ISSN (Online) 2747-0563
*Volume 2 Tahun 2021*

# Studi Literatur Tentang Performa Naïve Bayes dalam Klasifikasi Data

Andreas Nugroho Sihananto[1], Hendra Maulana[2]

[1,2] Program Studi Informatika, Universitas Pembangunan Nasional "Veteran" Jawa Timur
[1]andreas.nugroho.jarkom@upnjatim.ac.id
[2]hendra.maulana.if@upnjatim.ac.id

***Abstrak*— Klasifikasi data merupakan salah satu proses penting dalam pengolahan data. Salah satu metode klasifikasi data yang populer adalah Naïve Bayes yang dikenal karena mampu mengklasifikasi data secara cepat. Meski begitu ketika dihadapkan pada data-data tertentu akurasi klasifikasi dari Naïve Bayes sangat rendah. Dari hasil studi literatur yang telah dilakukan ditemukan fakta bahwa untuk data-data dengan atribut yang tidak terlalu banyak dan setiap atributnya memiliki bobot yang sama Naïve Bayes masih dapat melakukan klasifikasi dengan akurasi tinggi. Untuk data-data dengan bobot atribut yang berbeda-beda, dilakukan modifikasi atau hibridisasi terhadap Naïve Bayes. Hasil dari modifikasi ini berhasil meningkatkan akurasi klasifikasi Naïve Bayes.**

*Kata Kunci*— **Naïve Bayes, klasifikasi data, atribut, performa, studi literatur**

## I. PENDAHULUAN

Revolusi Industri 4.0, memunculkan adanya suatu fenomena koleksi data-data digital yang berukuran sangat besar yang disebut sebagai big data. Data-data mentah yang dikumpulkan ini akan memiliki nilai ekonomi setelah diolah. Kebutuhan pengolahan data yang umum dibutuhkan antara lain prediksi kondisi di masa depan, analisa sentimen, sistem pendukung keputusan, serta pendeteksian objek atau pengenalan pola. Salah satu algoritma pengenalan pola yang banyak diimplementasikan adalah Naïve Bayes.

Naive Bayes adalah algoritma yang lazim digunakan dalam Data Mining atau Machine Learning (ML). Algoritma ini didasarkan pada Teorema Bayesdan merupakan salah satu algoritma ML paling sederhana namun banyak digunakan serta diaplikasikan di banyak industri. Naive Bayes menggunakan Teorema Bayes dan mengasumsikan bahwa semua prediktor adalah independen. Dengan kata lain, pengklasifikasi ini mengasumsikan bahwa kehadiran satu fitur tertentu dalam suatu kelas tidak mempengaruhi kehadiran fitur lainnya.

Naïve Bayes sendiri memiliki rumus dasar sebagaimana ditunjukkan pada persamaan (1)

$$P(c|x) = \frac{(Px|c)*P(c)}{P(x)} \qquad (1)$$

Yang mana :

$$P(c|x) = P(x_1|c\ ) * P(x_2|c) * \ldots . P(x_n|c) * P(c)$$

P (c|x) adalah probabilitas posterior menurut prediktor (x) untuk kelas (c). P(c) adalah probabilitas sebelumnya dari kelas, P(x) adalah probabilitas sebelumnya dari prediktor, dan P(x|c) adalah probabilitas dari prediktor untuk kelas tertentu (c).

Naïve Bayes memiliki beberapa kelebihan antara lain memiliki kecepatan komputasi yang cepat, serta mampu menyelesaikan masalah berbasis *multi-class*. Meski begitu pada beberapa kasus ditemukan hasil klasifikasi dari Naïve Bayes ternyata jauh dari kenyataan lapangan. Hal ini biasanya disebabkan karena ada satu atau lebih prediktor yang memiliki bobot atau pengaruh yang lebih besar daripada prediktor yang lain. Kelemahan ini biasanya ditutupi dengan melakukan hibridisasi atau modifikasi pada Naïve Bayes.

