

Binary Number Classification Using Naïve Bayes for Digit Recognition

Klasifikasi Angka Biner dengan Metode Naive Bayes untuk Pengenalan Angka

Mirza Ardiana¹, Putri Nur Rahayu², Thomas Brian³, Indri ika Widyastuti⁴, Sekarsari Wibowo⁵

^{1,2}Prodi Manajemen Bisnis, Politeknik Perkapalan Negeri Surabaya

³Prodi Teknik Kelistrikan Kapal, Politeknik Perkapalan Negeri Surabaya

⁴Prodi Teknik Bangunan Kapal, Politeknik Perkapalan Negeri Surabaya

⁵Prodi Pengolahan Limbah, Politeknik Perkapalan Negeri Surabaya

E-mail: *¹mirzaardiana@ppns.ac.id, ²putri.nur@ppns.ac.id, ³thomasbrian@ppns.ac.id,
⁴indriwidyastuti@ppns.ac.id, ⁵sekarsari.wibowo@ppns.ac.id

Abstract – Automatic digit recognition is a crucial application in digital image processing and artificial intelligence. This study aims to evaluate the implementation of the Naïve Bayes method in classifying digits from 0 to 9 by assessing its effectiveness in recognizing binary digits based on a prepared dataset. Based on the test results, the Naïve Bayes model demonstrated a low accuracy of 30.0%, indicating that this method is less effective in handling the given digit dataset. Confusion matrix analysis revealed that certain digits, such as 6, 7, and 9, were entirely misclassified by the model, with an error rate of 100%. Digits 0 and 1 exhibited lower error rates; however, the model still struggled to classify other digits accurately. The error rate graph displayed varying misclassification patterns among digits, with certain numbers—especially those with similar shapes—being more challenging to distinguish. Although Naïve Bayes can be applied to digit recognition tasks, these results highlight the need for further improvements, such as using more relevant features or adopting more advanced classification methods, such as Support Vector Machines (SVM) or deep learning, to enhance classification accuracy. This study provides insights into the limitations of the Naïve Bayes method and serves as a foundation for further research exploring more effective classification techniques. Additionally, it encourages the development of more efficient models for real-world applications, such as handwriting recognition, biometric security systems, and document digitization.

Keywords — artificial intelligence, digit classification, naïve bayes

Abstrak – Pengenalan angka secara otomatis merupakan aplikasi penting dalam bidang pengolahan citra digital dan kecerdasan buatan. Penelitian ini bertujuan untuk mengevaluasi penerapan metode Naïve Bayes dalam klasifikasi angka 0 hingga 9, dengan menguji efektivitasnya dalam mengenali angka biner berdasarkan *dataset* yang telah disiapkan. Berdasarkan hasil pengujian, model Naïve Bayes menunjukkan akurasi rendah sebesar 30.0%, yang mengindikasikan bahwa metode ini kurang efektif dalam menangani *dataset* angka yang digunakan. Analisis matriks konfusi mengungkapkan bahwa beberapa angka, seperti 6, 7, dan 9, tidak dapat dikenali dengan benar oleh model, dengan tingkat kesalahan mencapai 100%. Angka 0 dan 1 memiliki tingkat kesalahan yang lebih rendah, namun model tetap kesulitan dalam mengklasifikasikan angka-angka lainnya secara akurat. Grafik *error rate* memperlihatkan pola kesalahan yang bervariasi antar angka, dengan angka-angka tertentu, terutama yang memiliki bentuk serupa, lebih sulit dibedakan. Meskipun Naïve Bayes dapat diterapkan dalam tugas pengenalan angka, hasil ini menunjukkan perlunya perbaikan lebih lanjut, seperti penggunaan fitur yang lebih relevan atau penerapan metode klasifikasi yang lebih canggih, seperti *Support Vector Machines* (SVM) atau *deep learning*, untuk meningkatkan akurasi klasifikasi angka. Penelitian ini memberikan wawasan mengenai keterbatasan metode Naïve Bayes dan dapat menjadi dasar untuk penelitian lebih lanjut yang mengeksplorasi teknik klasifikasi lainnya yang lebih

efektif, serta dapat mendorong pengembangan model yang lebih efisien dalam aplikasi nyata, seperti pengenalan tulisan tangan, sistem keamanan biometrik, dan digitalisasi dokumen.

