OPTIMASI PARAMETER ALGORITMA DECISION TREE C4.5 PADA KLASIFIKASI BLOGGER PROFESSIONAL

Fida Maisa Hana, Widya Cholid Wahyudin
Universitas Muhammadiyah Kudus
Jln. Ganesha I Purwosari Kudus 59316, Jawa Tengah, Indonesia
fidamaisa@umkudus.ac.id

Abstrak

Blogger merupakan salah satu pekerjaan yang paling dicari. Bisnis yang dapat menghasilkan banyak uang jika ditekuni secara mendalam. Pada saat ini teknologi informasi berkembang dengan cepat, kita bisa menghemat waktu untuk mengklasifikasikan blogger profesional atau bukan menggunakan metode data mining. Untuk memprediksi blogger profesional menggunakan data mining, diperlukan elemen pendukung untuk menentukan dan data yang valid. Dalam menentukan blogger profesional digunakan teknik klasifikasi data mining. Algoritma yang digunakan adalah decision tree C4.5. Penelitian ini Memanfaatkan 4-*Fold Cross Validation* dan Optimasi paramter *apply prepurning, apply purning,* dan *minimal leaf size* pada algoritma C4.5. Hasil akurasi Pengujian didapatkan 88,00 % dengan Precision sebesar 87,30% %, dan Recall sebesar 75,00%.

Kata Kunci: Data Mining, Klasifikasi, C4.5, *Cross Validation* , Blogger

Abstract

Blogger is one of the most sought-after jobs. A business that can make a lot of money if pursued in depth. In the era of computer technology that continues to develop rapidly, we can save time classifying professional bloggers or not using data mining methods. To predict professional bloggers using data mining, supporting elements are needed to determine and valid data. This research uses data mining and classification techniques to determine professional bloggers. The algorithm used is the C4.5 decision tree. This research uses 4-Fold cross-validation and optimizes the apply prepurning, apply purning, and minimum leaf size parameters in the C4.5 algorithm. The test results were 88.00% with Precision of 87.30%, and Recall of 75.00%.

Keywords*: Data Mining, Classification, C4.5, Cross Validation , Blogger*

I. PENDAHULUAN

Salah satu jenis media sosial adalah blog. Blog merupakan layanan internet dan yang menawarkan komponen halaman pembaca gratis kepada pengguna yang memungkinkan orang-orang untuk bergabung dalam komunitas virtual dan menjadi anggota jaringannya (Ardiyansyah, Rahayuningsih and Maulana, 2018). Blog adalah buku harian online yang lahir pada tahun 1997 sebagai situs web dinamis.

Seseorang yang melakukan aktivitas menulis atau blogging menggunakan platform blog tertentu disebut blogger. Di era teknologi saat ini, blogger adalah salah satu pekerjaan yang paling dicari. Bisnis yang dapat menghasilkan banyak uang jika blogger bisa konsisten dan belajar secara mendalam. Domain blog ada yang gratis dan berbayar sebagai lahan untuk blogger menulis pendapat atau menulis cerita yang menarik minat pembaca (Widiastuti, Hermawan and Avianto, 2023). Seorang Blogger yang sukses diwajibkan mempunyai ketrampilan menulis yang baik, sesuai dengan tata bahasa yang benar, jalan cerita, dan kemudahan tu-lisannya untuk dipahami pembaca. Ini tidak harus sama dengan sastrawan atau pujangga, tetapi cukup untuk membuat tu-lisannya dipahami pembaca.

Perkembangan teknologi informasi yang terus berkembang dengan cepat, kita dapat menghemat waktu untuk mengklasifikasikan blogger profesional atau bukan menggunakan metode data mining (Pratama, Prihandono and Ridwan, 2020). Bidang ilmu komputer yang digunakan untuk memprediksikan berbagai aspek kehidupan dan tujuan disebut Data mining. Metode pengolahan data besar yang memiliki manfaat untuk ekstraki informasi bergun dinamakan Data mining. Proses ini dapat dilakukan dengan menggunakan *software* yang menggunakan perhitungan matematika, statistika, atau kecerdasan buatan. Prediksi, deskripsi, klasifikasii, klasterisasi, asosiasi, dan estimasi merupakan teknik-teknik yang umum dipakai untuk pengolahan data mining (Larose and Larose, 2014). Untuk memprediksi bloger profesional menggunakan data mining, diperlukan elemen pendukung untuk menentukan dan data yang valid.

