JUTIF: Jurnal Teknologi Informasi,
Vol. 1, No. 2, Juli - Desember 2025, Hal. 41- 45

Analisis Perbandingan Kinerja Algoritma Klasifikasi Di Data *Deals* Dengan Menggunakan Metode *K-Fold Cross Validation*

Bi'ad Ardli Abrori[1], Ahmad Marsehan[2]
[1,2]Program Studi Teknologi Informasi, Universitas PGRI Silampari
e-mail: biadardliabrori@gmail.com

Abstract - Pada penelitian ini memiliki tujuan untuk membandingkan kinerja dari dua algoritma klasifikasi yaitu algoritma *Naïve Bayes* dan algoritma *K-Nearest Neigbor* dengan menggunakan metode *K-Fold Cross Validation*. Adapun data yang digunakan pada penelitian ini yaitu data publik *deals* yang mempunyai 1000 data dan mempunyai satu kelas target yaitu *Future Customer*. Data training yang digunakan sebesar 90% atau 900 data dari 1000 data dan data testing yang digunakan sebesar 10% atau 100 data dari 1000 data dan jumlah K pada *K-fold Cross Validation* sebesar 10 atau 10 kali tahap percobaan. Hasil penelitian yaitu diperoleh kinerja algoritma *Naïve Bayes* sebesar 92,60%, *recall* sebesar 98,65% dan presisi sebesar 92,56%. Sedangkan pada algortma K-*Nearest Neighbor* mendapatkan akurasi sebesar 93,80%, *recall* sebesar 94,06% dan presisi sebesar 94,05%.

Keywords: Naïve Bayes, K-Nearest Neighbor, K-Fold Cross Validation, Classification, Data Mining.

PENDAHULUAN

Algoritma klasifikasi merupakan salah satu bagian penting dalam data mining yang bertujuan untuk mengklasifikasikan data ke dalam kategori-kategori yang telah ditentukan. Penerapan algoritma klasifikasi pada data *deals* memegang peranan penting dalam menganalisis pola-pola transaksi untuk berbagai tujuan. Salah satu metode evaluasi yang umum digunakan dalam mengevaluasi kinerja algoritma klasifikasi adalah metode *K-Fold Cross Validation.*

Metode *K-Fold Cross Validation* merupakan teknik validasi model yang penting dalam pengembangan algoritma klasifikasi. Dengan metode ini, data dibagi menjadi k subset, di mana tiap subset digunakan sebagai data uji satu kali dan subset lainnya menjadi data latih. Hal ini dilakukan secara bergantian sehingga setiap subset berperan sebagai data uji sekaligus data latih. Penggunaan *K-Fold Cross Validation* memungkinkan evaluasi kinerja algoritma klasifikasi menjadi lebih dapat diandalkan dan konsisten.

Penelitian ini bertujuan untuk melakukan analisis perbandingan kinerja beberapa algoritma klasifikasi pada data *deals* dengan menggunakan metode *K-Fold Cross Validation*. Beberapa algoritma klasifikasi yang akan dievaluasi antara lain *Naïve Bayes* dan *K-Nearest Neighbor*. Data *deals* yang digunakan dalam penelitian ini merupakan data transaksi untuk tujuan analisis pasar dan pemasaran.

Dalam penelitian ini, dilakukan eksperimen dengan mengimplementasikan berbagai algoritma klasifikasi pada data *deals* dan mengukur kinerja masing-masing algoritma menggunakan metode *K-Fold* Cross Validation. Hasil analisis akan menunjukkan perbandingan kinerja antara algoritma klasifikasi yang dievaluasi dan memberikan pemahaman yang lebih mendalam terkait keunggulan dan kelemahan masing-masing algoritma dalam mengklasifikasikan data *deals*.

