SNESTIK

Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika

https://ejurnal.itats.ac.id/snestik dan https://snestik.itats.ac.id

Informasi Pelaksanaan :
SNESTIK III - Surabaya, 11 Maret 2023
Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:
DOI : 10.31284/p.snestik.2023.4157

Prosiding ISSN 2775-5126

Fakultas Teknik Elektro dan Teknologi Informasi-Institut Teknologi Adhi Tama Surabaya
Gedung A-ITATS, Jl. Arief Rachman Hakim 100 Surabaya 60117 Telp. (031) 5945043
Email : snestik@itats.ac.id

Perbandingan Kinerja Algoritma Klasifikasi Naive Bayes, k-Nearest Neighbor dan Logistic Regression pada Dataset Multiclass

WR Wahyudi, SA Adriko, MI Firdaust, MHA Harits, Dian Puspita Hapsari
Sistem Informasi, Institut Teknologi Adhi Tama Surabaya
e-mail:13201910849@mhs.ac.id

ABSTRACT

This study compares the performance of three classification algorithms: Naive Bayes, k- Nearest Neighbor, and Logistic Regression on a multiclass dataset. The performance of each algorithm was evaluated using metrics such as accuracy, precision, recall, and F1- score. The results of the study show that the performance of the three algorithms varies depending on the specific dataset used. Overall, the logistic regression algorithm performed the best, followed by k-Nearest Neighbor and Naive Bayes. The results of this study provide useful insights for researchers and practitioners looking to select an appropriate algorithm for multiclass classification problems.

Keywords: Classification algorithms; Naive Bayes; k-Nearest Neighbor; Logistic Regression; Multiclass.

ABSTRAK

Penelitian ini membandingkan kinerja tiga algoritma klasifikasi: Naive Bayes, k- Nearest Neighbor, dan Logistic Regression pada dataset multiclass. Kinerja masing-masing algoritma dievaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-skor. Hasil penelitian menunjukkan bahwa kinerja ketiga algoritma tersebut variatif tergantung pada dataset spesifik yang digunakan. Secara keseluruhan, algoritma regresi logistik yang memiliki kinerja terbaik, diikuti oleh k-Nearest Neighbor dan Naive Bayes. Hasil penelitian ini memberikan wawasan yang bermanfaat bagi para peneliti dan praktisi yang ingin memilih algoritma yang sesuai untuk masalah klasifikasi multiclass.

Kata Kunci: Algoritma Klasifikasi; Naive Bayes; K-Nearest Neighbour; Regresi Logistik; Multi Kelas.

PENDAHULUAN

Dalam dunia data mining, klasifikasi adalah salah satu teknik yang digunakan untuk mengkategorisasikan objek ke dalam kelas-kelas tertentu. Ada berbagai macam algoritma klasifikasi yang dapat digunakan, di antaranya adalah Naive Bayes, k-Nearest Neighbor, Linear Regression, Decision Tree, Random Forest, Neural Network, dan Logistic Regression. Dalam penelitian ini akan mencoba untuk membandingkan hasil kinerja dari tiga algoritma tersebut untuk lima dataset yang berbeda. Algoritma yang akan digunakan ialah algoritma Naive Bayes, K-Nearest Neighbour, dan Logistic Regression. Penelitian ini bertujuan untuk mengidentifikasi algoritma terbaik diantara tiga pilihan algoritma klasifikasi tersebut [1][2].

Algoritma Naive Bayes merupakan salah satu metode probabilistik yang digunakan untuk klasifikasi data. Naive Bayes membuat asumsi bahwa setiap fitur yang ada dalam data adalah independen satu sama lain, meskipun dalam kenyataannya mungkin terdapat korelasi antar fitur. K-Nearest Neighbor (k-NN) adalah metode non-parametrik yang digunakan untuk klasifikasi dan regresi data. Algoritma ini berdasarkan pada konsep bahwa sebuah data akan diklasifikasikan ke dalam kelas yang sama dengan data-data lain yang paling dekat dengannya dalam feature space. Sedangkan Regresi logistik adalah metode statistik untuk menganalisis sebuah dataset dimana terdapat satu atau lebih variabel independen yang menentukan hasil. Hasil tersebut diukur dengan variabel dikotom (dimana hanya ada dua kemungkinan hasil). Ini digunakan untuk memprediksi hasil biner (1/0, Ya/Tidak, Benar/Salah) dari sekumpulan variabel independen. Regresi Logistik memodelkan probabilitas kelas default [3].