Untuk perhitungan proses klasifikasi, ada beberapa algoritma, seperti k-nearest neighbor (KNN), decission Tree C4.5 dan Naive Bayes (Kurniawan and Ivandari, 2017). Berdasarkan analisis komparasi pembelajaran mesin dalam pengklasifikasian tenaga kerja non-aktif, penelitian (Kusrorong *et al.*, 2019) menemukan bahwa skenario pelatihan pasokan dan cross-validation memiliki akurasi tertinggi di C4.5 dibandingkan dengan K-Nearest Neighbors (KNN) dan Naïve Bayes (NB). Penelitan sebelumnya telah menggunakan algoritma decision tree C4.5 untuk pengklasifikasian seseorang terkena penyakit diabetes menghasilkan akurasi sebesar 97,12 % (Hana, 2020).

C4.5 dikembangkan oleh Ross Quinlan, yaitu algoritma yang termasuk dalam Decision Tree yang dipakai guna membuat sebuah pohon keputusan. C4.5 termasuk pertumbuhan dari algoritma ID3 Quinlan yang sudah ada sebelumnya. Menghasilkan sebuah pohon keputusan yang dapat dipakai untuk proses klasifikasi (Kusrorong *et al.*, 2019).

Riset ini memanfaatkan teknik klasifikasi data mining dalam menentukan bloger profesional. Algoritma yang dipakai adalah decision tree C4.5. Optimasi parameter dilakukan pada algoritma C4.5 agar mendapatkan akurasi yang maksimal.

II. LANDASAN TEORI

1. Data Mining

Teknik yang dipakai untuk mencari informasi yang dari jumlah data yang besar dinamakan Data mining (Dita Merawati, 2019). Istilah "pengolahan data" sudah dikenal sejak tahun 1990 karena mengolah data sangat penting untuk berbagai bidang, seperti pendidikan, industri, dan kesehatan (Gorunescu, 2011). data mining termasuk kombinasi keilmuan penelitian basis data, kecerdasan buatan, dan statistik. (Larose and Larose, 2014) Prediksi, deskripsi, klasifikasii, klasterisasi, asosiasi, dan estimasi adalah enam metode umum yang digunakan dalam data mining untuk menghasilkan hubungan yang memiliki hubungan dan pola dalam pengolahan data besar. (Larose and Larose, 2014).

a. Deskripsi

Teknik deskripsi dipakai untuk menciptakan kriteria dan aturan yang mudah dipahami untuk pola yang tercipta secara berulang pada sekelompok data.

b. Prediksi

Teknik ini mirip seperti klasifikasi, hanya saja data kelas seperti dengan nilai yang diramal di masa depan.

c. Klasifikasi

Metode klasifikasi, juga dikenal sebagai pengelompokan, memungkinkan data diidentifikasi dengan karakteristik tertentu dan kemudian dikelompokkan dalam kelas-kelas tertentu.

d. Estimasi

Tekinik stimasi mendekati algoritma klasifikasi, tetapi target merupakan bilangan numerik atau kontinyu.

e. Klastering

klastering adalah teknik pengelompokan atau klasterisasi data dengan kelompok data yang mengandung karakteristik yang sama di

sebuah himpunan dan yang berbeda di himpunan lain.

f. Asosiasi

Asosiasi merupakan teknik menemukan fitur yang terlihat dalam situasi tertentu atau membuat aturan yang menghubungkan antara kombinasi item.

2. Klasifikasi

Klasifikasi adalah proses menemukan karakteristik suatu objek dan memasukkannya ke suatu kelas yang telah didefinisikan sebelumnya (Larose and Larose, 2014). Menghitung data sebelumnya termasuk proses klasifikasi. Proses ini juga dikenal sebagai pelatihan data dengan data baru atau pengujian data. Proses ini akan membuat pengujian data. Setiap dataset yang dipakai wajib mempunyai label atau fitur yang dimaksudkan untuk diklasifikasikan. Tujuan klasifikasi adalah meramalkan sasaran kelas dalama setiap permasalahan data. Tugas yang memulainya pada satu set data yang memiliki kelas yang dikenal adalah tugas klasifikasi. (Putra and Chan, 2018).

