Jurnal Ilmiah Sinus (JIS) Vol : 22, No. 2, Juli 2024
ISSN (Print) : 1693-1173 , ISSN (Online): 2548-4028

Comparison of Sentiment Analysis from Twitter Data Collection with Naïve Bayes, Decision Tree, and k-Nearest Neighbor Methods

Erwin Apriliyanto[1)], Yayu Sri Rahayu[2)]
[1)] Teknik Komputer, Universitas Muhammadiyah Karanganyar
[2)] Teknik Komputer, Universitas Muhammadiyah Karanganyar
[1)] itanalisterwin@gmail.com 1, [2)] rahayupink2024@gmail.com 2

ABSTRACT

Dalam konteks perkembangan pesat pengguna media sosial di Indonesia, terutama di Twitter, muncul permasalahan dalam mengelola dan menganalisis volume besar data yang dihasilkan. Dengan jumlah pengguna yang terus meningkat, terutama dalam konteks penggunaan bahasa Indonesia, penting untuk dapat memahami dan mengklasifikasikan konten yang diposting ke dalam kategori yang relevan, seperti positif, negatif, atau netral. Permasalahan utama yang dihadapi adalah bagaimana mengklasifikasikan tweet dalam bahasa Indonesia secara otomatis sehingga dapat memberikan informasi yang berguna untuk berbagai keperluan, termasuk analisis sentimen, pemantauan opini publik, dan pengambilan keputusan yang berdasarkan data. Data yang dihasilkan memberikan informasi berharga untuk penelitian dan pengambilan keputusan. Penelitian ini bertujuan untuk mengklasifikasikan tweet berbahasa Indonesia ke dalam kategori positif, negatif, dan netral. Hasil pengujian menunjukkan bahwa metode Decision Tree memiliki rata-rata presisi kelas yang lebih tinggi dibandingkan dengan K-nearest neighbours (K-NN) dan Naïve Bayes. Algoritma Decision Tree mencapai presisi kelas sebesar 72.85%, sedangkan Algoritma K-NN 54.60%, dan Naïve Bayes 47.66%. Selain itu, penggunaan Decision Tree menghasilkan presisi yang tinggi untuk kelas Negatif (90,00%) dan Positif (81,82%).
Kata kunai: Sentiment Analysis, Twitter, Naïve Bayes, Decision Tree, k-Nearest Neighbor

I. PENDAHULUAN

Jejaring sosial telah merevolusi cara orang berkomunikasi. Informasi yang tersedia dari jejaring sosial berguna untuk menganalisis opini pengguna, misalnya mengukur umpan balik terhadap produk yang baru dirilis, melihat respons terhadap perubahan kebijakan, atau menikmati acara yang sedang berlangsung. Memfilter data ini secara manual bisa membosankan dan berpotensi mahal. Analisis sentimen adalah bidang yang relatif baru yang berhubungan dengan ekstraksi opini otomatis dari pengguna. Contoh sentimen positif adalah, “pemrosesan bahasa alami itu menyenangkan”, sementara sentimen negatif adalah “hari ini buruk, saya tidak keluar”. Teks objektif dianggap tidak mengungkapkan sentimen apa pun, misalnya headline seperti “rencana rak perusahaan sektor angin”. Ada banyak cara di mana data jejaring sosial dapat dimanfaatkan untuk memberikan pemahaman yang lebih baik tentang opini pengguna. Isu ini adalah salah satu yang dapat dimanfaatkan dengan menggunakan klasifikasi berdasarkan clastering data pengguna (Apriliyanto et al., 2020) (Khoirunisa et al., 2020) dan bisa menjadi topik penelitian data mining yang menarik.

Dalam tulisan ini, kami menyajikan analisis sentimen yang mampu menganalisis data Twitter menggunakan tiga metode: *Naïve Bayes*, *Decision Tree*, dan *K-Nearest Neighbor*. Tulisan ini menunjukkan cara mengumpulkan data sentimen secara otomatis untuk keperluan analisis dan pengumpulan opini. Dengan menggunakan data ini, pembaruan penelitian yang akan dilakukan menggunakan pengklasifikasi sentimen yang mampu menentukan sentimen positif, negatif, dan netral atau objektif terhadap data yang diolah.

Dalam penelitian yang dilakukan oleh (Ramadhan et al., 2020), akurasi KNN pada data yang telah diklasifikasikan sebelumnya adalah 98,94% dengan waktu pemrosesan sekitar 24

detik. Sementara itu, akurasi Decision Tree pada data yang sama mencapai 99,91% dengan waktu pemrosesan sekitar satu detik. Dari penelitian ini, beberapa kesimpulan dapat diambil. Pertama, algoritma Decision Tree memiliki akurasi yang lebih tinggi dalam mendeteksi serangan DDoS dibandingkan algoritma KNN. Kedua, waktu pemrosesan Decision Tree lebih efisien dibandingkan KNN karena durasinya lebih singkat.

II. TINJAUAN PUSTAKA

Menurut Tella, penelitiannya tentang pemodelan spasial hotspot PM10 menggunakan algoritma Naïve Bayes (NB), Random Forest (RF), dan K-Nearest Neighbor (KNN) menghasilkan kinerja yang baik. Evaluasi model untuk KNN, RF, dan NB menunjukkan spesifisitas masing-masing sebesar 98%, 99%, dan 92%; presisi sebesar 98%, 99%, dan 92%; recall sebesar 94%, 98%, dan 91%; serta akurasi keseluruhan sebesar 96%, 98%, dan 91% (Tella et al., 2021).

Kinerja setiap model bervariasi berdasarkan ukuran dan karakteristik dataset. Meskipun terdapat sedikit perbedaan dalam pengukuran kinerja antar algoritma, SVM adalah pengklasifikasi terbaik kedua, diikuti oleh K-Nearest Neighbor dan Decision Tree, ketika diterapkan pada setiap dataset (Sheth et al., 2022).

Dalam penelitian yang dilakukan oleh Akshay Gole, perbandingan antara tiga metode yaitu Naive Bayes, Decision Tree, dan K-Nearest Neighbor menunjukkan bahwa metode Naive Bayes merupakan pilihan terbaik. Hal ini didasarkan pada faktor-faktor seperti kecepatan pembelajaran, kecepatan klasifikasi, kinerja saat menghadapi data yang hilang, dan performa ketika berhadapan dengan fitur yang tidak relevan. (Akshay Gole, dkk, 2022).