Dalam jurnal ini, kami akan mengevaluasi dan membandingkan kinerja ketiga algoritma dengan menggunakan dataset yang berbeda-beda dan menganalisis hasilnya dari segi akurasi, precision, recall, dan F1-score. Selain itu, kami juga akan menganalisis kompleksitas algoritma dan waktu komputasi yang dibutuhkan untuk menjalankan setiap algoritma, serta menjelaskan pengukuran metrik kinerja menggunakan confusion matrix. Hasil dari jurnal ini diharapkan dapat memudahkan dalam memberikan informasi yang bermanfaat bagi peneliti dan praktisi dalam memilih algoritma yang sesuai untuk digunakan dalam klasifikasi data [4][5].

Ketiga algoritma diatas memiliki kelebihannya masing - masing dalam pengklasifikasian dataset. Naive Bayes memiliki kompleksitas rendah dan dapat digunakan untuk klasifikasi dataset besar dengan fitur banyak. Ia juga dapat digunakan untuk klasifikasi multi-kelas dan teks non- numerik. Algoritma Nearest Neighbors cocok untuk klasifikasi data tidak linear dan menghasilkan hasil baik pada dataset tidak homogen. Algoritma Regresi logistik memiliki kelebihan-kelebihan seperti mudah dipahami dan sederhana, dapat digunakan untuk data kategorik maupun numerik, dapat digunakan untuk data dengan lebih dari 2 kelas, dapat digunakan untuk data yang tidak seimbang, dan dapat digunakan untuk data dengan banyak variabel independen yang dapat menangani korelasi antar variabel.

METODE

Algoritma naïve bayes adalah algoritma pembelajaran mesin untuk masalah klasifikasi yang terutama digunakan untuk klasifikasi teks yang melibatkan kumpulan data pelatihan berdimensi tinggi. Beberapa contohnya adalah analisis sentimental penyaringan spam dan mengklasifikasikan tidak hanya dikenal karena Kesederhanaannya tetapi juga untuk Efektivitasnya. Dengan algoritma Naive bayes dapat membangun model dengan cepat dan menjadikannya algoritma prediksi yang paling cepat untuk dipelajari. Algoritme ini menggunakan probabilitas suatu objek. Mengapa disebut algoritma naïve bayes karena membuat asumsi bahwa kemunculan fitur tertentu tidak tergantung pada kemunculan fitur lain bahkan jika ciri-ciri ini bergantung satu sama lain atau pada keberadaan ciri-ciri lainnya, semua sifat ini secara individual berkontribusi pada probabilitas dan itulah mengapa disebut naif Algoritma ini mengacu pada ahli statistik dan filosof Thomas Bayes [6].

Dasar dari algoritma naïve bayes adalah teorema dasar yang secara alternatif dikenal sebagai aturan Bayes atau Hukum bayes algoritma ini adalah metode untuk menghitung probabilitas kondisional yaitu probabilitas suatu peristiwa berdasarkan pengetahuan sebelumnya yang tersedia. Secara singkat algoritma naïve bayes classification adalah pengklasifikasi kumpulan data statistika yang mana untuk memprediksi semua probabilitas tiap anggota suatu class. Neural Network Dan Decision Tree memiliki persamaan kekuatan klasifikasi dengan Naïve Bayes yang didasarkan pada teorema Bayes. Naïve Bayes memiliki bukti kecepatan dan akurasi yang tinggi saat digunakan ke dalam kumpulan data data yang besar.

$$P(X) = \frac{P(X|H)\,P(H)}{P(X)}$$

Dimana :
X = Data dengan class yang belum diketahui
H = Hipotesis data X merupakan suatu class spesifik
P(H|X) = Probabilitas hipotesis H berdasarkan kondisi x (posteriori prob.)
P(H) = Probabilitas hipotesis H (prior prob.)
P(X|H) = Probabilitas X berdasarkan kondisi tersebut
P(X) = Probabilitas dari X