Penelitian ini bertujuan untuk melakukan analisis perbandingan kinerja klasifikasi di data *deals* dengan menggunakan metode K-Fold Cross Validation. Klasifikasi merupakan salah satu teknik dalam data mining yang digunakan untuk mengelompokkan data ke dalam kategori-kategori yang telah ditentukan. Metode K-Fold Cross Validation sendiri merupakan metode validasi yang umum digunakan dalam pengembangan model klasifikasi untuk menghindari overfitting.

Dalam kondisi ideal, penggunaan metode K-Fold Cross Validation diharapkan dapat meningkatkan akurasi klasifikasi data *deals*. Namun, dalam kondisi faktual, masih terdapat beberapa kendala yang perlu diatasi. Salah satunya adalah ketidakpastian dalam pemilihan parameter yang optimal untuk model klasifikasi. Hal ini dapat mempengaruhi kinerja klasifikasi secara keseluruhan.

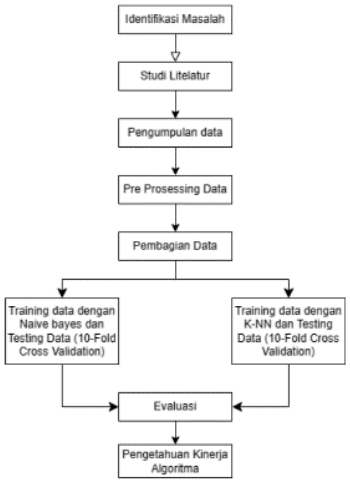
Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi dalam pemahaman dan penerapan algoritma klasifikasi pada data deals serta pentingnya penggunaan metode *K-Fold Cross Validation* dalam evaluasi kinerja algoritma klasifikasi. Hasil dari penelitian ini diharapkan dapat memberikan panduan yang berguna bagi para peneliti dan praktisi dalam bidang data mining untuk memilih algoritma klasifikasi yang optimal dalam menganalisis data *deals*.

METODOLOGI PENELITIAN

1. Tahapan Penelitian

Penelitian ini dilakukan untuk mengetahui tingkat kinerja dari algoritma *K-Nearest Neighbor* dan

algoritma *Naïve Bayes* dengan menggunakan dataset publik *deals* kemudian memilih atribut yang akan digunakan, menentukan variabel target (label) klasifikasi, mengkategorisasikan data-data sebanyak 2 kategori yaitu iya dan tidak, kemudian menggunakan metode *K-Fold* cross validation, dan terakhir membandingkan tingkat kinerja kedua algoritma tersebut. Adapun langkah-langkah yang dilakukan dalam penelitian ini ditunjukan pada Gambar 1. Langkah-langkah penelitian.



Gambar 1 : Langkah-Langkah Penelitian

2. Alur Klasifikasi

Alur klasifikasi pada penelitian ini dimulai dari tahap seleksi data, *preprocessing data*, transformasi data, data mining, dan evaluasi. Proses-proses tersebut dapat dijelaskan sebagai berikut:

a. Seleksi data
   Proses seleksi terhadap data yang akan digunakan untuk data mining yang dalam penelitian ini yaitu menggunakan algoritma klasifikasi.
b. *Preprocessing data*
   Dalam penelitian ini yaitu melakukan pemilihan atribut yang akan digunakan seperti variabel *Age, Gender, Payment Metod* dan *Future Custumer* sebagai label target.
c. *Transformasi data*
   Dalam penelitian ini menyalin data yang telah dipilih sebelumnya ke dalam format excel yang akan digunakan untuk proses data mining menggunakan aplikasi Rapidminer.
d. Data mining
   Pada penelitian ini yaitu menerapkan model algoritma *Naïve Bayes* dan algoritma *K-Nearest Neighbor* untuk melakukan klasifikasi terhadap data yang menjadi bagian dari data training pada setiap tahap sesuai dengan aturan *K-Fold* cross validation.
e. Evaluasi
   Pada penelitian ini yaitu melakukan pemeriksaan hasil klasifikasi dari model algoritma data mining sesuai dengan aturan *K-Fold* cross validation.