Penelitian ini bertujuan untuk membandingkan tiga model algoritma klasifikasi, yaitu Naive Bayes, regresi logistik, dan K-Nearest Neighbor (KNN). Hasil penelitian menunjukkan bahwa ketiga model tersebut dapat digunakan untuk memprediksi tingkat kesembuhan pasien Covid-19 di Indonesia. Namun, model (kNN) terbukti paling unggul dengan akurasi tertinggi sebesar 0,750, dibandingkan dengan regresi logistik dan Naive Bayes yang keduanya memiliki akurasi sebesar 0,703. Variabel yang mempengaruhi tingkat kesembuhan pasien Covid-19 meliputi usia, jenis kelamin, dan provinsi tempat tinggal pasien. Penelitian ini menggunakan variabel-variabel yang relatif sederhana, namun masih bisa ditingkatkan dengan menambahkan variabel lain seperti penyakit penyerta, pola makan, riwayat perjalanan, dan variabel relevan lainnya. (Romadhon & Kurniawan, 2021)

Dalam penelitian lain yang dilakukan oleh (Sianturi dan Yuhana 2022), tiga algoritma klasifikasi yang digunakan adalah Decision Tree, Naïve Bayes, dan K-Nearest Neighbor. Model-model ini dievaluasi menggunakan Python dan sklearn melalui dua jenis pengujian: 80:20 train split dan K-Fold 10 Cross-Validation. Hasil penelitian menunjukkan bahwa metode Decision Tree unggul dalam mendeteksi dan memprediksi gaya belajar. Pada pengujian dengan train split 80:20, akurasi yang diperoleh adalah 0,96 dengan waktu pemrosesan 0,000998 detik, sedangkan pada pengujian K-Fold 10 Cross-Validation, akurasinya mencapai 0,98 dengan waktu pemrosesan 0,04033 detik.

Dalam studi yang dilakukan oleh (Itoo et al., 2021), fokusnya adalah membandingkan efektivitas algoritma pembelajaran mesin dalam mengklasifikasikan transaksi penipuan dan non-penipuan pada data kartu kredit dengan menggunakan metode random under sampling (RUS). Temuan penelitian menunjukkan bahwa Regresi Logistik (LR) menonjol dalam kinerja dibandingkan dengan Naïve Bayes (NB) dan K-Nearest Neighbor (KNN), dengan tingkat akurasi tertinggi mencapai 95% untuk LR, 91% untuk NB, dan 75% untuk KNN. Selain itu, LR juga menunjukkan hasil yang lebih baik dalam hal Sensitivity, Specificity, Precision, dan F-Measure dibandingkan dengan NB dan KNN. Terdapat juga pengamatan bahwa teknik yang dimonitor (LR dan NB) menunjukkan kinerja yang lebih superior dalam berbagai situasi daripada teknik tanpa pemantauan seperti KNN.

 DOI : https://doi.org/ 10.30646/sinus.v22i2.833

Dalam penelitian yang dilakukan oleh (Jopri et al., 2021), disarankan untuk mengembangkan sistem kualitas daya yang lebih unggul dalam mengidentifikasi berbagai sumber harmonik. Prosesnya dimulai dengan pembuatan dan pengumpulan sinyal kualitas daya, diikuti oleh estimasi fitur tegangan dan arus. Langkah berikutnya adalah membentuk himpunan fitur tegangan dan arus. Sistem diagnostik yang diusulkan menggunakan algoritma pembelajaran mesin KNN dan NB. Hasil eksperimen menunjukkan bahwa penggunaan kombinasi fitur terbaru dengan KNN menghasilkan kinerja yang lebih baik dalam hal akurasi, presisi, sensitivitas, spesifisitas, dan F-measure. Untuk penelitian berikutnya, disarankan untuk mempertimbangkan penggunaan klasifikasi yang lebih populer seperti jaringan saraf konvolusional dalam mengidentifikasi sumber harmonik.

Dalam penelitian yang dilakukan oleh (Lestari et al., 2020), Penggunaan teknik pembelajaran mesin terbukti efektif dalam memisahkan data EEG antara kasus kejang dan non-kejang. Di antara tiga metode klasifikasi utama, yaitu KNN, Naïve Bayes, dan hutan pohon acak, KNN menonjol dengan kinerja yang superior (akurasi: 92,7%, presisi: 82,5%, sensitivitas: 73,2%, dan spesifisitas: 96,7%). Disusul oleh hutan pohon acak (akurasi: 86,6%, presisi: 68,2%, sensitivitas: 42,2%, dan spesifisitas: 96,7%), dan kemudian Naïve Bayes (akurasi: 55,6%, presisi: 25,3%, sensitivitas: 80,3%, dan spesifisitas: 50,4%). Proses pelatihan berbeda untuk setiap metode, dengan Naïve Bayes hanya membutuhkan waktu 0.166030 detik, hutan pohon acak memerlukan waktu 2.4094 detik, dan KNN adalah yang paling lambat dengan waktu pelatihan 4.789 detik.

Dalam penelitian yang dilakukan oleh Nurdina & Puspita bahwa hasil pengujian algoritma Naive Bayes dan K-NN dalam mengklasifikasikan kepuasan pelanggan maskapai penerbangan menggunakan aplikasi RapidMiner Studio versi 10.1 menunjukkan bahwa metode Naive Bayes memiliki akurasi yang lebih tinggi daripada metode K-NN. Dengan akurasi sebesar 84,48%, Naive Bayes mengungguli K-NN yang hanya mencapai 65,38%. Selain itu, presisi Naive Bayes sebesar 82,25% lebih tinggi dibandingkan dengan K-NN yang sebesar 67,35%. Begitu pula dengan nilai recall, di mana Naive Bayes mencapai 82,43% dan K-NN sebesar 74,33% (Nurdina & Puspita, 2023).