Algoritma K-Nearest Neighbor adalah pendekatan untuk menemukan kasus dengan menghitung kedekatan antara kasus baru dan lama. Jarak data yang ada ditentukan oleh pengguna yang diwakili oleh k. Nilai k terbaik untuk algoritma ini bergantung pada datanya. Secara umum, nilai k yang tinggi mengurangi efek noise pada klasifikasi, tetapi membuat batas antara setiap klasifikasi menjadi lebih kabur [7]. Ada banyak cara untuk mengukur kedekatan antara data baru dan data lama (data latih), antara lain jarak Euclidean dan jarak Manhattan (city distance), yang paling umum adalah jarak Euclidean. Untuk data yang memiliki kelas lebih dari satu atau multiclass menggunakan kernel berikut;

$$R^{*} \leq R_{kNN} \leq R^{*}\left(2 - \frac{M R^{*}}{M-1}\right)$$

Dimana:
R^{*} adalah Bayes error rate/rerata kesalahan Bayes (yang mana notasi tersebut dianggap sebagai rerata minimal kesalahan yang mungkin terjadi). Jika nilainya mendekati nol, batas ini berkurang menjadi "tidak lebih dari dua kali tingkat kesalahan Bayesian".
R_{kNN} merupakan rerata kesalahan k-NN

M merupakaan jumlah kelas

Selanjutnya Algoritma Logistic Regression adalah algoritma klasifikasi machine learning yang menganalisis menggunakan metode statistik untuk memprediksi hasil biner, seperti ya atau tidak, berdasarkan pengamatan sebelumnya dari kumpulan data [8][9]. Model regresi logistik memprediksi variabel data dependen dengan menganalisis hubungan antara satu atau lebih variabel independen yang ada. Fungsi Logistik dapat dijelaskan sebagai berikut;

$$p(x) = \frac{1}{1+ e^{-(x-\mu)/s}}$$

Dimana:

μ merupakan parameter lokasi (midpoint/titik temu dari kurva, dimana $p(\mu) = \frac{1}{2}$

Untuk meminimalkan *The Loss* maka dapat memaksimumkan estimasi *Likelihood*.

Kelompok data gambar yang digunakan

Kelompok data yang berbeda sebagai data uji coba, terdiri dari 5 kelompok daat. Pertama adalah kelompok data Bigmart Sales yang memiliki record/instance sebanyak 5681, dataset ini akan diklasifikasikan berdasarkan ukuran toko yaitu dibagi antara besar, sedang dan kecil. Dataset kedua merupakan dataset mengenai Superstore memiliki record sebanyak 9995, dataset ini akan diklasifikasikan berdasarkan kategori produk yang dijual yaitu dibagi antara produk furniture, office supplies dan technology. Dataset ketiga merupakan dataset mengenai Store sales memiliki record sebanyak 4249, dataset ini akan diklasifikasikan berdasarkan tipe produk yang dijual yaitu dibagi antara coffe, tea, espresso, dan herbal tea. Dataset keempat merupakan dataset mengenai Supermarket sales memiliki record sebanyak 1001, dataset ini akan diklasifikasikan berdasarkan lokasi supercenter yaitu dibagi antara yangon, naypyidaw, dan mandalay. Dataset terakhir merupakan dataset mengenai US store memiliki record sebanyak 3256, dataset ini akan diklasifikasikan berdasarkan Wilayah Toko yaitu dibagi antara timur, barat, utara dan selatan.