3. Model Algoritma Naïve Bayes

Klasifikasi merupakan proses untuk menentukan sebuah kategori dari sekumpulan objek yang kategorinya tidak diketahui. Pengkategorisasian teks menjadi suatu hal yang penting dan kebutuhannya yang semakin meningkat seiring berjalannya waktu, karena data semakin lama akan semakin bertambah. Sehingga perlu digunakan metode untuk mengklasifikasi data uji untuk menghasilkan kategori yang sesuai(Hardianti et al., 2018). Adapun rumus perhitungan pada algoritma *Naïve Bayes* adalah sebagai berikut :

$$P\ (h \mid D) = \frac{P(\ h \mid D\ )P(h)}{P(D)}$$

Keterangan :

h : Hipotesis data dengan suatu class tertentu.
D : Data yang belum memiliki class.
P(*h*) : Probabilitas hipotesis.
P(*D*) : Probabilitas D.
P(*h* | *D*) : Probabilitas *h* berdasarkan kondisi *D.*
P(*D* | *h*) : Probabilitas *D* berdasarkan kondisi *h.*

4. Model Algoritma *K-NEAREST NEIGHBOR*

Algoritma *K-Nearest Neighbor* adalah algoritma klasifikasi atau regresi nonparametrik dalam bidang pengenalan pola. Pada sub bagian ini, kami secara singkat memperkenalkan algoritma klasifikasi *K-Nearest Neighbor*. Asumsikan ada beberapa data pelatihan yang memiliki beberapa atribut dan label. Selanjutnya ada kelompok data testing yang hanya memiliki beberapa atribut saja tanpa label (Purwanto et al., 2023). Adapun rumus perhitungan pada algoritma *K-Nearest Neighbor* adalah sebagai berikut:

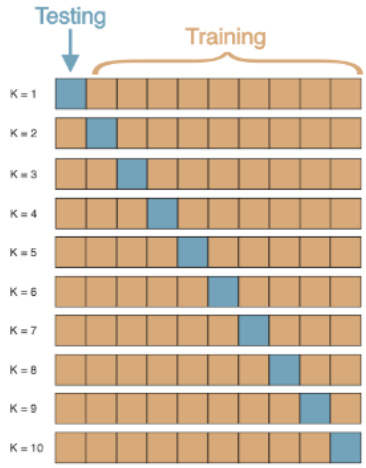
$$Similarity\ (T,S) = \frac{\sum_{i=1}^{n} f\ (Ti, Fi)\ wi}{wi}$$

Nilai kedekatan berada diantara 0 sampai 1. Nilai 0 artinya kedua kasus tidak memiliki similarity, sedangkan nilai 1 artinya kedua kasus memiliki similarity mutlak.

5. Alur Validasi Pengujian

Metode validasi pengujian yang digunakan yaitu metode *Cross Validation* yaitu metode validasi model untuk menilai dan mengetahui berapa hasil statistik analisis yang akan menggeneralisasi himpunan data independe (Tuntun et al., 2022) n. Dalam penelitian ini, data yang digunakan dibagi menjadi beberapa subset data. Pada tahap pertama, subset data pertama menjadi data uji dan subset data lainnya menjadi data pelatihan. Kemudian pada tahap kedua, subset kedua menjadi data uji dan subset data lainnya menjadi data latih. Proses pada tahap selanjutnya sama dengan

penggunaan subset data sesuai dengan tahap yang dilakukan. Ilustrasi pembagian subset data dapat dilihat pada Gambar 2.



Gambar 2 : Alur Validasi Pengujian

HASIL DAN PEMBAHASAN

Pada bagian ini, dijelaskan hasil penelitian dan pada saat yang sama diberikan pembahasan yang komprehensif. Hasil dapat disajikan dalam angka, grafik, tabel, dan lain-lain yang membuat pembaca memahami dengan mudah. Pada bagian ini ditekankan nilai baru dari penelitian yang memuat inovasi, serta implikasinya. Pembahasan dapat dibuat dalam beberapa sub-bab.