Kata Kunci — kecerdasan buatan, klasifikasi angka, *naïve bayes*

1. PENDAHULUAN

Pengenalan angka secara otomatis merupakan bagian dari kajian dalam pengolahan citra digital dan kecerdasan buatan yang bertujuan untuk mengekstraksi, menganalisis, dan menginterpretasikan pola numerik dari berbagai bentuk visual, seperti tulisan tangan, cetakan, atau tampilan digital. Dalam pengolahan citra, proses ini mencakup tahapan seperti *preprocessing*, segmentasi, ekstraksi fitur, dan klasifikasi untuk meningkatkan akurasi identifikasi angka dalam berbagai kondisi. Sementara itu, di bidang kecerdasan buatan, berbagai model pembelajaran mesin, seperti *Naïve Bayes*, Jaringan Saraf Tiruan (ANN), dan *Deep Learning*, diterapkan guna mengoptimalkan kinerja sistem pengenalan angka secara otomatis. Teknologi ini memiliki berbagai penerapan strategis, antara lain dalam digitalisasi dokumen melalui pengenalan tulisan tangan, pemrosesan cek dan faktur dalam sektor keuangan, serta sistem verifikasi identitas berbasis biometrik. Seiring dengan perkembangan teknologi digital yang pesat, permintaan terhadap sistem pengenalan angka dengan tingkat akurasi yang tinggi semakin meningkat, terutama di sektor keuangan, pendidikan, keamanan, dan administrasi [1].

Salah satu pendekatan yang umum digunakan dalam klasifikasi angka adalah *Naïve Bayes*. Metode *Naïve Bayes* adalah salah satu teknik yang sering digunakan dalam klasifikasi karena pendekatannya yang berbasis probabilitas. Teorema *Naïve Bayes* memungkinkan pengambilan keputusan berdasarkan probabilitas posterior yang diperbarui sesuai dengan data yang diamati [2]. Dalam konteks klasifikasi angka biner, metode *Naïve Bayes* dapat membantu menentukan probabilitas suatu angka termasuk dalam salah satu dari dua kategori berdasarkan fitur yang diekstrak dari data.

Penelitian dalam bidang pengenalan angka umumnya menggunakan berbagai metode kecerdasan buatan, seperti *Support Vector Machine (SVM)*, *K-Nearest Neighbors (KNN)*, dan Jaringan Saraf Tiruan [3]. Namun, metode *Naïve Bayes* tetap menjadi salah satu alternatif yang menarik karena kemudahannya dalam implementasi dan interpretasi hasil klasifikasi. Selain itu, metode ini tidak memerlukan banyak parameter dan dapat diterapkan pada dataset dengan dimensi tinggi tanpa mengalami *overfitting* yang signifikan.

Beberapa penelitian telah menunjukkan efektivitas metode *Naïve Bayes* dalam berbagai tugas klasifikasi, termasuk pengenalan tulisan tangan dan klasifikasi karakter optik. Keunggulan utama metode ini adalah kemampuannya untuk bekerja dengan dataset kecil serta interpretabilitas yang tinggi. Namun, tantangan utama dalam penggunaannya adalah asumsi independensi antar fitur dalam model *Naïve Bayes* yang dapat mempengaruhi akurasi klasifikasi dalam kasus tertentu [4].

Oleh karena itu, penelitian ini berfokus pada penerapan metode *Naïve Bayes* dalam mengklasifikasikan angka biner guna mengevaluasi efektivitasnya. Dengan melakukan evaluasi terhadap teknik *preprocessing* dan jumlah data latih yang digunakan, penelitian ini diharapkan dapat memberikan pemahaman yang lebih mendalam terkait potensi serta keterbatasan metode *Naïve Bayes* dalam pengenalan angka biner.