HASIL DAN PEMBAHASAN

Hasil simulasi untuk nilai recall, precision dan F1 score kelompok data gambar

Berikut ini merupakan hasil simulasi 3 algoritma pengklasifikasi yang digunakan dalam penelitian ini. Berawal dari Confusion Matrix sebagai pengukuran kinerja untuk masalah klasifikasi pada pembelajaran mesin dimana keluaran dapat berupa dua kelas atau lebih. Hasil akurasi algoritma pengklasifikasi tersebut diukur menggunakan alat ukur kinerja antara lain *Recall, Precision*, dan F1 *Score*. Masing-masing alat ukur kinerja tersebut digunakan sesuai dengan fungsinya. *Recall* atau *Sensitivity (True Positive Rate) Recall* menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi. *Precision* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Skor F1 adalah nilai yang memberi tahu seberapa tepat algoritma pengklasifikasi (berapa banyak instance yang diklasifikasikan dengan benar), serta seberapa robust model (tidak melewatkan sejumlah besar instance).

Tabel 1. *Klasifikasi Algoritma Naive Bayes*

Data set	Correctly Classified Instances	Incorrectly Classified Instances	Kappa static	Mean Error	Root mean Error	Relative Error	Root Relativ e Error

1	100%	0%	1	0,0015	0,0059	0,3756%	1,3078%
2	98,69 %	1,3008 %	0.9768	0.0125	0.0891	3.3645 %	20.66%
3	99.34 %	0.6591 %	0.9912	0.0037	0.0542	0.9782 %	12.53%
4	100%	0%	1	0.0114	0.0265	2.5527 %	5.6063 %
5	77.52%	22.47%	0.7016	0.1176	0.2808	32.16%	65.65 %

Tabel 2. *Kinerja rata-rata Algoritma Naive bayes*

Data set	Precision	Recall	F-Measure	ROC Area
1	1,000	1,000	1,000	0.984
2	0.988	0.987	0.987	0.992
3	0.993	0.993	0.993	0.999
4	1,000	1,000	1,000	1,000
5	0.860	0.775	0.793	0.977

Tabel 3. *Klasifikasi Algoritma Logistic Regression*

Data set	Correctly Classified Instances	Incorrectly Classified Instances	Kappa static	Mean Error	Root mean Error	Relative Error	Root Relative Error
1	99.12%	0.88%	0.9856	0.006	0.0741	1.4742 %	16.37 %
2	98.69 %	1,3008%	0.9768	0.0125	0.0891	3.3645 %	20.66 %
3	100%	0%	1	0	0	0%	0%
4	92,5%	7,5%	0.8878	0.0503	0.2237	11.31 %	47.39 %
5	100%	0%	0	0	0	0%	0%

Tabel 4. *Kinerja rata-rata Algoritma Logistic Regression*

Data set	Precision	Recall	F-Measure	ROC Area
1	0.991	0.991	0.991	0.938
2	0.988	0.987	0.987	0.992
3	1,000	1,000	1,000	1,000
4	0.932	0.925	0.924	0.986
5	1,000	1,000	1,000	1,000

Tabel 5. *Klasifikasi Algoritma k-Nearest Neighbor*

Data set	Correctly Classified Instances	Incorrectly Classified Instances	Kappa static	Mean Error	Root mean Error	Relative Error	Root Relative Error
1	100%	0%	1	0.0074	0.0274	1.7991 %	6.0524 %
2	91.12 %	8.87 %	0.83	0.1402	0.2294	37.73 %	53.22 %

3	100%	0%	1	0.0028	0.0177	0.7524 %	4.099 %
4	100%	0%	1	0.0881	0.1312	19.81 %	27.80 %
5	100%	0%	1	0.0266	0.0666	7.2817 %	15.56 %

Tabel 6. *Kinerja rata-rata Algoritma k-Nearest Neighbor*

Dataset	Precision	Recall	F-Measure	ROC Area
1	1,000	1,000	1,000	0.954
2	0.917	0.911	0.909	0.984
3	1,000	1,000	1,000	1,000
4	1,000	1,000	1,000	1,000
5	1,000	1,000	1,000	1,000

Berdasarkan kinerja rata-rata dari ketiga model yang telah dijelaskan sebelumnya, selanjutnya melakukan perbandingan mengenai kinerja model terbaik hingga terburuk dalam melakukan klasifikasi kelima dataset tersebut. Hasil dari ketiga model yang digunakan adalah seperti tabel berikut:

Tabel 7. *Perbandingan hasil dari ketiga Model*

Model	Naive Bayes	Logistic Regression	K-Nearest Neighbor
Dataset yang diklasifikasi 100%	2 Dataset	2 Dataset	4 Dataset
Dataset yang memiliki 0% Error	0 Dataset	2 Dataset	0 Dataset

Dari ketiga algoritma klasifikasi yang digunakan, K-Nearest Neighbor merupakan algoritma dengan hasil dan tingkat akurasi paling tinggi. K-Nearest Neighbor menghasilkan akurasi terbaik karena mendapatkan hasil 100% pada pengklasifikasian instans dalam 4 dataset, sedangkan kedua algoritma lain memiliki 2 dataset dengan hasil klasifikasi 100%. Berdasarkan evaluasi proses pengklasifikasian pada model algoritma Naïve Bayes, Logistic Regression dan k-Nearest Neighbor, dapat dinyatakan bahwa kinerja terbaik dihasilkan oleh algoritma K-Nearest Neighbor.

KESIMPULAN

Algoritma K-Nearest Neighbor pada klasifikasi 5 dataset menghasilkan kinerja terbaik dengan hasil 4 dataset yang dapat diklasifikasikan dengan nilai 100%, akan tetapi dalam pengukuran error didapatkan oleh Logistic Regression dengan 2 dataset yang memiliki error 0%. Hasil tersebut membuat k-Nearest neighbor memiliki nilai terbaik jika dibandingkan Naïve Bayes dan Logistic Regression untuk Correctly Classified Instance. Sedangkan dalam pengukuran error paling kecil didapatkan oleh Logistic Regression karena kedua model lain memiliki error dalam keseluruhan dataset yang diuji.

DAFTAR PUSTAKA

[1] A. Robles-Guerrero, T. Saucedo-Anaya, E. González-Ramírez, and J. I. De la Rosa-Vargas, "Analysis of a multiclass classification problem by Lasso Logistic Regression and Singular Value Decomposition to identify sound patterns in queenless bee colonies," *Comput. Electron. Agric.*, vol. 159, no. February, pp. 69–74, 2019, doi: 10.1016/j.compag.2019.02.024.

[2] N. Singh and P. Singh, "A novel Bagged Naïve Bayes-Decision Tree approach for multi-class classification problems," *J. Intell. Fuzzy Syst.*, vol. 36, no. 3, pp. 2261–2271, 2019, doi: 10.3233/JIFS-169937.

[3] L. Mandal and N. D. Jana, "A comparative study of naive bayes and k-NN algorithm for multi-class drug molecule classification," *2019 IEEE 16th India Counc. Int. Conf. INDICON 2019 - Symp. Proc.*, pp. 12–15, 2019, doi: 10.1109/INDICON47234.2019.9029095.

[4] U. Bentkowska, J. G. Bazan, M. Mrukowicz, L. Zareba, and P. Molenda, "Multi-class classification problems for the k-NN algorithm in the case of missing values," *IEEE Int. Conf. Fuzzy Syst.*, vol. 2020-July, 2020, doi: 10.1109/FUZZ48607.2020.9177592.

[5] M. Dreisig, M. H. Baccour, T. Schack, and E. Kasneci, "Driver Drowsiness Classification Based on Eye Blink and Head Movement Features Using the k-NN Algorithm," *2020 IEEE Symp. Ser. Comput. Intell. SSCI 2020*, pp. 889–896, 2020, doi: 10.1109/SSCI47803.2020.9308133.

[6] "Bayes-Not So," vol. 69, no. 3, pp. 385–398, 2014.

[7] D. Coomans and D. L. Massart, "Alternative k-nearest neighbour rules in supervised pattern recognition," *Anal. Chim. Acta*, vol. 136, pp. 15–27, 1982, doi: 10.1016/s0003-2670(01)95359-0.

[8] Q. Mary, "Single-Label Multi-Class Image Classification by Deep Logistic Regression," no. 2.

[9] S. H. Walker and D. B. Duncan, "Estimation of the probability of an event as a function of several independent variables.," *Biometrika*, vol. 54, no. 1, pp. 167–179, 1967, doi: 10.1093/biomet/54.1-2.167.