## II. METODE PENELITIAN

ISSN (Online) 2747-0563
*Volume 2 Tahun 2021*

## III. HASIL DAN PEMBAHASAN

Temuan dari studi literatur yang telah dilakukan dapat dilihat pada Tabel 1.

TABEL I
HASIL TEMUAN REVIEW

| No. | Pengarang | Permasalahan yang diangkat | Hasil | Saran atau kekurangan |
|---|---|---|---|---|
| 1 | [1] | Class-specific attribute weighted naive Bayes (CAWNB) | CAWNB mampu mempertahankan kualitas kecepatan Naïve Bayes sekaligus meningkatkan akurasinya jika dibandingkan metode modifikasi Naïve Bayes lainnya | Pada dataset penyakit jantung, diabetes, 6 tipe gelas, dan ionosphere tidak menunjukkan perubahan akurasi yang berarti jika dibandingkan metode lainnya |
| 2 | [2] | Prediksi Produksi Karet | Peneliti mengklaim hasil akurasinya dapat diterima dan digunakan di lapangan | Peneliti tidak menyebutkan angka pasti akurasi sistem yang mereka buat |
| 3 | [3] | E-commerce product review sentiment | Peneliti mengusulkan modifikasi bernama Naïve Bayes continuous learning framework yang berhasil meningkatkan 4akurasi klasifikasi antara 1,08 hingga 5,27 persen pada data produk yang jumlahnya lumayan besar | Pengujian pada data berukuran ribuan telah menunjukkan hasil memuaskan namun pengujian pada *domain-specific sentiment* tidak memberikan perubahan yang signifikan |
| 4 | [4] | Class-specific attribute value weighting for Naive Bayes (CAVWNB) | Terdapat peningkatan akurasi secara eksperimental jika dibandingkan CAWNB milik [1] | Dalam beberapa kasus dataset uji seperti *balance-scale* kenaikan akurasinya nyaris tidak signifikan jika dibandingkan CAWNB |
| 5 | [5] | Penggunaan Naïve Bayes dalam Klasifikasi Big Data | Naïve Bayes di-hybrid-kan dengan algoritma Cuckoo–Grey wolf dan disebut sebagai CG-CNB. Pengujian pada kasus data berukuran besar menghasilkan tingkat akurasi, sensitivitas dan ketegasan yang lebih tinggi | Peneliti sebaiknya juga mencantumkan perbandingan waktu eksekusi antara CG-CNB dengan algoritma-algoritma pembandingnya |
| 6 | [6] | Prediksi Tingkat Penyebaran Covid-19 Di Indonesia | Akurasi yang dihasilkan sangat rendah, hanya 48, 48% | Naïve Bayes sebaiknya digabungkan dengan metode lainnya guna meningkatkan akurasi |
| 7 | [7] | Pendeteksian Penyakit Kulit | Penyakit kulit keratosis serta tumor jinak serta kanker kulit melanoma berhasil diprediksi dengan tingkat akurasi di atas 90% | Akurasi yang tinggi ini tercapai salah satunya karena dukungan dataset yang sudah diseleksi. Penggunaan dataset lain sebagai dataset uji sebaiknya dilakukan |
| 8 | [8] | Sistem Pendukung Keputusan Penerimaan Karyawan Baru | Dengan 20 data uji, akurasi yang dihasilkan Naïve Bayes hanya mencapai 60% | Dataset kecil dan akurasi yang dihasilkan rendah menunjukkan untuk kasus ini Naïve Bayes memerlukan hibridisasi dengan algoritma lain |
| 9 | [9] | Pendeteksian Penyakit Kulit Kucing | Algoritma Naïve Bayes dikombinasikan dengan Certainty Factor berhasil menghasilkan tingkat akurasi 100% | Perlu dicoba dengan dataset yang lebih besar. |
| 10 | [10] | Usulan *novelty* dalam Naïve Bayes yakni Selective Naïve Bayes | Selective Naïve Bayes berhasil mengatasi masalah akurasi rendah pada 65 dataset yang diuji | Pengukuran akurasi datanya menggunakan RMSE bukan hasil akurasi yang real |
| 11 | [11] | Analisa Sentimen Pengguna e-Money | Hasil data uji sentimen menggunakan Naïve Bayes menghasilkan akurasi 84%, lebih tinggi daripada algoritma C4.5 yang hanya 80% atau metode *Voting by majority* | Metode voting by majority sebaiknya hanya diberlakukan jika algoritma yang digunakan berjumlah 3 atau 4 |
| 12 | [12] | Klasifikasi | *Multinomial* | Metode |