2. METODE PENELITIAN

Metode *Naïve Bayes* digunakan untuk menghitung probabilitas secara statistik [5]. Metode ini biasa digunakan dalam metode data *mining* untuk mengklasifikasikan data. Klasifikasi data dapat digunakan untuk menghitung peluang dari satu kelas masing-masing kelompok dan menentukan kelompok yang optimal [6], [7]. Contoh rumus *Naïve Bayes* dapat dilihat pada Persamaan 1 dibawah ini.

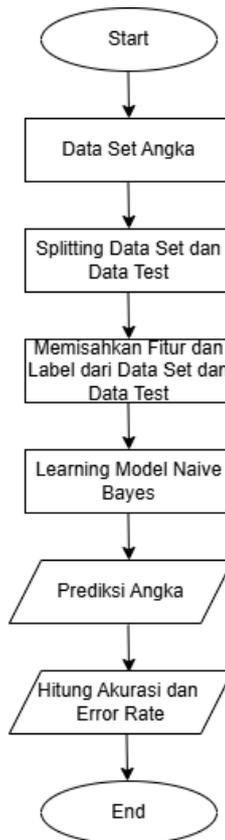
$$P(H|X) = \frac{P(H) P(X|H)}{P(X)} \dots\dots\dots (1)$$

Dimana X adalah data dengan kelas yang belum diketahui, H adalah Hipotesis data X merupakan kelas spesifik, $P(H|X)$ adalah probabilitas hipotesis H berdasar kondisi X , $P(H)$ adalah

probabilitas hipotesis H (*prior probability*), $P(X/H)$ adalah Probabilitas X berdasarkan kondisi pada kondisi pada hipotesis H , $P(X)$ adalah probabilitas dari X [6].

Metode *Naïve Bayes* bisa digunakan untuk probabilitas yang lebih kompleks bila dibandingkan dengan metode *classifier* lainnya [8], dikarenakan pada penelitian ini karakter dari biner angka memiliki sifat yang *dependent* dan tidak bebas antara variable satu dengan yang lainnya seperti di *Naïve Bayes* (bersifat *independent*) [9], [10]. Algoritma *Naïve Bayes* sangat iteratif untuk menemukan hiperparameter data. Metode ini menentukan nilai hiperparameter dari beberapa iterasi [11], [12]. Cara pemodelan metode *Naïve Bayes* sangat banyak digunakan untuk memecahkan masalah pada kehidupan sehari-hari dengan berbagai persoalan [13].

Gambar 1 berikut adalah alur dari penelitian kali ini, yaitu sebagai berikut :