Penelitian yang dilakukan oleh Kinanti Kumarahadi et al. mengimplementasikan aplikasi sistem pendukung keputusan untuk mengklasifikasikan mahasiswa yang mengajukan keringanan menggunakan metode KNN, dan kemudian melakukan pemeringkatan dengan metode SAW. Hasil pengujian terhadap keringanan UKT menunjukkan bahwa sistem pendukung keputusan ini menghasilkan output serupa dengan perhitungan manual yang dilakukan menggunakan Microsoft Excel. Tingkat akurasi metode ini dapat mencapai 100% tergantung pada jumlah data yang digunakan. Sebagai pengembangan lebih lanjut, penelitian dapat mengeksplorasi penambahan kriteria yang dapat disesuaikan dengan kebutuhan dan kondisi instansi terkait. (Kinanti Kumarahadi et al., 2020)

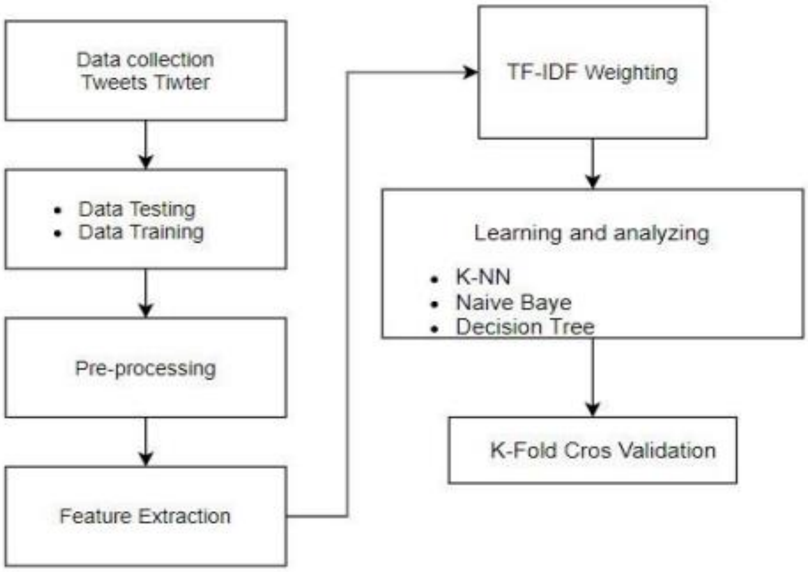
Berdasarkan hasil penelitian, klasifikasi pada Mushroom Data Set menunjukkan bahwa metode Decision Tree memberikan nilai akurasi tertinggi. Metode Cross Validation dianggap lebih dapat diandalkan karena setiap sub data memiliki kesempatan menjadi bagian dari data latih dan data uji. Pengujian menggunakan Decision Tree menghasilkan akurasi yang lebih baik dibandingkan dengan K-Nearest Neighbor. Decision Tree mencapai akurasi sebesar 0.9193, presisi sebesar 0.9227, recall sebesar 0.9193, dan skor F1 sebesar 0.9210. Sementara itu, K-Nearest Neighbor memperoleh akurasi sebesar 0.8961, presisi sebesar 0.8983, recall sebesar 0.8961, dan skor F1 sebesar 0.8972. (Chitayae & Sunyoto, 2020).

Penelitian ini hanya memfokuskan pada pengukuran kinerja algoritma data mining berdasarkan aspek akurasi, tanpa mempertimbangkan faktor-faktor lain seperti kecepatan komputasi, ketahanan, skalabilitas, dan interpretabilitas. Hasil pengujian akurasi algoritma dievaluasi dengan menggunakan konfusi matriks, yang mencerminkan nilai akurasi dari

setiap algoritma. Sebagai hasilnya, algoritma Naive Bayes mencapai tingkat akurasi tertinggi sebesar 65,59%, melebihi tingkat akurasi algoritma k-NN dengan berbagai nilai k (5: 57,88%, 10: 59,49%, 15: 59,38%, 20: 60,18%, dan 25: 61,57%) serta Algoritma *Decision Tree* yang mencapai 60,30% (Wibowo & Oesman, 2020).

III. METODE PENELITIAN

Penelitian ini melakukan analisis sentimen dengan membandingkan algoritma *naïve Bayes*, *Decision Tree*, dan *K-NN*. Pada penelitian ini dataset diperoleh dari https://github.com/ridife/dataset-idsa sebanyak 10806 *record*. Metode yang digunakan untuk proses analisis sentimen pada penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Proses Analisis Sentimen

3.1. *K-Fold Cross Validation* (*CV K-fold*)

Dalam metrik kualitatif, fakta non-numerik merupakan dasar informasi kualitatif, dan ketidakakuratan dalam kumpulan data ini disebut sebagai noise yang dapat sangat mempengaruhi prediksi informasi penting. Untuk mengatasi masalah ini, pemisahan Train-Test dan *Cross-Validation* diterapkan dalam penelitian ini. Kami memilih lima dataset secara acak dari Kaggle dan menjalankan setiap dataset melalui berbagai pengklasifikasi, membandingkan hasil untuk menentukan pengklasifikasi mana yang memberikan nilai paling akurat dalam sebagian besar kasus. Untuk setiap dataset, matriks konfusi dihasilkan, dan kami menghitung serta melaporkan setiap ukuran kinerja dalam tabel menggunakan matriks konfusi yang dihasilkan oleh masing-masing pengklasifikasi (Romadhon & Kurniawan, 2021).

1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10

Gambar 2. 10 Skema Validasi Silang K-Fold

CV 10-fold adalah salah satu metode cross-validation yang direkomendasikan untuk memilih model terbaik karena cenderung memberikan estimasi akurasi yang lebih akurat dibandingkan dengan *cross-validation* biasa, *holdout cross-validation*, dan *bootstrap*. Dalam *CV 10-fold*, data dibagi menjadi 10 bagian yang kurang lebih sama besar, sehingga terdapat 10 subset data untuk mengevaluasi kinerja model atau algoritma. Pada setiap subset dari 10 subset tersebut, *cross-validation* akan menggunakan 9 lipatan untuk pelatihan dan 1 lipatan untuk pengujian, seperti yang diilustrasikan pada Gambar 2.

3.2. Metode *Naïve Bayes*

Naïve Bayes Classifier adalah metode statistik Bayesian yang sederhana (Lestari et al., 2020). Metode ini disebut "*naif*" karena mengasumsikan bahwa semua variabel berkontribusi terhadap klasifikasi dan saling berkorelasi, yang dikenal sebagai asumsi independensi bersyarat kelas.