3. *Optimize Paramater*

Nilai parameter yang paling optimal dipilih operator untuk subproses. Operator *optimize parameter* adalah operator yang memiliki sarang. Tahap ini menjalankan subproses pada setiap gabungan nilai parameter yang ditunjuk dan selanjutnya nilai parameter optimal dikirimkan melalui port set parameter. Pengiriman vektor performa untuk nilai parameter optimal melalui port performa, jika ada, model yang terkait dikirim melalui port model. Proses terbaik dikirim melalui port keluaran jika ada Hasil tambahan. Nilai performa yang dikirimkan ke port performa bagian dalam menentukan parameter mana yang paling cocok..

4. Algoritma C4.5

Pada saat melakukan tugas klasifikasi, algoritma C4.5, yang sering dimanfaatkan oleh peneliti. Algoritma ini termasuk perbaikan dari algoritma ID3. Hasil akhir dalam algoritma C4.5 yaitu berupa pohon keputusan.

Tahapan pada algoritma C4.5 untuk menciptakan pohon keputusan adalah seperti dibawah ini:

- penunjukan atribut yang akan dijadikan sebagai akar
- pembuatan cabang pohonpada setiap nilai
- pembagian masalah kedalam cabang-cabang
- pengulangan prosedur dalam setiap cabang sampai semuad masalah pada cabang mempunyai kelas yang sama dalam menunjuk atributt sebagai akar, didasarkann pada nilaii gain terbesar dari atribut-atributt yang tersedia.

$$Gain(S, A) = \text{Entropi}(S) - \sum_{i=1}^{n} - \frac{|s_i|}{|s|} * Entropi\ S_i \quad (1)$$

Penjelasan:
S = himpunan kasus
A = Atribut
N = jumlah partisi atribut A
Si = Jumlah Kasus pada Partisi ke-i

$\left|\frac{\text{Si}}{\text{S}}\right|$ = proporsi Si terhadap S
S = jumlah kasus dalam S‘

Adapunn untuk mencari nilaii Entropy, digunakan rumuss sebagai berikut:

$$\text{Entropy}(S) = \sum_{i=1}^{k} -P_i \ log_2 P_i \qquad (2)$$

Penjelasan:
S = himpunan kasus
k = jumlah partisi S
pi = probabilitas yang didapat dari jumlah (ya/tidak) dibagi total kasus 3.

5. *K-Fold Cross Validation*

Pembagian data menjadi dua bagian, yaitu data proses latih dan data evaluasi. Subset validasi melatih model/algoritma dan subset validasi memvalidasi model/algoritma. Jenis cross-validation dapat dipilih berdasarkan ukuran dataset. Penelitian ini menggunakan 4 Fold, Dalam 4-Fold Cross Validation, pembagian data menjadi 4 fold dengan ukuran yang sama, sehingga mempunyai 4 subset data untuk proses evaluasi model atau

algoritma dalam bekerja. Dalam penelitian ini, hasil klasifikasi blogger profesional dengan algoritma C4.5 akan diuji dengan 4-Fold Cross Validation.

6. *Confussion Matrix*

Hasil dari analisis klasifikasi dalam data miningg yang ditampilkan pada sebuah tabel disebut *Confusion Matrix* (Gorunescu, 2011). Teknik ini sering digunakan untuk menghitung akurasi. Empat istilah dalam pengujian kinerja pada *confusion matrix* yaitu:

1. *False Positive* (FP), adalah data negatif yang diramal sebagai data positif.
2. *False Negative* (FN), adalah data positif yang diramal sebagai data negatif.
3. *True Positive* (TP), adalah dataa positif yang diramal benarr.
4. *True Negative* (TN), adalah dataa negatif yang diramal dengan benar.