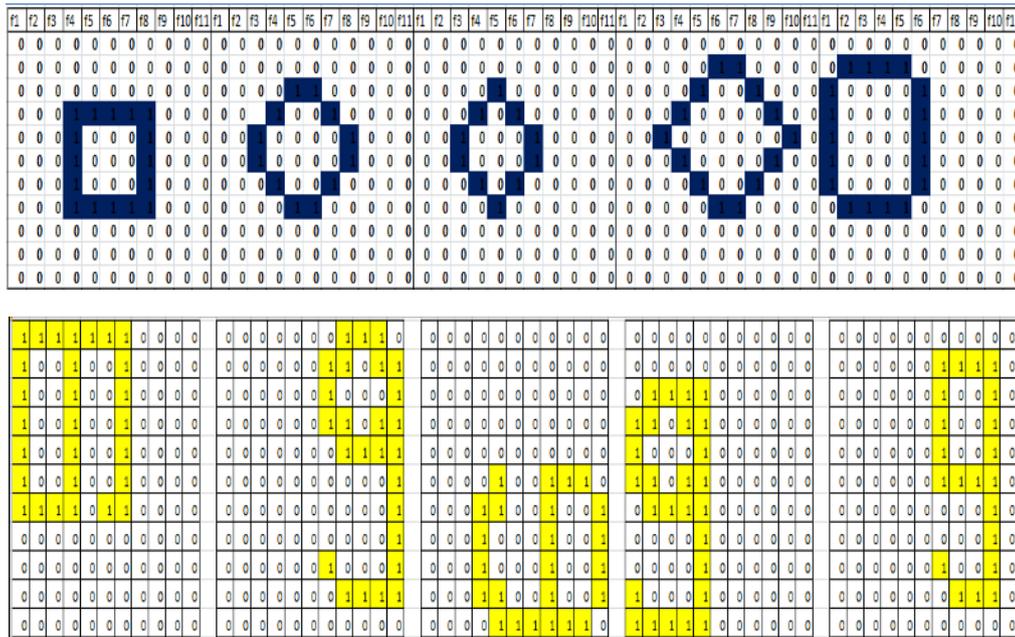
Gambar 1 Flowchart Penelitian

Flowchart pada Gambar 1 tersebut menggambarkan tahapan utama dalam proses pengenalan digit menggunakan algoritma *Naïve Bayes*. Proses dimulai dengan membaca *dataset* angka, yang terdiri dari data pelatihan dan data pengujian. Setelah itu, *dataset* dibagi menjadi dua bagian: data latih (*training set*) dan data uji (*test set*), yang kemudian dipisahkan menjadi fitur (representasi angka dalam bentuk numerik) dan label (kelas angka 0-9). Pemisahan ini bertujuan untuk membangun model klasifikasi berdasarkan data pelatihan yang diberikan.

Setelah tahap pemrosesan data, model *Naïve Bayes* dilatih menggunakan data latih untuk mempelajari pola angka berdasarkan fitur yang tersedia. Model kemudian digunakan untuk melakukan prediksi angka pada data uji, dan hasilnya dibandingkan dengan label yang sebenarnya. Akurasi model dihitung dengan membandingkan jumlah prediksi yang benar terhadap total data uji, serta dihitung juga error rate untuk mengevaluasi tingkat kesalahan model. Terakhir, sistem menampilkan hasil evaluasi, termasuk tingkat akurasi dan confusion matrix, untuk menganalisis performa model dalam mengenali digit 0 hingga 9.

3.2.2. Test set

Selain *Training Set*, bagian lain dari *dataset* yang digunakan adalah *Test Set*, yang berfungsi untuk menguji kemampuan prediksi model setelah melalui proses pelatihan. Dalam penelitian ini, data uji terdiri dari angka 0 hingga 9, di mana setiap angka memiliki lima jenis sampel yang berbeda dari data yang digunakan dalam pelatihan. Sama seperti *training set*, setiap sampel direpresentasikan dalam bentuk *grid* berukuran 11×11 . Penggunaan *test set* ini bertujuan untuk mengevaluasi sejauh mana model *Naïve Bayes* dapat menggeneralisasi pola dari data yang belum pernah dilihat sebelumnya seperti pada Gambar 4.



Gambar 4 Contoh Data Tes Penelitian

3.3. Fase Pelatihan (Training)

Dalam proses klasifikasi menggunakan *Naïve Bayes*, langkah penting yang dilakukan adalah ekstraksi fitur. Setiap gambar dianggap sebagai vektor fitur, di mana fitur utama yang digunakan adalah intensitas piksel dari gambar tersebut. Dalam konteks *Naïve Bayes*, fitur ini akan direpresentasikan dalam bentuk distribusi probabilitas berdasarkan kelas angka yang ada pada gambar.