Ini didasarkan pada *teorema Bayes* mengenai probabilitas bersyarat:

$$P(C_i|X) = \frac{P(C_i|X)P(C_i)}{P(X)} \quad \text{.....(1)}$$

Dimana $P(C_i|X)$ = Probabilitas posterior
$P(C_i).$ = Probabilitas apriori
$P(C_i|X).$ = Probabilitas
$P(X).$ = Probabilitas prediktor kelas sebelumnya

3.3. Metode *Decision Tree*

Menurut Akshay Gole (2022), Decision Tree adalah struktur klasifikasi berbentuk pohon yang menyerupai diagram alir seperti pada flowchart. Setiap node internal mewakili pengujian terhadap suatu atribut, setiap cabang menunjukkan hasil dari pengujian tersebut, dan setiap daun atau node akhir mewakili kelas atau distribusi kelas (Akshay Gole, dkk, 2022). Untuk menentukan atribut yang akan digunakan untuk membagi pohon, digunakan konsep entropi. Semakin tinggi entropi dari sebuah sampel, semakin tidak murni sampel tersebut. Rumus untuk menghitung entropi sampel S adalah

$$\text{Entropy(S)} = \sum_i^c - P_i \, Log \, P_i \quad \text{.....(2)}$$

Di mana c adalah jumlah nilai yang terdapat pada atribut target (jumlah kelas). Sementara pi menunjukkan porsi atau perbandingan antara jumlah sampel dalam kelas i dengan jumlah total sampel dalam kumpulan data (Ramadhan et al., 2020)

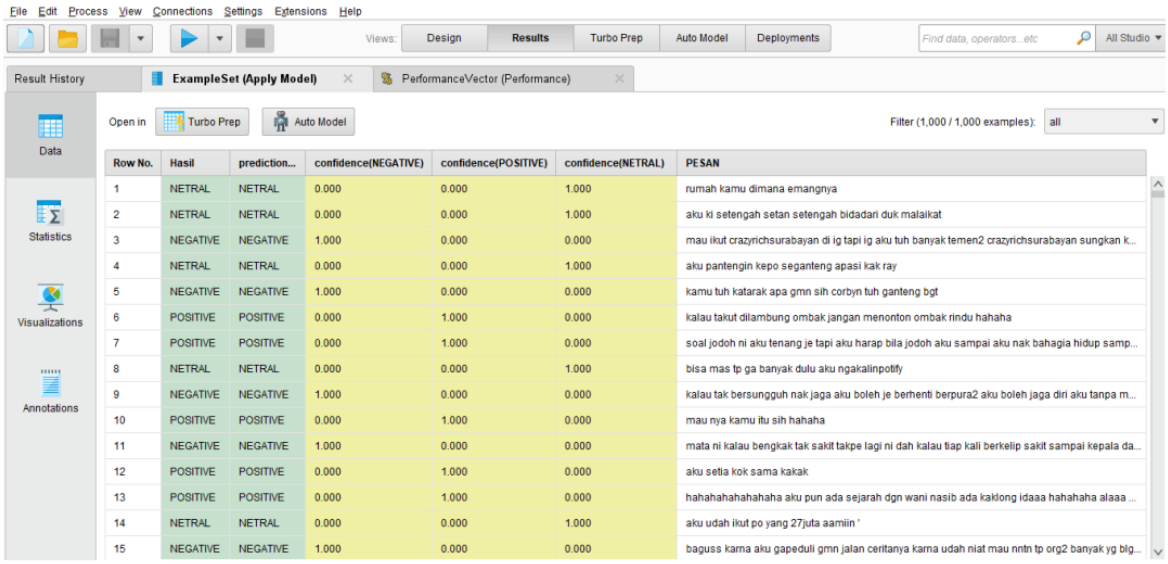
3.3 Metode *K-Nearest Neighbor*

K-nearest neighbour (k-NN) adalah algoritma supervised yang klasifikasinya didasarkan pada mayoritas (Romadhon & Kurniawan, 2021). Algoritma k-NN biasanya menggunakan jarak Euclidean atau Manhattan, tetapi jarak lain seperti norma Chebyshev atau Mahalanobis juga dapat digunakan

$$X^2 = (c-a)^2 + (d-b)^2. \quad \text{.....(3)}$$

IV. HASIL DAN PEMBAHASAN

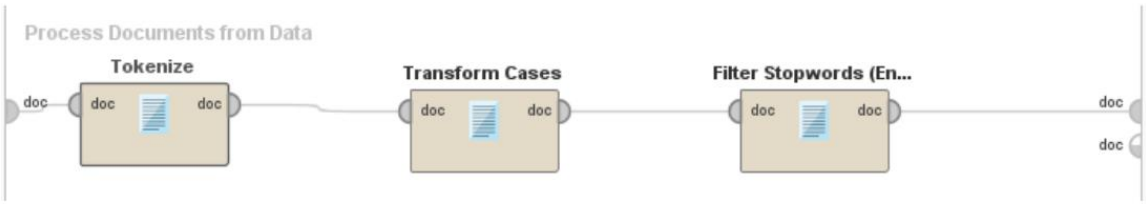
Dalam penelitian ini, peneliti menggunakan 5000 data latih dan 100 data uji. Contoh data yang digunakan dalam dataset ditampilkan pada Gambar 3.



Row No.	Hasil	prediction...	confidence(NEGATIVE)	confidence(POSITIVE)	confidence(NETRAL)	PESAN
1	NETRAL	NETRAL	0.000	0.000	1.000	rumah kamu dimana emangnya
2	NETRAL	NETRAL	0.000	0.000	1.000	aku ki setengah setan setengah bidadari duk malaikat
3	NEGATIVE	NEGATIVE	1.000	0.000	0.000	mau ikut crazyrichsurabayan di ig tapi ig aku tuh banyak temen2 crazyrichsurabayan sungkan k...
4	NETRAL	NETRAL	0.000	0.000	1.000	aku pantengin kepo seganteng apasi kak ray
5	NEGATIVE	NEGATIVE	1.000	0.000	0.000	kamu tuh katarak apa gmn sih corbyn tuh ganteng bgt
6	POSITIVE	POSITIVE	0.000	1.000	0.000	kalau takut dilambung ombak jangan menonton ombak rindu hahaha
7	POSITIVE	POSITIVE	0.000	1.000	0.000	soal jodoh ni aku tenang je tapi aku harap bila jodoh aku sampai aku nak bahagia hidup samp...
8	NETRAL	NETRAL	0.000	0.000	1.000	bisa mas tp ga banyak dulu aku ngakalinpotify
9	NEGATIVE	NEGATIVE	1.000	0.000	0.000	kalau tak bersungguh nak jaga aku boleh je berhenti berpura2 aku boleh jaga diri aku tanpa m...
10	POSITIVE	POSITIVE	0.000	1.000	0.000	mau nya kamu itu sih hahaha
11	NEGATIVE	NEGATIVE	1.000	0.000	0.000	mata ni kalau bengkak tak sakit takpe lagi ni dah kalau tiap kali berkelip sakit sampai kepala da...
12	POSITIVE	POSITIVE	0.000	1.000	0.000	aku setia kok sama kakak
13	POSITIVE	POSITIVE	0.000	1.000	0.000	hahahahahahahaha aku pun ada sejarah dgn wani nasib ada kaklong idaaa hahahaha alaaa ...
14	NETRAL	NETRAL	0.000	0.000	1.000	aku udah ikut po yang 27juta aamiin '
15	NEGATIVE	NEGATIVE	1.000	0.000	0.000	baguss karna aku gapeduli gmn jalan ceritanya karna udah niat mau nntn tp org2 banyak yg blg...