ISSN (Online) 2747-0563
*Volume 2 Tahun 2021*

| | | | | |
|---|---|---|---|---|
| | | Berita Akun Twitter Suara Surabaya dengan Multinomial Naïve Bayes | *Naïve Bayes* menghasilkan akurasi klasifikasi sebesar 89% | *Multinomial Naïve Bayes* belum terlalu tepat menentukan lokasi kejadian |
| 13 | [13] | Analisa sentimen vaksinasi Covid-19 di Indonesia dengan Naïve Bayes | Peneliti berhasil mengklasifikasi twit bernuansa negatif sebanyak 56%, twit positif 39% dan twit netral sejumlah 1% | Ada sejumlah twit (4%) yang gagal diklasifikasi |
| 14 | [14] | Diagnosa ADHD dengan Naïve Bayes | Data uji pertama menghasilkan nilai akurasi 100%, nilai sensitifitas 100% dan nilai spesifitas 100%. Adapun data ujia kedua menghasilkan nilai akurasi 93,3%, nilai sensitifitas 100% dan nilai spesifitas 87,5%. | Variasi jawaban pada data latih maupun data uji sangat kurang ditambah lagi jumlah data amat kecil |
| 15 | [15] | Analisa hashtag kuliner | Hasil terbaik didapat dengan metode Multinomial Naïve Bayes dengan nilai akurasi 0.71, rata-rata precision 0.80, rata-rata recall 0.53, dan rata-rata f-score 0.53. | Ada sejumlah twit yang hanya menggunakan gambar dan tidak ada fungsi pengolahan gambar menjadi teks sehingga terjadi sejumlah kesalahn prediksi |

Hasil penelitian oleh [7], [11], [12], [14] menunjukkan bagaimana pada Naïve Bayes sukses mengklasifikasikan data dengan akurasi klasifikasi yang tinggi. Namun penelitian-penelitian tersebut biasanya menggunakan data-data dengan fitur yang sedikit atau jumlah data yang kecil. Pada kenyataannya temuan [6], [8] menunjukkan bahwa ketika data-data yang harus diklasifikasi memiliki banyak fitur dan ditambah adanya fitur yang memiliki bobot lebih besar daripada fitur yang lain, algoritma Naïve Bayes menghasilkan akurasi yang rendah[10].

Rendahnya akurasi Naïve Bayes memunculkan gagasan sejumlah peneliti untuk memodifikasi algoritma Naïve Bayes supaya dapat mengenali pembobotan fitur. Beberapa algoritma *Selective Naïve Bayes*[10], *Naïve Bayes continuous learning* [3], *Class-specific attribute weighted naive Bayes* (CAWNB)[1] *dan Class-specific attribute value weighting for Naive Bayes* (CAVWNB)[4] melakukan perubahan dengan mengubah atau memodifikasi formula klasifikasi. Adapun pendekatan lain seperti yang dilakukan [5] dan [9], menggunakan hibridisasi dengan algoritma lain.