Setelah fitur diekstraksi, tahap berikutnya adalah estimasi probabilitas, yang terdiri dari dua komponen utama:

- Probabilitas *Prior*, yaitu estimasi probabilitas awal dari setiap kelas (angka 0 hingga 9). Probabilitas ini dihitung berdasarkan proporsi kemunculan setiap angka dalam dataset pelatihan.
- Probabilitas *Likelihood*, yaitu estimasi probabilitas untuk setiap fitur (piksel) yang terkait dengan kelas tertentu. Probabilitas ini diperoleh dengan menganalisis distribusi data fitur untuk setiap kelas. Pada tahap ini, dilakukan perhitungan *mean* dan standar deviasi untuk setiap fitur guna menentukan pola distribusi dalam dataset.

Tahap terakhir adalah model pembelajaran, dimana model *Naïve Bayes* dibangun dengan menggunakan probabilitas kelas yang telah dihitung sebelumnya. Model ini kemudian digunakan untuk mengklasifikasikan input gambar baru berdasarkan probabilitas yang paling sesuai dengan data pelatihan.

3.4. Fase Prediksi (Testing)

Dalam tahap pengujian model, data uji mengalami proses ekstraksi fitur yang serupa dengan data pelatihan. Gambar yang diuji akan melalui proses *flattening* untuk mengubahnya menjadi vektor

fitur serta normalisasi agar nilai piksel berada dalam rentang [0,1]. Hal ini memastikan bahwa data uji memiliki format yang sama dengan data pelatihan, sehingga dapat dianalisis oleh model *Naïve Bayes*.

Selanjutnya, dilakukan perhitungan probabilitas kelas, yang melibatkan beberapa langkah utama antara lain:

- a. Probabilitas *Posterior*, yang dihitung menggunakan Teorema *Naïve Bayes* dengan Persaman 2:

$$P(\text{Class}|\text{Feature}) = \frac{P(\text{Features}|\text{Class}) \cdot P(\text{Class})}{P(\text{Feature})} \dots\dots\dots (2)$$

- b. Di sini, $P(\text{Class})$ adalah *prior* (probabilitas kelas), dan $P(\text{Features} | \text{Class})$ adalah *likelihood* yang dihitung selama fase pelatihan.
- c. Perhitungan *Likelihood* untuk Setiap Fitur: Setiap fitur (*pixel*) dihitung dengan distribusi yang telah dipelajari pada fase pelatihan.

Tahap terakhir adalah klasifikasi, dimana setelah menghitung probabilitas *posterior* untuk semua kelas, *Naïve Bayes* akan memilih kelas dengan probabilitas tertinggi sebagai hasil prediksi. Dengan cara ini, model dapat menentukan angka yang paling mungkin sesuai dengan pola dalam gambar uji.

3.5. Evaluasi Hasil

Untuk mengevaluasi kinerja model *Naïve Bayes* pada data uji, digunakan *Confusion Matrix*. Matriks ini menampilkan jumlah prediksi yang benar dan salah untuk setiap kelas, memungkinkan analisis lebih mendalam terhadap performa model. Dengan matriks kebingungan, dapat diketahui apakah model cenderung mengalami kesalahan dalam mengklasifikasikan angka tertentu dan seberapa baik model membedakan antar kelas.

Selain itu, kinerja model juga diukur menggunakan akurasi, yang dihitung dengan menggunakan Persamaan 3:

$$\text{Akurasi} = \frac{\text{Jumlah Prediksi Benar}}{\text{Jumlah Total Data Uji}} \dots\dots\dots (3)$$

