) - \sum \boldsymbol{veValues} \frac{|\mathbf{Yv}|}{|\mathbf{Ya}|}\ \boldsymbol{Entropy}\ (\boldsymbol{Yv}) \quad (2)$$

Dimana Nilai (a) mewakili semua nilai yang mungkin dalam himpunan kasus a. Yv adalah *subclass* dari Y, dengan kelas v setara dengan kelas a. Y,a semua nilai yang sesuai dengan a.

2.3. Penerapan *Gain Ratio*

Nilai *information gain* terbesar dari atribut saat ini digunakan untuk memilih atribut sebagai *node*, baik akar (*root*) maupun *internal node*. Nilai gain ratio dihitung dengan membagi information gain dengan *split information*. Dimana *split information* (S, A) adalah estimasi nilai entropi dari variabel input S dengan kelas c dan |Si|/|S| adalah probabilitas kelas i dari atribut tersebut.[6][8] Nilai *split information* dan *gain ratio* dapat dilihat pada persamaan 3 dan 4 sebagai berikut.

$$\boldsymbol{Split\ Information}\ (\boldsymbol{S}, \boldsymbol{A}) = \sum_{i}^{C} \frac{|\mathbf{Si}|}{|\mathbf{S}|}\ \mathbf{log2}\ \frac{|\mathbf{Si}|}{|\mathbf{S}|} \quad (3)$$

$$\boldsymbol{Gain\ Ratio}\ (\boldsymbol{S}, \boldsymbol{A}) = \frac{\mathbf{Information\ Gain}\ (\mathbf{S},\mathbf{A})}{\mathbf{Split\ Information}\ (\mathbf{S},\mathbf{A})} \quad (4)$$

2.4. Confusion Matrix

Kinerja model klasifikasi pada kumpulan data testing dengan nilai sebenarnya yang diketahui sering digambarkan menggunakan *Confusion Multiclass Matrix.* Tabel 2 di bawah ini menunjukkan persamaan *Confusion Multiclass Matrix.*

Tabel 2. *Confusion Multiclass Matrix*

		Predicted		
		A	B	C
Actual	A	TP_A	E_{AB}	E_{AC}
	B	E_{BA}	TP_B	E_{BC}
	C	E_{CA}	E_{CB}	TP_C

Asumsikan pada tabel di atas bahwa terdapat kelas prediksi terdiri dari variabel A, B, dan C. *True Positive* (TP): Ketika kita mengantisipasi ya dan nilai sebenarnya adalah benar. Confusion Multiclass Matrix berkembang dari *Confusion Biner Matrix*, yang sebelumnya menyertakan nilai *False Negative* (FN), *False Positive* (FP), dan *True Negative* (TN). Pada *Confusion Multiclass Matrix,* hanya TP yang disebutkan karna FN ditentukan oleh jumlah baris per variabel, sedangkan FP ditentukan oleh jumlah kolom per variabel, dan TN merupakan situasi ketika tak dapat ditafsir dan nilai aktual yang salah.[6]

2.5. Perhitungan Akurasi dan Presisi

Melakukan perhitungan akurasi sesudah proses klasifikasi telah selesai yang bertujuan untuk menunjukan keakuratan dalam melakukan klasifikasi data terhadap data sebenarnya[6] dengan rumus pada persamaan 5 berikut ini.

$$\boldsymbol{Akurasi} = \frac{\sum \textbf{data uji benar klasifikasi}}{\sum \textbf{jumlah total data uji}} \textbf{ X 100} \quad (5)$$

Setelah menentukan tingkat akurasi, kemudian dilakukan perhitungan nilai presisi. Dimana dalam menentukan nilai presisi, terdapat tp adalah nilai *true positive* dan fp adalah nilai *false negative*. Nilai tp sama untuk data latih (*predictive*) dan data uji (*reference*). Nilai tp + fp merupakan jumlah keseluruhan data uji.[6] Rumus untuk menghitung nilai presisi sebagai berikut pada persamaan 6.

$$\boldsymbol{precision} = \frac{\textbf{tp}}{\textbf{tp+fp}} \textbf{ X 100}\% \quad (6)$$

2.6. ROC (*Receiver Operating Characteristic*)

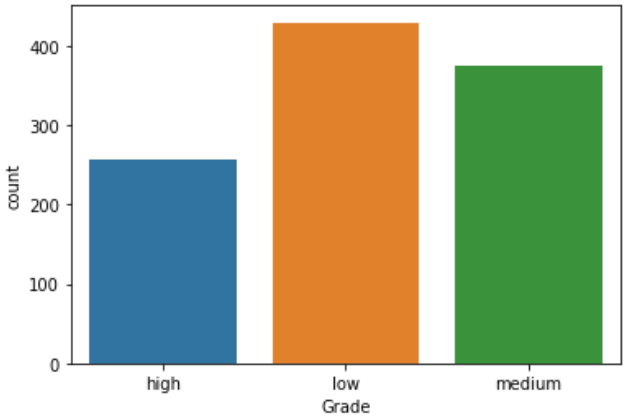
ROC melakukan perbandingan klasifikasi serta menampilkan akurasi secara grafis dengan garis kurva, yang mengekspresikan *Confusion matrix*. ROC merupakan grafik dua dimensi dengan garis horizontal mewakili *false positives* dan *true positive* mewakili garis vertikal.[6][9][10] Angka-angka yang ditunjukkan pada kurva adalah nilai *True Positive Rate* (TPR) dan *False Positive Rate* (FPR), yang dapat dihitung sebagai berikut menggunakan persamaan 7 dan 8 sebagai berikut.