Modifikasi dan hibridisasi memang berhasil meningkatkan tingkat akurasi Naïve Bayes, terutama pada data-data yang memiliki atribut dan kelas yang tidak terlalu banyak. Beberapa modifikasi seperti Selective Naïve Bayes [10] mampu mereduksi jumlah atribut tanpa mengurangi akurasi. Namun metode *Naïve Bayes continuous learning* tidak melakukan reduksi atribut karena metode ini hanya dapat dioperasikan pada analisis sentimen [1], [3].

Usulan dari [1], [5] dan [4] merupakan usulan modifikasi Naïve Bayes yang dapat diaplikasikan ke berbagai macam dataset. Hanya saja usulan *Hybrid Cuckoo–Grey wolf dan Correlative Naïve Bayes(*CG-CNB*)*[5] tidak melakukan pengujian kepada data-data dengan atribut dan kelas yang beragam. Adapun usulan CAWNB[1] dan CAVWNB[4] meski sukses dalam sejumlah besar dataset, namun tidak menunjukkan peningkatan yang signifikan ketika diuji dengan data-data kompleks seperti pengenalan objek gelas serta diagnosa penyakit medis jantung dan hepatitis. Eksplorasi lebih jauh mengenai modifikasi algoritma Naïve Bayes masih sangat dimungkinkan mengingat algoritma ini sangat populer dan mudah dimodifikasi sesuai kebutuhan penggunanya.

## IV. KESIMPULAN

Naïve Bayes telah lama dikenal sebagai algoritma klasifikasi yang memiliki performa komputasi yang cepat, namun kecenderungan Naïve Bayes untuk menyamaratakan semua fitur yang mempengaruhi klasifikasi membuat Naïve Bayes sering tidak cocok diterapkan di dunia nyata. Pada kasus-kasus dengan fitur-fitur yang sedikit seperti analisa sentimen atau diagnosa ADHD Naïve Bayes mampu memberikan hasil yang memuaskan. Namun jika kasus yang diklasifikasi memiliki banyak fitur dengan bobot yang berbeda-beda, maka perlu dilakukan modifikasi pada Naïve Bayes. Kasus-kasus yang cenderung sulit diklasifikasi oleh Naïve Bayes biasa antara lain berhubungan dengan prediksi sebaran wabah, prediksi keputusan, serta pendeteksian penyakit medis

Beberapa modifikasi yang berhasil memberikan perubahan signifikan terhadap akurasi Naïve Bayes pada data dengan banyak fitur antara lain Selective Naïve Bayes, *Hybrid Algoritma Cuckoo–Grey wolf dan Correlative Naïve Bayes(*CG-CNB*), Naïve Bayes continuous learning, Class-specific attribute weighted naive Bayes* (CAWNB) *dan Class-specific attribute value weighting for Naive Bayes* (CAVWNB). Dengan berkembangnya tren data yang semakin lama berukuran semakin besar, kebutuhan algoritma klasifikasi dengan kemampuan komputasi cepat seperti Naïve Bayes tentu saja semakin diperlukan. Penelitian-penelitian tentang modifikasi Naïve Bayes yang menunjukkan

ISSN (Online) 2747-0563
*Volume 2 Tahun 2021*

peningkatan akurasi yang lumayan signifikan menunjukkan bahwa Naïve Bayes masih merupakan algoritma klasifikasi yang fungsinya masih signifikan sampai sejauh ini.

## UCAPAN TERIMA KASIH

Ucapan terimakasih kepada Progdi Informatika Fakultas Ilmu Komputer UPN "Veteran" Jawa Timur yang telah mendanai artikel ini.

## REFERENSI

[1] L. Jiang, L. Zhang, L. Yu, and D. Wang, "Class-specific attribute weighted naive Bayes," *Pattern Recognit.*, vol. 88, pp. 321–330, 2019, doi: 10.1016/j.patcog.2018.11.032.

[2] M. Hawari and B. Sinaga, "Naïve Bayes Algorithm Implementation To Predict Gum Production at PT. Sri Rahayu Court," *J. Mantik*, vol. 3, no. 3, pp. 40–45, 2019.