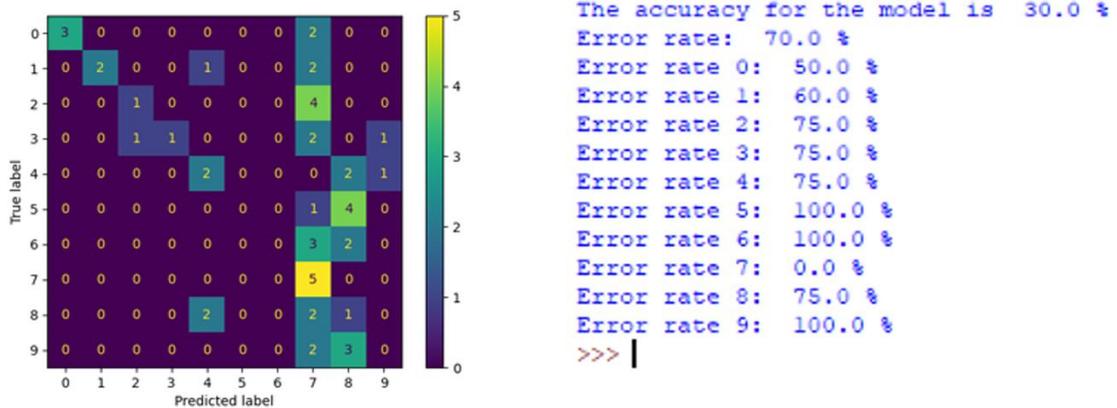
Akurasi memberikan gambaran tentang seberapa sering model membuat prediksi yang benar dibandingkan dengan keseluruhan data uji. Semakin tinggi nilai akurasi, semakin baik model dalam mengenali angka berdasarkan pola yang telah dipelajari selama pelatihan.

3.6. Output Prediksi

Dalam tahap prediksi akhir, model *Naïve Bayes* akan menentukan angka yang paling mungkin untuk setiap gambar uji berdasarkan kelas dengan probabilitas tertinggi. Prediksi ini dihasilkan setelah menghitung probabilitas posterior untuk semua kelas, lalu memilih kelas dengan nilai probabilitas terbesar sebagai hasil akhir.

Untuk lebih memahami hasil yang diperoleh, dilakukan visualisasi atau penyajian hasil. Prediksi yang dihasilkan model dapat disajikan dalam bentuk teks, yang menunjukkan angka yang dikenali oleh model. Selain itu, hasil prediksi juga dapat divisualisasikan dengan menampilkan gambar input beserta angka hasil klasifikasinya. Dengan cara ini, dapat dilakukan analisis terhadap gambar yang berhasil diklasifikasikan dengan benar maupun yang mengalami kesalahan prediksi.

Dari proses yang telah dilakukan tersebut, berikut adalah hasil pengujian yang didapatkan, yaitu sebagaimana disajikan oleh Gambar 5 berikut:

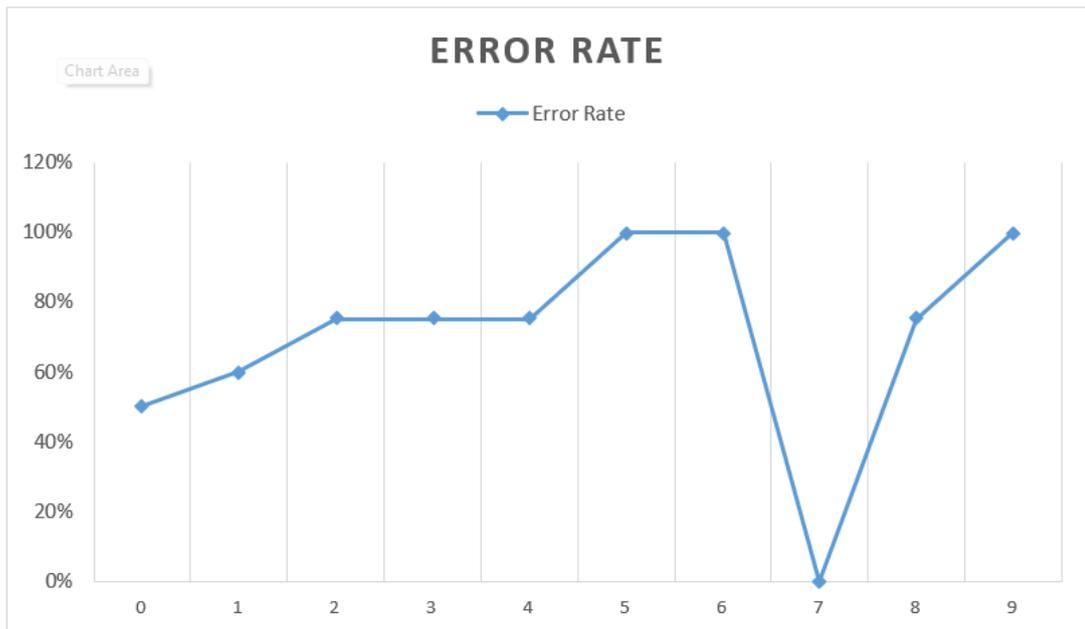


Gambar 5 Hasil *Confusion Matrix*

Berdasarkan hasil pengujian dengan metode Naïve Bayes untuk mengenali angka 0 hingga 9, model ini menunjukkan akurasi sebesar 30.0%. Hal ini berarti model hanya mampu mengenali sekitar 30% dari angka yang diuji dengan benar, sementara sisanya diklasifikasikan secara keliru. Akurasi yang rendah ini mengindikasikan bahwa metode Naïve Bayes mungkin kurang efektif dalam menangani dataset angka yang digunakan.

Matriks konfusi yang ditampilkan menunjukkan bahwa banyak angka yang salah diklasifikasikan ke angka lain. Beberapa digit memiliki tingkat kesalahan yang lebih tinggi dibandingkan digit lainnya. Dari hasil pengujian, angka 6, 7, dan 9 memiliki tingkat error 100%, yang berarti model gagal mengenali angka-angka ini dengan benar sama sekali. Sementara itu, angka 0 dan 1 memiliki error rate lebih rendah, masing-masing 50% dan 60%, sehingga lebih mudah dikenali dibandingkan yang lain. Angka lainnya, seperti 2, 3, 4, dan 8, memiliki error rate sekitar 75%, menunjukkan bahwa model juga mengalami kesulitan dalam membedakan angka-angka ini.

Gambar 6 menunjukkan grafik hasil *Error Rate* dari penelitian yang dilakukan. Grafik ini menampilkan tingkat kesalahan model dalam mengklasifikasikan angka dari dataset uji.



Gambar 6 Grafik Hasil *Error Rate* Penelitian

Grafik *error rate* yang ditunjukkan oleh Gambar 6 memperlihatkan pola perubahan tingkat kesalahan untuk setiap angka dari 0 hingga 9. Terdapat tren peningkatan *error rate* dari angka 0 hingga

sekitar angka 6, lalu terjadi penurunan drastis pada angka 7, sebelum kembali meningkat pada angka 9. Pola ini mencerminkan bahwa beberapa angka memiliki fitur yang lebih sulit dibedakan oleh metode *Naïve Bayes*, terutama angka yang memiliki bentuk serupa dengan angka lainnya dalam *dataset*.

Secara keseluruhan, model *Naïve Bayes* yang digunakan dalam pengujian ini masih memiliki performa yang kurang baik dalam mengenali angka. Kesalahan klasifikasi yang tinggi menunjukkan bahwa metode ini mungkin tidak cukup efektif untuk tugas pengenalan angka, terutama pada angka yang bentuknya lebih kompleks. Untuk meningkatkan kinerja model, perlu dilakukan perbaikan, seperti menggunakan fitur yang lebih relevan atau menerapkan metode klasifikasi lain yang lebih canggih, seperti *Support Vector Machine* (SVM) atau *Deep Learning*.

4. KESIMPULAN

Berdasarkan hasil pengujian, model *Naïve Bayes* untuk mengenali angka 0 hingga 9 menunjukkan akurasi yang rendah sebesar 30.0%, dengan kesalahan klasifikasi yang tinggi. Beberapa angka, seperti 6, 7, dan 9, tidak dapat dikenali dengan benar oleh model, yang mencerminkan kesulitan dalam membedakan angka dengan bentuk serupa. Meskipun angka 0 dan 1 memiliki error rate yang lebih rendah, model ini masih mengalami kesulitan dalam mengklasifikasikan angka-angka lainnya dengan akurat. Pola *error rate* yang terlihat juga menunjukkan bahwa angka-angka tertentu lebih sulit dibedakan daripada yang lainnya.