1. Pengumpulan Data

Data yang digunakan adalah dataset yang bersifat publik yaitu dataset *deals*. Dataset *deals* adalah dataset yang berisi 2 jenis future customer beserta usia, jenis kelamin dan metode pembayaran. Data ini tersedia dalam bentuk CSV pada aplikasi RapidMiner dan dalam format excel yang digunakan sebagai tabel. setiap baris menunjukkan jenis customer yang berbeda, sedangkan kolom menunjukkan fitur data, yaitu: usia customer, jenis kelamin customer dan metode pembayaran. Ada 2 jenis customer yaitu future customer Yes dan future customer No.

Tabel 1. Dataset *Deals*

No	Row No	Future Customer	Age	Gender	Payment Method
1	1	yes	64	Male	Credit card
2	2	yes	35	Male	Cheque
3	3	yes	25	Female	Credit card
4	4	no	39	Female	Credit card
5	5	yes	39	Male	Credit card
6	6	no	28	Female	Cheque
7	7	yes	21	Female	Credit card

2. Seleksi Data

Pada tahap ini dilakukan proses pemilihan variabel atribut yang akan digunakan dalam proses data mining menggunakan algoritma *K-Nearest Neighbor* dan algoritma *Naïve Bayes*. Variabel atribut yang digunakan adalah variabel *Age*, variabel *Gender*, variabel *Payment Method*, dan *Future Customer* yang menjadi label atau kelas target dari klasifikasi yang terdiri dari 2 target kelas No dan Yes. Bentuk data hasil seleksi dapat dilihat pada Tabel 2. Hasil seleksi.

Tabel 2. Hasil Seleksi

No	Future Customer	Age	Gender	Payment Method
1	Yes	64	Male	Credit card
2	Yes	35	Male	Cheque
3	Yes	25	Female	Credit card
4	No	39	Female	Credit card
5	Yes	39	Male	Credit card
6	No	28	Female	Cheque
7	Yes	21	Female	Credit card

3. Pre-prosessing data

3.1. Pembersihan Data (Data Cleaning)

Setelah proses pengumpulan dan penyaringan data selesai, langkah berikutnya adalah membersihkan data untuk menghilangkan duplikasi, mendeteksi ketidaksesuaian data, dan memperbaiki kesalahan seperti kesalahan penulisan, sehingga data dapat diproses menggunakan teknik data mining. Data cleaning adalah sebuah konsep yang sering dijumpai dalam bidang data science. Menurut (Setiawan et al., 2021) data cleaning merujuk pada upaya meningkatkan kualitas data yang digunakan. Seorang ahli data sebelum memilih model akan menghabiskan setengah dari waktunya untuk membersihkan data, karena kualitas data berperan penting dalam hasil analisis yang dihasilkan. Data cleaning sendiri merupakan suatu prosedur yang bertujuan untuk memastikan kualitas dari suatu set data.Setelah semua data yang diperlukan melalui proses pembersihan, data akan disimpan ke dalam dataset baru menggunakan *Microsoft Office Excel* dengan format csv.

3.2. Integrasi dan Transformasi Data (*Data Integration and Transformation*)

Menurut (Giordano et al., 2011 merupakan sebuah rangkaian langkah, metode, dan teknologi yang dipergunakan untuk menciptakan dan mengembangkan tahapan yang mengekstrak, mentransformasi, mengolah, dan memuat data secara praktis, serta melakukan analisis penyimpanan data dalam waktu nyata atau dalam mode batch. Transformasi data dapat dijelaskan sebagai proses di mana data awal yang berupa data mentah hasil

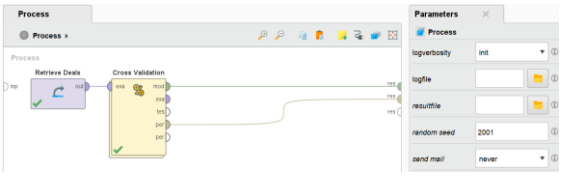
ekstraksi kemudian disaring dan diubah sesuai dengan aturan bisnis yang berlaku (Dwi Handoko et al., 2022). Beberapa teknik data mining memerlukan format data tertentu sebelum dapat diaplikasikan. Pada penelitian ini, data yang akan diproses dari aplikasi *Excel* akan diubah menjadi file CSV (comma delimited) yang dapat digunakan untuk pengolahan data di Software *RapidMiner*.