Secara umum, bentuk *Confusion Matrix* bisa dipahami pada tabel dibawah ini:

Tabel 1. Tabel *Confussion Matrix* (Gorunescu, 2011)

klasifikasi		**Kelas prediksi**	
		Kelas: ya	**Kelas: tidak**
Kelas observasi	Kelas Ya	**A(Benar Positif)**	**B(Salah Negatif)**
	kelas tidak	**C(Salah Positif)**	**D(Benar Negatif)**

Penghitungan akurasi menggunakan persaman dibawah ini:

$$Akurasi = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \quad (3)$$

Pada saat metode klasifikasi digunakan, mereka selalu memaksimalkan untuk mendapatkan model yang memproduksi akurasi yang tinggi. Kemampuan model ini ditunjukan pada saat model disandingkan pada pengujian data, rata-rata model yang digunakan dapat meramalkan dengan baik pada semua data yang digunakan untuk latih. (Istiawan and Khikmah, 2019).

Rasio antara jumlah prediksi positif yang benar (TP) dan jumlah kasus positif total (TP + FN) pada matriks confusion disebut dengan sensitivitass atau recall. Ini mengukur seberapa baik model kita dapat menemukan semua kasus positif yang ada. Kemampuan pengujiann untuk mengidentifikasi hasil positif dari sejumlah data yang seharusnya positif adalah sensitivitas. Untuk menghitungg sensitivitass atau recall memamkai rumus berikut:

$$Sensitivitas = \frac{TP}{TP+FN} \quad (4)$$

Namun, ketepatan adalah tingkat keakuratan antara data yang diambil dan hasil ramalan model. Ketepatan didefinisikan sebagai rasioo prediksi benar yang positif dibandingkann dengan keseluruhann hasil prediksi yang positif. Berapa banyak data yang benar-benar positif yang telah diprediksi dari semua kelas. Dalam pengolahan data, ketepatan adalah hasil dari membagi jumlah data yang benar-benar positif dengan jumlah data yang diakui sebagai positif. Ketepatan dihitung dengan persamaan berikut.:

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

III. METODE PENELITIAN

Metode eksperimen digunakan dalam penelitian ini, di mana algoritma klasifikasi C4.5 akan dimanfaatkan dalam membuat perhitungan pada dataset..

Menyiapkan dataset adalah tahap pertama yang disiapkan dalam penelitian ini. Penelitian ini menggunakan 4 Fold, Dalam 4-Fold Cross Validation, artinya pembagian data dilakukan dengan 4 fold yang berukuran sama, sehingga menghasilkan 4 subset data sebagai bahan evaluasi kinerja model atau algoritma. Dalam penelitian ini, hasil klasifikasi blogger profesional dengan algoritma C4.5 akan diuji dengan 4-Fold Cross Validation. Proses pengujian dan evaluasi untuk mendapatkan nilai akurasi mengguanakn *confussion*.

A. Sumber Data

Website dengan alamat https://archive.ics.uci.edu/dataset/255/blogger adalah sumber data yang dipakai dalam riset ini. Ini merupakan data repositori pembelajaran mesin UCI yang bisa digunakan secara terbuka untuk penelitian.

Dataset yang dipakai merupakan data Blogger dengan nama file kohkiloyeh.xls. terdapat 6 variabel yang digunakan pada data yang berjumlah 100 sampel. Daftar pertanyaan digunakan untuk mengumpulkan informasi untuk pembentukan database. Selain dikirim secara lisan atau tertulis, kuesioner ini dikirim melalui pemrograman situs web kuesioner yang terhubung ke internet, di mana pengguna dapat menjawab pertanyaan dengan cara apa pun yang mereka inginkan. Data ini dikumpulkan dari Boyer Ahmad dan Provinsi Kohkiloye, yang terletak di Iran.

B. Metode

Metode yang diusulkan pada tahap klasifikasi adalah Decision Tree C4.5 dengan optimasi parameter dalam pembentukan pohon keputusan pada algoritma C4.5. tahap-tahapnya adalah seperti dibawah ini: Pembentukan akar adalah langkah pertama dalam membangun pohon keputusan. Data kemudian dibagi sesuai dengan atribut yang cocok untuk daun.

1. Pemotongan pohon yang sudah dibentuk, yang berarti mengambil cabang yang tidak diperlukan dari pohon yang telah dibentuk dan memangkasnya. Selain mengurangi ukuran pohon, pemecahan pohon juga dilakukan dengan tujuan untuk meminimalkan prediksi salah dalam perkara baru dari hasil pemisahan divide and conquer. Ada dua metode untuk pruning, yaitu :

   a) Pre-pruning, yang berarti penghentian pembentukan suatu subtree lebih awal dengan menghindari pembagian data pelatihan lebih jauh. Saat segera berhenti, node berubah menjadi cabang, atau node akhir. Node terakhir adalah kelas dengan kemunculan di antara subset sampel yang paling sering.

   b) Post-pruning—penyederhanaan pohon setelah pohon selesai dibangun. Leaf (node akhir) dengan kelas yang paling sering nampak adalah node yang jarang dipotong.