Keseluruhan hasil pengujian ini menunjukkan bahwa *Naïve Bayes* tidak cukup efektif untuk tugas pengenalan angka, terutama untuk angka yang memiliki kesamaan visual atau struktur kompleks. Oleh karena itu, untuk meningkatkan performa model, diperlukan upaya lebih lanjut, seperti menggunakan fitur yang lebih relevan atau beralih ke metode klasifikasi yang lebih canggih seperti *SVM* atau *Deep Learning*, yang dapat menangani kesulitan dalam membedakan angka yang serupa dan memberikan hasil yang lebih akurat.

DAFTAR PUSTAKA

- [1] Jain, A. K., & Zongker, D. (2014). Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2), 153-158.
- [2] Jain, A. K., Duin, R. P. W., & Mao, J. (2015). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 4-37.
- [3] King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, 9(2), 137-163.
- [4] Meertens, Q. A., Diks, C. G. H., van den Herik, H. J., & Takes, F. W. (2019). A Bayesian Approach for Accurate Classification-Based Aggregates. *arXiv preprint arXiv:1902.02412*.
- [5] Heliyanti Susana, "Penerapan Model Klasifikasi Metode Naive Bayes Terhadap Penggunaan Akses Internet," *J. Ris. Sist. Inf. dan Teknol. Inf.*, vol. 4, no. 1, pp. 1-8, 2022, doi: 10.52005/jursistekni.v4i1.96.
- [6] A. Tangkelayuk, "The Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes, dan Decision Tree," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 2, pp. 1109-1119, 2022, doi: 10.35957/jatisi.v9i2.2048.
- [7] I. P. Sari, L. Elvitaria, and Izah Yoelanda, "Metode Bayesian Network Untuk Menentukan Probabilitas Indikasi Gangguan Bipolar," *JEKIN - J. Tek. Inform.*, vol. 4, no. 2, pp. 196-208, 2024, doi: 10.58794/jekin.v4i2.717.



-
- [8] D. Erwanto, P. N. Rahayu, dan Y. B. Utomo, “Klasifikasi Cacat Pada Kaleng Kemasan Menggunakan Metode Lacunarity Dan Naïve Bayes,” *Electro Luceat*, vol. 7, no. 2, hlm. 142–150, 2021.
- [9] S. T. A. Yosua, S. Verra, dkk., “Penggunaan Metode Naïve Bayes Dalam Mengklasifikasi Pengangguran Pada Desa Bojong Kulur”, *Biaglala Informatika*, Vol. 10 No. 1 – Tahun 2022.
- [10] R. Arifuddin, D. S. Utami, Erfiani, “Effectiveness of Machine Learning Models with Bayesian Optimization-Based Method to Identify Important Variables that Affect GPA”, *JTAM (Jurnal Teori dan Aplikasi Matematika)*, Vol. 8, No. 3, Juli 2024.
- [11] Syaharuddin, “Integration Of Bayesian Methods In Machine Learning A Theoretical And Empirical Review”, *INSERT: Information System and Emerging Technology Journal*. Vol. 5, No. 2, Desember 2024
- [12] F. Nur, E. P. Enda, S. Rina, “Optimasi Hyperparameter pada Neural Network (Studi Kasus Identifikasi Komentar Cyberbullying Instagram)”, *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, Vol. 9 No. 2 Agustus 2023
- [13] P. P. Yudha, Aradea, Rianto, “Pemodelan Bayesian Network untuk Menentukan Probabilitas Penyebab Pelanggaran dalam Pertandingan Karate Kumite”, *INNOVATION IN RESEARCH OF INFORMATICS - VOL. 3 NO. 2 (2021)* 54-57