3.3. Pengrangan Data (*Data Reduction*)

Data yang dikumpulkan dari lapangan memiliki banyak atribut, oleh karena itu perlu dicatat dengan teliti dan rinci serta segera dilakukan analisis data melalui reduksi data. Mereduksi data berarti merangkum, memilih hal-hal yang pokok, memfokuskan pada hal-hal yang penting, dicari tema dan polanya (Azhari et al., 2021). Dengan demikian, pengurangan data akan membantu peneliti dalam memperoleh gambaran yang lebih jelas. Hal ini akan memudahkan dalam pengumpulan data selanjutnya dan pencariannya jika dibutuhkan.

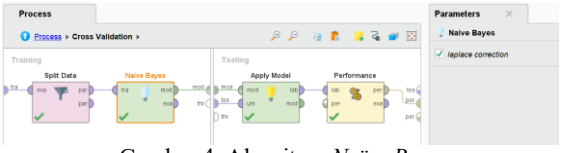
4. Algoritma *Naïve Bayes*

Data yang telah selesai dipersiapkan selanjutnya akan dilakukan mining menggunakan algoritma data mining. Yang pertama menggunakan algoritma *Naïve Bayes*. *Naïve Bayes* merupakan metode yang dapat digunakan untuk memperkirakan probabilitas kemunculan suatu kelas (Novianto et al., 2023). Secara sederhana, dalam pengklasifikasi *Naïve Bayes* diasumsikan bahwa keberadaan suatu fitur dalam kelas tidak berkaitan dengan keberadaan fitur lainnya. Model *Naïve Bayes* mudah dibuat dan sangat berguna untuk kumpulan data yang besar. Meskipun sederhana, *Naïve Bayes* terbukti unggul bahkan dibandingkan dengan metode klasifikasi yang lebih canggih. Berikut ini adalah implementasi algoritma *Naïve Bayes* menggunakan aplikasi *RapidMiner* yang dapat dilihat pada gambar 3.



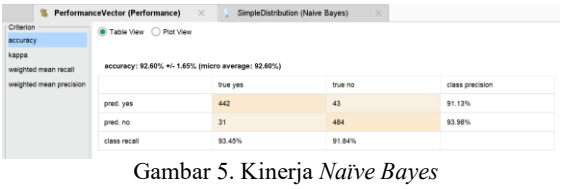
Gambar 3. Data Ke *Cross Validation*

Selanjutnya, dalam operator *Cross Validation* akan ditambahkan operator *Split* Data guna membagi partisi data menjadi data training dan data testing. Pembagian data dilakukan dengan rasio 90% untuk data training dan 10% untuk data testing. Tahapan berikutnya adalah menghubungkan model *Naïve Bayes* dengan operator *Apply Model*, kemudian dari Apply Model dihubungkan ke operator Performance untuk mengevaluasi kinerja algoritma *Naïve Bayes* pada data *deals* menggunakan metode *K-Fold Cross Validation* dengan nilai 'k' sebesar 10. Kinerja algoritma dievaluasi berdasarkan tingkat akurasi, *classification error*, *recall*, dan presisi. Detail proses langkah ini dalam aplikasi Rapidminer dapat dilihat pada Gambar 4 Algoritma *Naïve Bayes* .



Gambar 4. Algoritma *Naïve Bayes*

Setelah proses perancangan proyek selesai, langkah berikutnya adalah menjalankan desain proyek yang telah dibuat. Evaluasi tingkat kinerja bisa ditemukan pada Gambar 5, yaitu evaluasi kinerja *Naïve Bayes* .