$$\textbf{TPR} = \frac{\textbf{tp}}{\textbf{tp+fp}} \quad (7)$$

$$\textbf{FPR} = \frac{\textbf{fp}}{\textbf{tp+fp}} \quad (8)$$

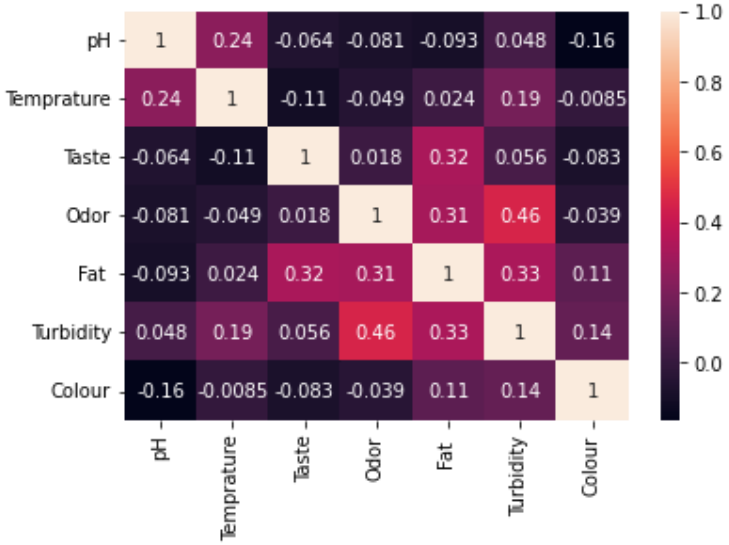
3. Hasil dan Diskusi

Pada pembahasan mengenai hasil dari penelitian tersebut berdasarkan data yang didapatkan sebanyak 1059 data telah diidentifikasi, mendapatkan bahwa variabel respon (*label*) pada kategori *"Grade"* yaitu high berjumlah 256 data, low berjumlah 429 data, dan medium berjumlah 374 data. Berikut hasilnya pada gambar 3.



Gambar 3. Perhitungan jumlah data tiap Grade

Pada nilai korelasi matriks antara *Grade* dengan nilai pengukuran pada kadar susu sapi menunjukan bahwa korelasi pada masing-masing variabel ada yang memiliki nilai tertinggi dan terendah. Pada nilai tertinggi yaitu korelasi antara "*Odor*" dengan "*Turbidity*" sebesar 0.46. sedangkan pada nilai terendah yaitu korelasi antara "*pH*" dengan "*Colour*" sebesar -0.16. Berikut hasil keseluruhan korelasi matriks pada gambar 4.



Gambar 4. Matriks Korelasi Grade dengan Nilai Kadar pada Kandungan Susu

Kemudian pembagian data yang dilakukan dengan perbandingan 70% data latih (*training*) dan juga 30% data uji (*testing*). Selanjutnya dilakukan klasifikasi dengan Algoritma *Random Forest* dengan data yang telah dibagi, maka nilai yang dihasilkan berupa *precision, recall,* dan *f1-score* dari setiap variabel bernilai sama sebesar 98% dengan nilai akurasi juga mencapai 98%. Berikut data nilai yang dihasilkan pada tabel 3.

Tabel 3. Hasil *Accuracy, Precision, Recall, F1-score*

	Precision	***Recall***	***F1-Score***
0	0.97	0.98	0.98
1	1.00	0.97	0.98
2	0.98	1.00	0.99
Avg	0.98	0.98	0.98
Accuracy			0.98

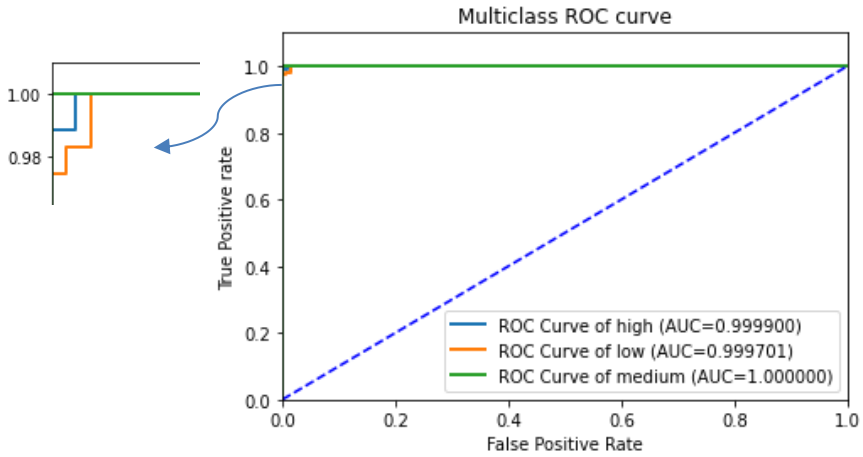
Pada perhitungan akurasi yang lain menggunakan *Confusion Matrix* pada tiga kelas prediksi yaitu variabel 0 (*high*), 1 (*low*) dan 2 (*medium*) dengan membandingkan dari segi *Predicted* dan *Actual*.