Gambar 3. Himpunan Data

Gambar 4 memperlihatkan proses ekstraksi fitur yang digunakan sebagai dasar untuk proses klasifikasi dan tokenisasi.



Gambar 4. Proses Ekstraksi Fitur

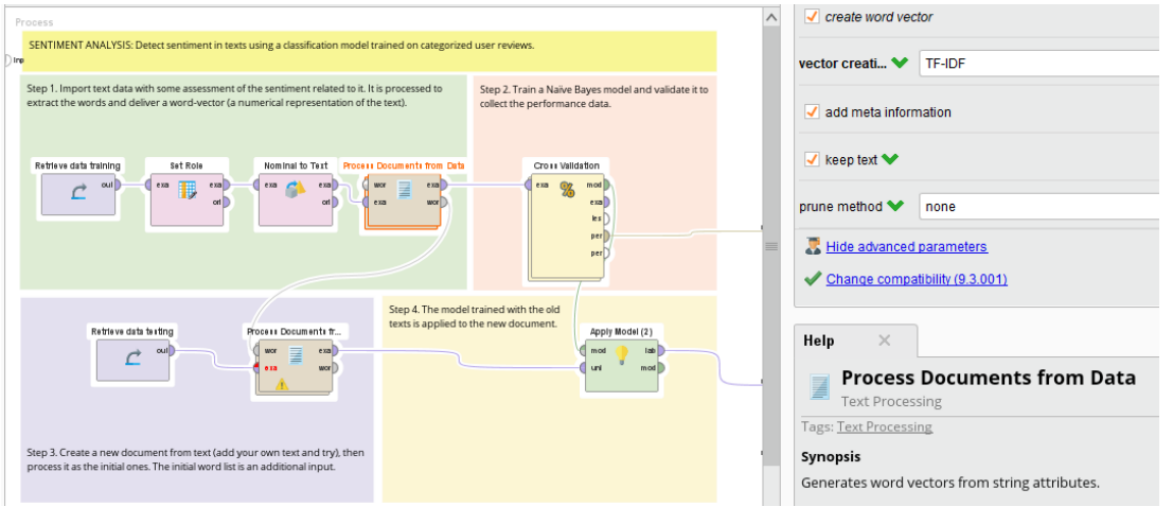
Untuk mengekstrak data menggunakan metode TF-IDF Weighting, digunakan rumus berikut:

$$TF{-}IDF(t,d) = TF(t,d)x\ IDF(t)$$

Di mana:

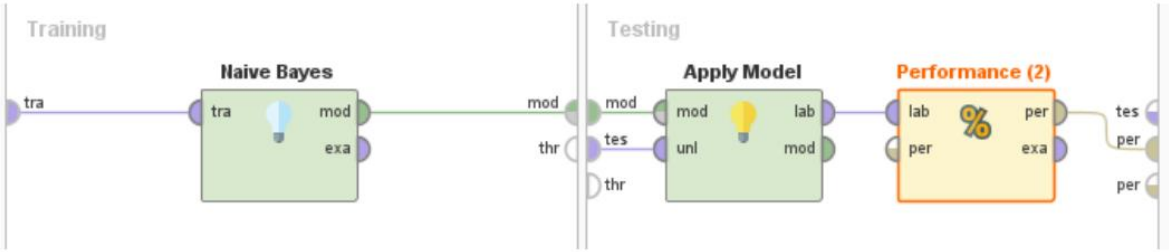
$$\text{IDF}(t) = \log(\frac{N}{df(t}$$

Desain skema model dengan metode naïve Bayes yang menggunakan 5.000 data latih dan 100 data uji ditampilkan pada Gambar 5 dan Gambar 6.



Gambar 5. Model Proses dengan Metode Naive Bayes

 DOI : https://doi.org/ 10.30646/sinus.v22i2.833



Gambar 6. Detail Model Proses Metode Naive Bayes

Penelitian yang menggunakan Metode Naïve Bayes mencapai akurasi 47,58%, sebagaimana ditunjukkan pada Gambar 7, perhitungan tersebut menggunakan rumus $P(C_x\text{-}\,X) = \frac{P(X|C_x).P(C_x)}{P(X)}$

PerformanceVector

```
PerformanceVector:
accuracy: 47.58% +/- 2.01% (micro average: 47.58%)
ConfusionMatrix:
True:   NEGATIVE        POSITIVE        NETRAL
NEGATIVE:       735     283     605
POSITIVE:       376     643     694
NETRAL: 344     319     1001
kappa: 0.214 +/- 0.031 (micro average: 0.214)
ConfusionMatrix:
True:   NEGATIVE        POSITIVE        NETRAL
NEGATIVE:       735     283     605
POSITIVE:       376     643     694
NETRAL: 344     319     1001
```