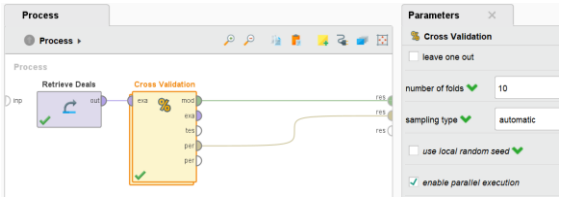


Gambar 5. Kinerja *Naïve Bayes*

Berdasarkan data yang tertera pada Gambar 5, dapat disimpulkan bahwa algoritma *Naïve Bayes* memiliki tingkat kinerja yang tinggi pada dataset *deals* dengan metode *Cross Validation* dan jumlah 'k' sebesar 10. Hasilnya menunjukkan akurasi sebesar 92,60%, *recall* sebesar 98,65%, dan presisi sebesar 92,56%.

5. Algoritma *K-Nearest Neighbor*

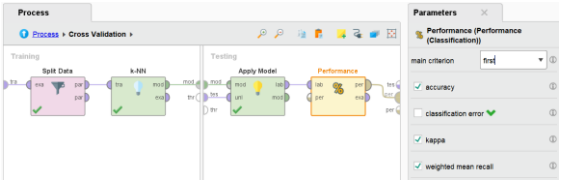
Selanjutnya data publik *deals* akan digunakan untuk proses data mining dengan menggunakan Algoritma *K-Nearestt Neighbor*. Algoritma *K-Nearest Neighbor* merupakan metode klasifikasi penambangan data, digunakan untuk membuat prediksi masa depan (Lusiandro et al., 2020). Data disimpan dalam format file *Microsoft Excel*, lalu diimpor ke aplikasi Rapidminer. Setelah itu, data yang telah diimpor akan disambungkan dengan operator *Cross Validation* menggunakan 10 k-fold. Proses ini dapat dilihat pada Gambar 6 di aplikasi *Rapidminer*. Data akan melalui proses *cross validation*.



Gambar 6. Data Ke *Cross Validation*

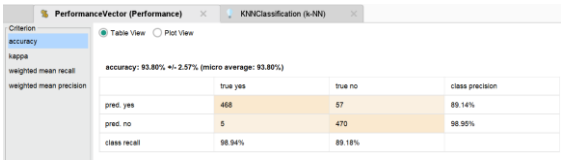
Selanjutnya, dalam proses *Cross Validation* akan dimasukkan operator Split Data. Operator ini bertujuan untuk membagi data menjadi data training (90%) dan data testing (10%). Setelah itu, operator Split Data dihubungkan ke model *K-Nearestt Neighbor* dengan nilai k=10. Tahap berikutnya adalah menghubungkan model *K-Nearestt Neighbor* ke operator *Apply Model*, kemudian dari *Apply Model* dihubungkan ke operator *Performance* untuk mengevaluasi kinerja algoritma *K-Nearest Neighbor* pada data *deals* menggunakan metode *K-Fold Cross Validation* dengan k=10. Penilaian kinerja algoritma

fokus pada akurasi, *classification error, recall*, dan presisi. Proses ini dapat dilihat pada Gambar 7 dalam aplikasi *Rapidminer*.



Gambar 7. Algoritma *K-Nearest Neighbor*

Setelah desain proyek telah tersusun, langkah selanjutnya adalah mengimplementasikan desain proyek tersebut. Hasil kinerja dapat dilihat pada Gambar 8. Kinerja *K-Nearest Neighbor*.



Gambar 8. Kinerja *K-Nearest Neighbor*

Berdasarkan data pada Gambar 8, terlihat bahwa algoritma *K-Nearest Neighbor* menunjukkan tingkat kinerja yang baik pada dataset *deals*. Algoritma ini dievaluasi dengan metode *Cross Validation* menggunakan 10 fold, menghasilkan akurasi sebesar 93,80%, *recall* sebesar 94,06%, dan presisi sebesar 94,05%.