Pada variabel 0 sebesar 85, variabel 1 sebesar 113, dan variabel 2 sebesar 116. Berikut hasilnya ditampilkan pada Tabel 4 .

Tabel 4. *Confusion Matrix* data

		Predicted		
		0	1	2
Actual	0	85	0	1
	1	2	113	1
	2	0	0	116

Adapun pengujian yang menggunakan ROC *Curve* untuk menampilkan akurasi secara visual pada setiap kelas kategori dari model yang dihasilkan sangat memuaskan. Pada gambar terdapat sebuah garis putus-putus sebagai acuan (*baseline*) jika semakin besar jarak antara garis-garis tersebut tehadap acuannya maka semakin baik tingkat prediksi. Tiga garis kurva berwarna digunakan untuk mewakili setiap kategori (*class*) sebagai variabel respon. Dimana garis biru menunjukkan kurva ROC *"high"* dengan tingkat prediksi AUC sebesar 0.9999, garis orange menunjukkan kurva ROC *"low"* dengan tingkat prediksi AUC sebesar 0.999701, dan garis hijau menunjukkan kurva ROC *"medium"* dengan tingkat prediksi AUC sebesar 1.00 berarti memiliki tingkat prediksi yang teritinggi. Meskipun dibandingkan dengan *class* lainnya tidak memiliki perbedaan yang signifikan. Berikut hasil ROC pada gambar 5.



Gambar 5. Kurva ROC dari dataset

4. Kesimpulan

Berdasarkan hasil percobaan yang telah dilakukan, maka diperoleh tingkat akurasi pada klasifikasi dengan algoritma *Random Forest* sebesar 98%. Pada pengujian *precision, recall,* dan *f1-score* yang dihasilkan dari setiap variabel bernilai sama sebesar 98%. kemudian tingkat akurasi variabel respon yang tertinggi terdapat pada kategori "medium" dengan AUC sebesar 1.00, meskipun perbedaan hasil akurasi dengan kategori yang lainnya tidak berbeda cukup jauh. Maka dari itu algoritma *Random Forest* sangat cocok untuk digunakan dalam melakukan klasifikasi pada data kualitas susu sapi.

Referensi

[1] E. Situmorang, "Studi Perbandingan Kandungan Kalsium dan Fosfor dalam Susu Kambing Etawa Murni dan Susu Kambing Etawa Kemasan," Thesis, Universitas Sumatera Utara, 2019. Accessed: Oct. 19, 2022. [Online]. Available: https://repositori.usu.ac.id/handle/123456789/24643

[2] A. Wanto *et al.*, *Data Mining : Algoritma dan Implementasi*. Yayasan Kita Menulis, 2020.

[3] A. Damuri, U. Riyanto, H. Rusdianto, and M. Aminudin, "Implementasi Data Mining dengan Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako," *JURIKOM J. Ris. Komput.*, vol. 8, no. 6, Art. no. 6, Dec. 2021, doi: 10.30865/jurikom.v8i6.3655.

[4] A. Ramadhan, B. Susetyo, and Indahwati, “PENERAPAN METODE KLASIFIKASI *RANDOM FOREST* DALAM MENGIDENTIFIKASI FAKTOR PENTING PENILAIAN MUTU PENDIDIKAN,” *J. Pendidik. Dan Kebud.*, vol. 4, no. 2, pp. 169–182, Dec. 2019, doi: 10.24832/jpnk.v4i2.1327.
[5] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, “Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi *Random Forest*,” *SISTEMASI*, vol. 10, no. 1, p. 163, Jan. 2021, doi: 10.32520/stmsi.v10i1.1129.
[6] V. W. Siburian and I. E. Mulyana, “Prediksi Harga Ponsel Menggunakan Metode *Random Forest*,” p. 4, 2018.
[7] N. L. W. S. R. Ginantra *et al.*, *Data Mining dan Penerapan Algoritma*. Yayasan Kita Menulis, 2021.
[8] R. C. Barros, A. C. P. L. F. de Carvalho, and A. A. Freitas, *Automatic Design of Decision-Tree Induction Algorithms*. Springer, 2015.
[9] C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making*. John Wiley & Sons, 2011.
[10] A. Amrin, “Perbandingan Metode Neural Network Model Radial Basis Function Dan Multilayer Perceptron Untuk Analisa Risiko Kredit Mobil,” *Paradigma*, vol. 20, no. 1, Art. no. 1, Apr. 2018, doi: 10.31294/p.v20i1.2783.

halaman ini sengaja dibiarkan kosong