Gambar 7. Akurasi Naïve Bayes

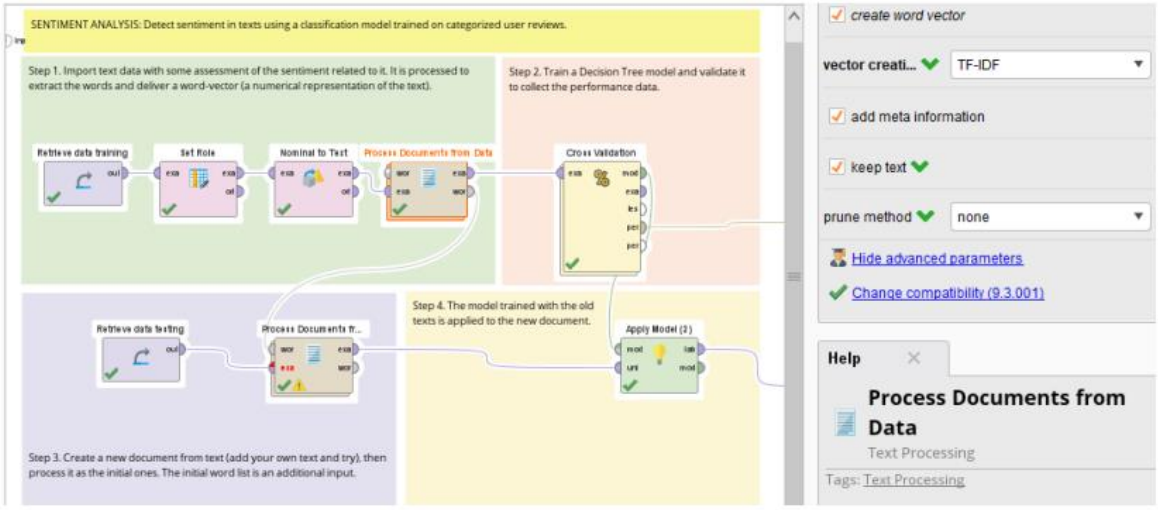
Gambar 8 menampilkan hasil uji fungsi menggunakan naïve bayes, dengan presisi untuk kelas Negatif sebesar 45,29%, kelas Positif sebesar 37,54%, dan kelas Netral sebesar 60,16%.

accuracy: 47.58% +/- 2.01% (micro average: 47.58%)

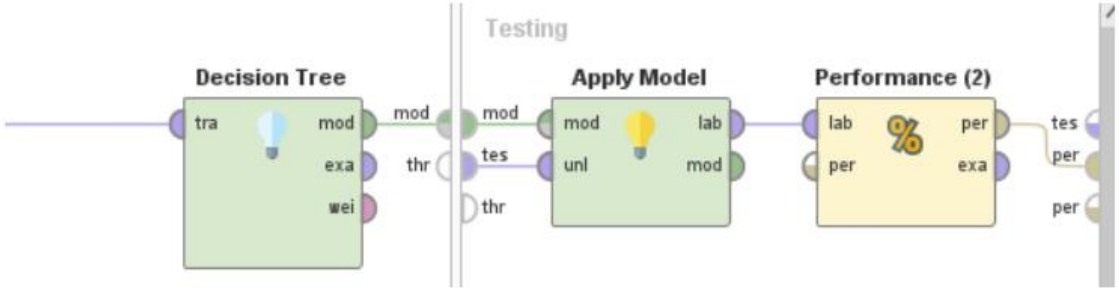
	true NEGATIVE	true POSITIVE	true NETRAL	class precision
pred. NEGATIVE	735	283	605	45.29%
pred. POSITIVE	376	643	694	37.54%
pred. NETRAL	344	319	1001	60.16%
class recall	50.52%	51.65%	43.52%	

Gambar 8. Kontingensi Naive Bayes

Desain skema model dengan metode *Decision Tree* yang menggunakan 5.000 data pelatihan dan 100 data pengujian ditampilkan pada Gambar 9 dan Gambar 10.



Gambar 9. Model Proses dengan Metode Decision Tree



Gambar 10. Detail Proses Model Metode Decision Tree

Pada penelitian dengan menggunakan Metode Decision Tree diperoleh akurasi sebesar 47,58% seperti terlihat pada Gambar 11, perhitungan tersebut menggunakan rumus $H(S) = -\sum_{i=1}^{c} p_i \, log2(p_i)$ untuk *entropy* sednan dengan *information Gain* mengunakan rumus $IG(T,a) = H(T) - \sum_{v£Values(a)} \frac{T_v}{T} HT_v$

PerformanceVector

```
PerformanceVector:
accuracy: 47.48% +/- 0.43% (micro average: 47.48%)
ConfusionMatrix:
True:   NEGATIVE        POSITIVE        NETRAL
NEGATIVE:       45      2       3
POSITIVE:       4       36      4
NETRAL: 1406    1207    2293
kappa: 0.034 +/- 0.009 (micro average: 0.034)
ConfusionMatrix:
True:   NEGATIVE        POSITIVE        NETRAL
NEGATIVE:       45      2       3
POSITIVE:       4       36      4
NETRAL: 1406    1207    2293
```

Gambar 11. Akurasi Decision Tree

Gambar 12 menunjukkan hasil uji fungsi menggunakan *Decision Tree*, dengan presisi untuk kelas Negatif sebesar 90,00%, kelas Positif sebesar 81,82%, dan kelas Netral sebesar 46,74%.

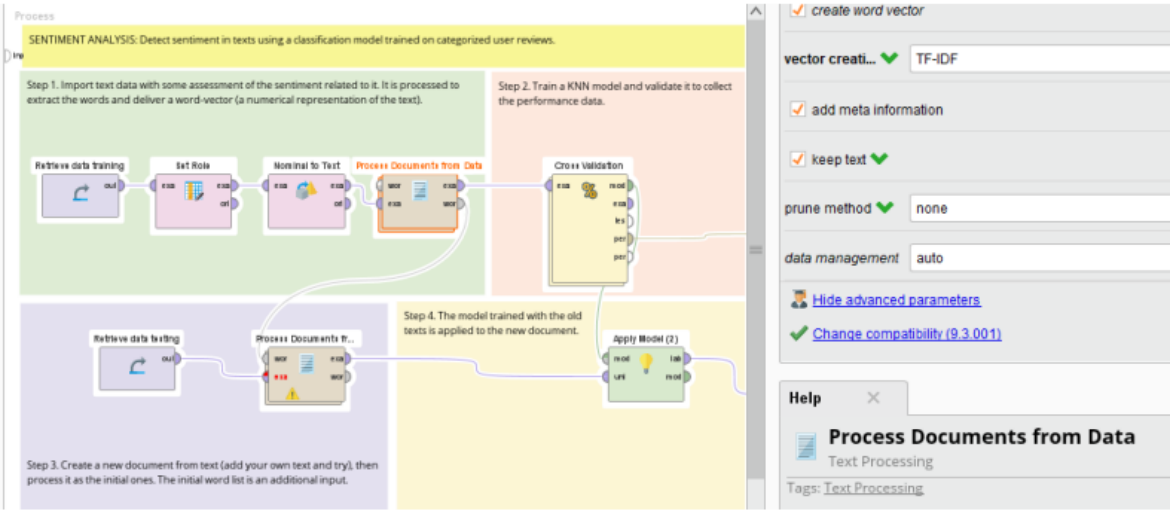
Table View Plot View

accuracy: 47.48% +/- 0.43% (micro average: 47.48%)