Tabel 3. Hasil Perbandingan

No	Algoritama	Kinerja		
		Akurasi	Recall	Presisi
1	*Naïve Bayes*	92,60%	98,65%	92,56%
2	*K-Nearest Neighbor*	93,80%	94,06%	94,05%

KESIMPULAN

Berdasarkan penelitian yang didokumentasikan dalam bagian Hasil dan Pembahasan, dapat disimpulkan bahwa dalam uji dan validasi tingkat kinerja algoritma data mining pada data publik *deals* menggunakan metode *K-Fold Cross Validation* dengan jumlah 'k' sebanyak 10, Algoritma *K-Nearest Neighbor* menunjukkan tingkat akurasi lebih tinggi yaitu 93,80% dibandingkan dengan *Naïve Bayes* yang mencapai 92,60%. Adapun tingkat *recall* lebih diunggulkan oleh *Naïve Bayes* dengan persentase 98,65% dibandingkan dengan *K-Nearest Neighbor* yang mencapai 94,06%. Sedangkan untuk tingkat presisi, *Naïve Bayes* memiliki tingkat sebesar 92,56%, sedangkan *K-Nearest Neighbor* memiliki tingkat presisi sebesar 94,05%.

REFERENSI

Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(2), 640. https://doi.org/10.30865/mib.v5i2.2937

Dwi Handoko, F., Fauzi, A., Ryan, D., Kurniasih, F., Mutiara, P., Taqwaning Afifi, S., & Author, C. (n.d.). *TRANSFORMASI DATA MENJADI INFORMASI PADA BISNIS INTELIJEN*. *2*(3), 2022. https://doi.org/10.38035/jihhp.v2i3

Giordano AL, Data Integration : Blueprint and Modeling Techniques for a Scalable and Sustainable Architecture, IBM Press, Boston, 2011

Hardianti, A. T., Manga, A. R., & Darwis, H. (2018). Penerapan Metode Naïve Bayes pada Klasifikasi Judul Jurnal. *Prosiding Seminar Nasional Ilmu Komputer Dan Teknologi Informasi*, *3*(2).

Lusiandro, M. A., Nasution, S. M., & Setianingsih, C. (2020, October). Implementation of the advanced traffic management system using k-nearest neighbor algorithm. In 2020 International Conference on Information Technology Systems and Innovation (ICITSI) (pp. 149-154). IEEE.

Novianto, E., Hermawan, A., & Avianto, D. (2023). KLASIFIKASI ALGORITMA K-NEAREST NEIGHBOR, NAIVE BAYES, DECISION TREE UNTUK PREDIKSI STATUS KELULUSAN MAHASISWA S1. *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, *8*(2), 146–154. https://doi.org/10.36341/rabit.v8i2.3434

Purwanto, A., Widi Nugroho, H., Ilmu Komputer, F., Darmajaya, I., Pagar Alam No, J. Z., Meneng, G., Rajabasa, K., & Bandar Lampung, K. (2023). *ANALISA PERBANDINGAN KINERJA ALGORITMA C4.5 DAN ALGORITMA K-NEAREST NEIGHBORS UNTUK KLASIFIKASI PENERIMA BEASISWA* (Vol. 17, Issue 1). https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Setiawan, I. (2021). PERBEDAAN DATA ENGINEER, DATA SCIENTIST DAN DATA ANALYST. *Jurnal Kajian Pendidikan FKIP Universitas Dwijendra*, *12*(2). http://ejournal.undwi.ac.id/index.php/widyaaccarya/index

Tuntun, R., Kusrini, K., & Kusnawi, K. (2022). Analisis Perbandingan Kinerja Algoritma Klasifikasi dengan Menggunakan Metode K-Fold Cross Validation. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *6*(4), 2111. https://doi.org/10.30865/mib.v6i4.4681