[3] F. Xu, Z. Pan, and R. Xia, "E-commerce product review sentiment classification based on a Naïve Bayes continuous learning framework," *Inf. Process. Manag.*, vol. 57, no. 5, p. 102221, 2020, doi: 10.1016/j.ipm.2020.102221.

[4] H. Zhang, L. Jiang, and L. Yu, "Class-specific attribute value weighting for Naive Bayes," *Inf. Sci. (Ny).*, vol. 508, pp. 260–274, 2020, doi: 10.1016/j.ins.2019.08.071.

[5] C. Banchhor and N. Srinivasu, "Integrating Cuckoo search-Grey wolf optimization and Correlative Naive Bayes classifier with Map Reduce model for big data classification," *Data Knowl. Eng.*, vol. 127, no. November 2018, p. 101788, 2020, doi: 10.1016/j.datak.2019.101788.

[6] A. F. Watratan, A. B. Puspita, and D. Moeis, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," *J. Appl. Comput. Sci. Technol.*, vol. 1, no. 1, pp. 7–14, 2020, doi: 10.52158/jacost.v1i1.9.

[7] V. R. Balaji, S. T. Suganthi, R. Rajadevi, V. Krishna Kumar, B. Saravana Balaji, and S. Pandiyan, "Skin disease detection and segmentation using dynamic graph cut algorithm and classification through Naive Bayes classifier," *Meas. J. Int. Meas. Confed.*, vol. 163, p. 107922, 2020, doi: 10.1016/j.measurement.2020.107922.

[8] T. D. Pangestuti, F. T. Anggraeny, and E. P. Mandyartha, "Rancang Bangun Sistem Pendukung Keputusan Penerimaan Karyawan Baru Menggunakan Metode Naive Bayes Classifier (Studi …," *J. Inform. dan Sist. Inf.*, vol. 1, no. 3, pp. 1072–1080, 2020, [Online]. Available: http://jifosi.upnjatim.ac.id/index.php/jifosi/article/view/236.

[9] F. Rahmawati, Y. V. Via, and E. Y. Puspaningrum, "Implementasi Metode Naive Bayes Dan Certainty Factor Dalam Mendiagnosa Penyakit Kulit," *J. Inform. dan Sist. Inf.*, vol. 1, no. 1, pp. 631–641, 2020.

[10] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Syst.*, vol. 192, p. 105361, 2020, doi: 10.1016/j.knosys.2019.105361.

[11] Z. E. Sholikha, E. Y. Puspaningrum, and W. S. JS, "Analisa Sentimen Pengguna E-Money Pada Twitter Menggunakan Algoritma C4.5 dan Naive Bayes," *J. Inform. dan Sist. Inf.*, vol. 1, no. 3, 2020.

[12] Qonita, E. D. Wahyuni, and A. A. Arifiyanti, "Klasifikasi Berita Pada Akun Twitter Suara Surabaya Menggunakan Metode Naïve Bayes," *J. Inform. dan Sist. Inf.*, vol. 1, no. 2, pp. 573–577, 2020.

[13] Pristiyono, M. Ritonga, M. A. Al Ihsan, A. Anjar, and F. H. Rambe, "Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 012045, 2021, doi: 10.1088/1757-899x/1088/1/012045.

[14] B. C. Nevianing, E. P. Mandyartha, and B. Nugroho, "DIAGNOSIS ATTENTION DEFICIT HYPERACTIVITY DISORDER ( ADHD ) BERDASARKAN DSM-IV MENGGUNAKAN ALGORITMA NAIVE BAYES," *J. Inform. dan Sist. Inf.*, vol. 2, no. 1, pp. 1–8, 2021.

[15] A. F. Rahma, Agussalim, and D. S. Y. Kartika, "Analisis Sentimen Hashtag Kuliner Di Indonesia Menggunakan Naive Bayes," *J. Inform. dan Sist. Inf.*, vol. 2, no. 1, pp. 19–25, 2021.