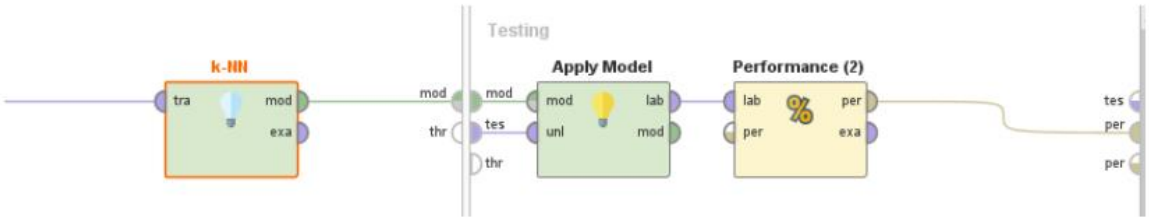
	true NEGATIVE	true POSITIVE	true NETRAL	class precision
pred. NEGATIVE	45	2	3	90.00%
pred. POSITIVE	4	36	4	81.82%
pred. NETRAL	1406	1207	2293	46.74%
class recall	3.09%	2.89%	99.70%	

Gambar 12. Kontingensi Decision Tree

Desain skema model dengan metode K-NN yang menggunakan 5000 data latih dan 100 data uji ditampilkan pada Gambar 13 dan Gambar 14.

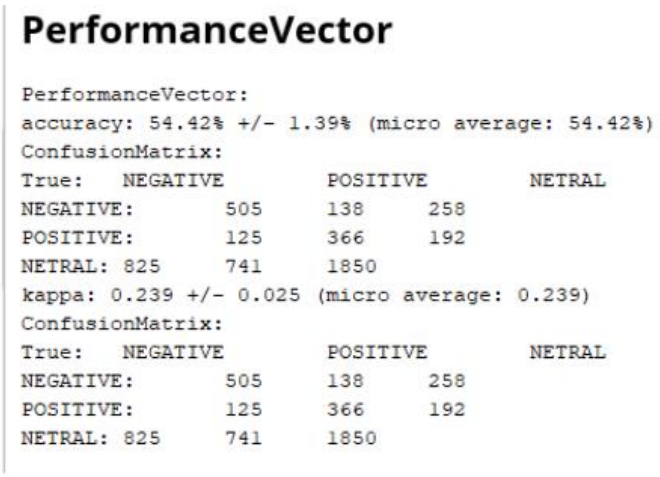


Gambar 13. Model Proses dengan Metode K-NN



Gambar 14. Detail Proses Model Metode K-NN

Penelitian yang menggunakan Metode K-NN menunjukkan akurasi sebesar 54,42%, seperti yang terlihat pada Gambar 15, perhitungan tersebut menggunakan rumus $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

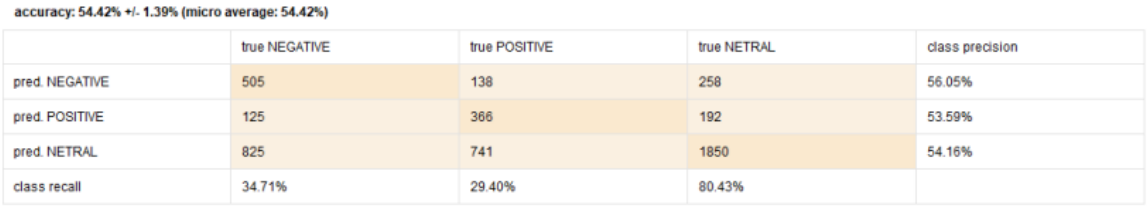


PerformanceVector

```
PerformanceVector:
accuracy: 54.42% +/- 1.39% (micro average: 54.42%)
ConfusionMatrix:
True:   NEGATIVE        POSITIVE        NETRAL
NEGATIVE:       505     138     258
POSITIVE:       125     366     192
NETRAL: 825     741     1850
kappa: 0.239 +/- 0.025 (micro average: 0.239)
ConfusionMatrix:
True:   NEGATIVE        POSITIVE        NETRAL
NEGATIVE:       505     138     258
POSITIVE:       125     366     192
NETRAL: 825     741     1850
```

Gambar 15. Akurasi K-NN

Gambar 16 menampilkan hasil uji fungsi menggunakan K-NN, dengan presisi untuk kelas Negatif sebesar 90,00%, kelas Positif sebesar 81,82%, dan kelas Netral sebesar 46,74%.



accuracy: 54.42% +/- 1.39% (micro average: 54.42%)

	true NEGATIVE	true POSITIVE	true NETRAL	class precision
pred. NEGATIVE	505	138	258	56.05%
pred. POSITIVE	125	366	192	53.59%
pred. NETRAL	825	741	1850	54.16%
class recall	34.71%	29.40%	80.43%	

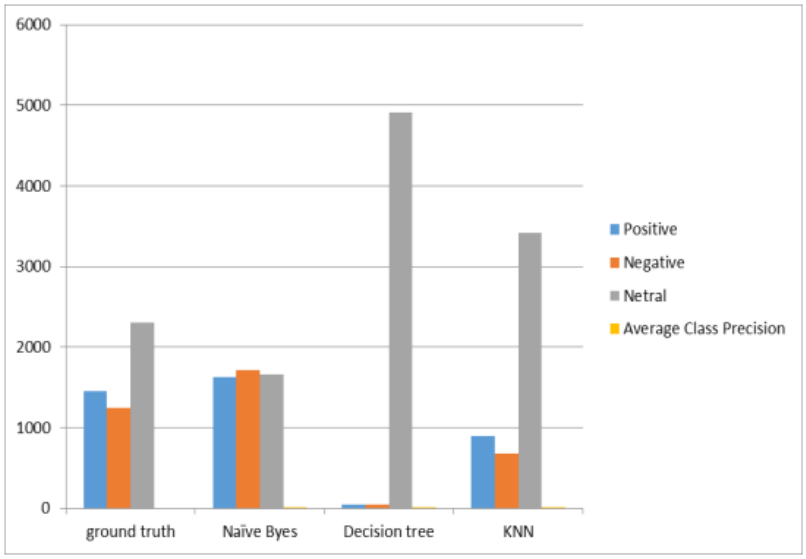
Gambar 16. Kontingensi K-NN

Penelitian ini melibatkan 5.000 data latih dan 100 data uji. Untuk pengujian, digunakan metode Naïve Bayes, Decision Tree, dan K-NN, dengan hasil akurasi ketiga metode tersebut tercantum dalam Tabel 1 dan ditunjukkan pada Gambar 17.