Pembentukan aturan keputusan adalah proses mengambil aturan keputusan dari pohon yang telah dibangun sebelumnya. Aturan dalam bentuk if—then dari pohon keputusan bisa didapatkan dengan melakukan penelusuran dari akar sampai ke daun.

IV. HASIL DAN PEMBAHASAN

Tabel 2 menunjukkan variabel data yang dipakai dalam riset ini.

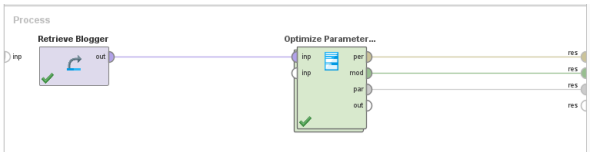
Tabel 2. Variabel data penelitian

No.	Atribut	Value
1	Professional Blogger	Yes & no
2	*Degree*	Low, Medium, & High
3	*Caprice*	Left, right, & middle
4	*Topic*	Political, impression, news, tourism, & scientic
5	*Lmt*	Yes & no
6	*LPSS*	Yes & no

Pada tabel 2, terdapat 6 variabel dataset blogger dengan 1 variabel class penentu klasifikasi yaitu professional blogger.

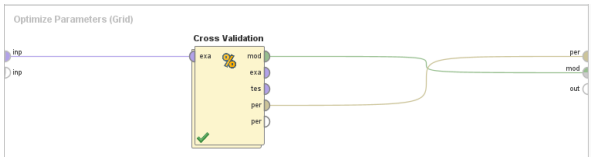
1. Optimasi Parameter

Penerapan dan pengujian dataset pada penelitian ini menggunakan aplikasi rapidminer. Proses Optimasi paramater klasifikasi blogger professional dengan klasifikasi algoritma C4.5 pada aplikasi *RapidMiner* 10.2 adalah sebagai berikut:



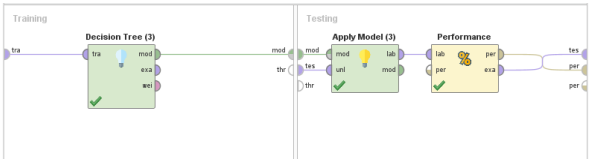
Gambar 1. Proses Optimasi Parameter menggunakan *RapidMiner* 10.2.

Proses pertama pada riset ini adalah menyediakan dataset, setelahnya optimasi parameter *apply prepurning, apply purning,* dan *minimal leaf size* pada algortima C4.5 dilakukan.



Gambar 2. Proses *Cross Validation* menggunakan *RapidMiner* 10.2.

Selanjutnya dilakukan pengujian dengan 4-Fold Cross Validation.



Gambar 3. Proses *Apply Model* Klasifikasi C4.5 dan *Confussion Matrix*

Pada *apply model,* algoritma C4.5 yang dipakai untuk klasifikasi, dihitung akurasinya menggunakan *Confusion Matrix* untuk menghitung akurasinya. *Performance vector* dan *Confusion Matrix dihasilkan* dari pengujian di aplikasi rapid miner.

ParameterSet

```
Parameter set:

Performance:
PerformanceVector [
-----accuracy: 88.00% +/- 0.00% (micro average: 88.00%)
ConfusionMatrix:
True:   yes     no
yes:    64      8
no:     4       24
-----precision: 87.30% +/- 9.26% (micro average: 85.71%) (positive class: no)
ConfusionMatrix:
True:   yes     no
yes:    64      8
no:     4       24
-----recall: 75.00% +/- 10.21% (micro average: 75.00%) (positive class: no)
ConfusionMatrix:
True:   yes     no
yes:    64      8
no:     4       24
-----AUC (optimistic): 0.943 +/- 0.033 (micro average: 0.943) (positive class: no)
-----AUC: 0.874 +/- 0.059 (micro average: 0.874) (positive class: no)
-----AUC (pessimistic): 0.827 +/- 0.085 (micro average: 0.827) (positive class: no)
]
Decision Tree (3).apply_prepruning      = false
Decision Tree (3).apply_pruning = true
Decision Tree (3).minimal_leaf_size     = 11
```