Tabel 1. Perbandingan Akurasi

Sentiment	ground truth	Naïve Byes	Decision tree	KNN
Positive	1455	1623	50	901
Negative	1245	1713	44	683
Netral	2300	1664	4906	3416
Average Class Precision		47.66%	72.85%	54.60%

 DOI : https://doi.org/ 10.30646/sinus.v22i2.833



Gambar 17. Grafik Perbandingan Akurasi

V. KESIMPULAN DAN SARAN

5.1 KESIMPULAN

Berdasarkan hasil pengujian, disimpulkan bahwa *Decision Tree* memberikan rata-rata presisi kelas yang lebih baik dibandingkan dengan Naïve Bayes dan algoritma K-nearest neighbour. Eksperimen dilakukan dengan menggunakan 5.100 tweet yang terbagi menjadi 5.000 data latih dan 100 data uji. Algoritma K-NN menghasilkan rata-rata presisi kelas sebesar 54,60%, Decision Tree menghasilkan rata-rata presisi kelas sebesar 72,85%, sedangkan Naïve Bayes menghasilkan rata-rata presisi kelas sebesar 47,66%, sedangkan dengan menggunakan *Decision Tree*, presisi untuk kelas Negatif mencapai 90,00%, sementara presisi untuk kelas Positif mencapai 81,82%.

5.2 SARAN

Penelitian selanjutnya mencari data dari sumber yang lain, sehingga didapat dataset yang lebih besar.

DAFTAR PUSTAKA

Akshay Gole, Sankalp Singh, Prathmesh Kanherkar, P.R.Abhishek, P. W. (2022). Comparative Analysis of Machine Learning Algorithms : Random Forest algorithm, Naive Bayes Classifier and KNN - A survey. *International Journal Research Publication & Seminat*, *13*(03). https://jrps.shodhsagar.com/index.php/j/article/view/556

Apriliyanto, E., Kusrini, K., & Arief, R. (2020). Identification Of Diseases In Rice Plant Using Chatbot With Methode Artificial Intelligence Markup Language and Normalization. *RESEARCH : Journal of Computer, Information System & Technology Management*. https://doi.org/10.25273/research.v3i2.7060

Chitayae, N., & Sunyoto, A. (2020). Performance Comparison of Mushroom Types Classification Using K-Nearest Neighbor Method and Decision Tree Method. *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, 308–313. https://doi.org/10.1109/ICOIACT50329.2020.9332148

Itoo, F., Meenakshi, & Singh, S. (2021). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, *13*(4), 1503–1511. https://doi.org/10.1007/s41870-020-00430-y

Jopri, M. H., Ab Ghani, M. R., Abdullah, A. R., Manap, M., Sutikno, T., & Too, J. (2021). K-nearest neighbor and naïve Bayes based diagnostic analytic of harmonic source identification. *Bulletin of Electrical Engineering and Informatics*, *9*(6), 2650–2657. https://doi.org/10.11591/eei.v9i6.2685

Khoirunisa, R., Apriliyanto, E., Sandi, A. S., & Kusrini, K. (2020). Penggunaan Natural Language Processing Pada Chatbot Untuk Media Informasi Pertanian. *Indonesian Journal of Applied Informatics*, *4*(2), 55. https://doi.org/10.20961/ijai.v4i2.38688

Kinanti Kumarahadi, Y., Apriliyanto, E., Yulianto, D., & Kusrini. (2020). Decision Support System For Determining The Provision Of Single Tuition Relief Using KNN and SAW Methods. *2020 8th International Conference on Cyber and IT Service Management (CITSM)*, 1–6. https://doi.org/10.1109/CITSM50537.2020.9268886

Lestari, F. P., Haekal, M., Edmi Edison, R., Ravi Fauzy, F., Nurul Khotimah, S., & Haryanto, F. (2020). Epileptic Seizure Detection in EEGs by Using Random Tree Forest, Naïve Bayes and KNN Classification. *Journal of Physics: Conference Series*, *1505*(1), 012055. https://doi.org/10.1088/1742-6596/1505/1/012055

Nurdina, A., & Puspita, A. B. I. (2023). Naive Bayes and KNN for Airline Passenger Satisfaction Classification: Comparative Analysis. *Journal of Information System Exploration and Research*, *1*(2). https://doi.org/10.52465/joiser.v1i2.167

Ramadhan, I., Sukarno, P., & Nugroho, M. A. (2020). Comparative Analysis of K-Nearest Neighbor and Decision Tree in Detecting Distributed Denial of Service. *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 1–4. https://doi.org/10.1109/ICoICT49345.2020.9166380

Romadhon, M. R., & Kurniawan, F. (2021). A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of Covid-19 Patients in Indonesia. *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 41–44. https://doi.org/10.1109/EIConCIT50028.2021.9431845

Sheth, V., Tripathi, U., & Sharma, A. (2022). A Comparative Analysis of Machine Learning Algorithms for Classification Purpose. *Procedia Computer Science*, *215*, 422–431. https://doi.org/10.1016/j.procs.2022.12.044

Sianturi, S. T., & Yuhana, U. L. (2022). Student Behaviour Analysis To Detect Learning Styles Using Decision Tree, Naïve Bayes, And K-Nearest Neighbor Method In Moodle Learning Management System. *IPTEK The Journal for Technology and Science*, *33*(2), 94. https://doi.org/10.12962/j20882033.v33i2.13665

Tella, A., Balogun, A.-L., Adebisi, N., & Abdullah, S. (2021). Spatial assessment of PM10 hotspots using Random Forest, K-Nearest Neighbour and Naïve Bayes. *Atmospheric Pollution Research*, *12*(10), 101202. https://doi.org/10.1016/j.apr.2021.101202

Wibowo, A. H., & Oesman, T. I. (2020). The comparative analysis on the accuracy of k-NN, Naive Bayes, and Decision Tree Algorithms in predicting crimes and criminal actions in Sleman Regency. *Journal of Physics: Conference Series*, *1450*(1), 012076. https://doi.org/10.1088/1742-6596/1450/1/012076