Gambar 4. Hasil *Performance Vector* Klasifikasi C4.5 dengan Optimasi Paramter

Hasil pengujian menghasilkan akurasi, recall, dan precision seperti tabel dibawah ini:

Tabel 3. Hasil Akurasi Klasifikasi C4.5 dengan Optimasi Paramter

Akurasi: 88.00%

	Benar Ya	**Benar Tidak**	***class precision***
pred. ya	64	8	88.89%
pred. tidak	4	24	85.71%
class recall	94.12%	75.00%	

Pengujian menghasilkan akurasi 88,00 % Precision sebesar 87,30% %, dan Recall sebesar 75,00%.

V. KESIMPULAN

Pada data Blogger yang terdiri dari 5 atribut, diantaranya pendidikan (degree), local media turnover (LMT), tingkah politik (caprice), topik, local, political and social space (LPSS) dan penentunya adalah Professional blogger (pb). Data ini dapat dilakukan klasfikasi menggunakan algoritma decision tree C4.5.

Penelitian ini Memanfaatkan 4-*Fold Cross Validation* dan Optimasi paramter *apply prepurning, apply purning,* dan *minimal leaf size* dalam algoritma C4.5

Hasil akurasi Pengujian sebesar 88,00 % dengan Precision sebesar 87,30% %, dan Recall sebesar 75,00%.

DAFTAR PUSTAKA

Ardiyansyah, Rahayuningsih, P. A. and Maulana, R. (2018) 'Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner', *Jurnal Khatulistiwa Informatika*, VI(1), pp. 20–28.

Dita Merawati, R. (2019) 'Penerapan Data Mining Penentu Minat Dan Bakat Siswa Smk Dengan Metode C4 . 5', *Jurnal Algor*, 1(1), pp. 28–37.

Gorunescu, F. (2011) *Data Mining: Concepts, Models, and Techniques*. Springer.

Hana, F. M. (2020) 'Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4 . 5'.

Istiawan, D. and Khikmah, L. (2019) 'Implementation of C4.5 Algorithm for Critical Land Prediction in Agricultural Cultivation Areas in Pemali Jratun Watershed', *Indonesian Journal of Artificial Intelligence and Data Mining*, 2(2), p. 67. doi: 10.24014/ijaidm.v2i2.7569.

Kurniawan, F. and Ivandari (2017) 'Komparasi Algoritma Data Mining Untuk Klasifikasi Penyakit Kanker Payudara', *Jurnal Stmik*, XII(1), pp. 1–8.

Kusrorong, N. S. B. *et al.* (2019) 'Kajian Machine Learning Dengan Komparasi Klasifikasi Prediksi Dataset Tenaga Kerja Non-Aktif', 7(1), pp. 37–49.

Larose, D. T. and Larose, C. D. (2014) *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition*, *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition*. doi: 10.1002/9781118874059.

Pratama, T. G., Prihandono, A. and Ridwan, A. (2020) 'Penerapan Teknik Bagging Untuk Meningkatkan Akurasi Klasifikasi Pada Algoritma C4.5 Dalam Menentukan Blogger Profesional', *Jurnal Bisnis Digitasl dan Sistem Informasi*, 1, pp. 49–55.

Putra, P. P. and Chan, A. S. (2018) 'Pengembangan Aplikasi Perhitungan Prediksi Stock Motor Menggunakan Algoritma C 4.5 Sebagai Bagian dari Sistem Pengambilan Keputusan (Studi Kasus di Saudara Motor)', *INOVTEK Polbeng - Seri Informatika*, 3(1), p. 24. doi: 10.35314/isi.v3i1.296.

Widiastuti, N., Hermawan, A. and Avianto, D. (2023) 'IMPLEMENTASI METODE NAÏVE BAYES UNTUK KLASIFIKASI DATA BLOGGER', 8(3), pp. 985–994. Available at: https://doi.org/10.29100/jipi.v8i3.3713.