Implementasi Algoritma Naive Bayes Untuk Klasifikasi Data Berbasis Pendekatan Probabilistik Matematis

Jhoanne Fredricka[1], Nofi Qurniati[2], Desi Mahdalena[3]

[1,3] *Program Studi Informatika – Universitas Dehasen Bengkulu*

[2] *Program Studi Pendidikan Profesi Guru - Universitas Dehasen Bengkulu*

Jl. Meranti Raya No. 32 Kota Bengkulu 38228- Bengkulu - Indonesia

fredrickajhoanne@gmail.com, nofi.qurniati@gmail.com, desimahdalena08@unived.ac.id

Abstrak - Perkembangan teknologi informasi menyebabkan peningkatan volume data yang sangat besar sehingga dibutuhkan metode klasifikasi yang mampu menghasilkan informasi secara cepat dan akurat. Penelitian ini bertujuan untuk mengimplementasikan algoritma Naive Bayes dalam proses klasifikasi data berbasis pendekatan probabilistik matematis. Metode penelitian menggunakan pendekatan kuantitatif eksperimental dengan tahapan pengumpulan dataset, preprocessing data, pembagian data training dan testing menggunakan rasio 80:20, implementasi algoritma menggunakan Python pada Google Colab, dan evaluasi model menggunakan confusion matrix, accuracy, precision, dan recall. Dataset terdiri dari atribut usia, pendapatan, status mahasiswa, dan keputusan pembelian. Hasil penelitian menunjukkan bahwa algoritma Naive Bayes menghasilkan nilai accuracy sebesar 85%, precision sebesar 90,9%, dan recall sebesar 83,3%. Pendekatan matematis diterapkan melalui perhitungan probabilitas posterior berdasarkan Teorema Bayes untuk melakukan penentuan kelas data. Dengan demikian, algoritma Naive Bayes terbukti efektif dalam meningkatkan performa klasifikasi data dan mendukung pengambilan keputusan secara lebih akurat dan efisien.

Kata Kunci— Naive Bayes, Data Mining, Klasifikasi Data, Probabilitas.

Abstract - The development of information technology has caused a very large increase in the volume of data, so classification methods are needed that are able to produce information quickly and accurately. This research aims to implement the Naïve Bayes algorithm in a data classification process based on a mathematical probabilistic approach. The research method uses an experimental quantitative approach with stages of data collection, data preprocessing, dividing training and testing data using a ratio of 80:20, implementing algorithms using Python on Google Colab, and evaluating models using confusion, accuracy, precision and recall matrices. The dataset consists of the attributes age, income, student status, and purchasing decisions. The research results show that the Naive Bayes algorithm produces an accuracy value of 85%, precision of 90.9%, and recall of 83.3%. A mathematical approach is applied through posterior probability calculations based on Bayes' Theorem to determine data classes. Thus, the Naive Bayes algorithm is proven to be effective in improving data classification performance and supporting more accurate and efficient decision making.

Keywords— *Naive Bayes, Data Mining, Data Classification, Probability.*

I. PENDAHULUAN

Perkembangan teknologi informasi telah menghasilkan pertumbuhan data yang sangat besar pada berbagai sektor seperti pendidikan, bisnis, kesehatan, dan industri [1]. Peningkatan volume data tersebut memerlukan metode pengolahan yang mampu mengekstraksi informasi penting secara cepat dan akurat. Tanpa metode analisis yang tepat, data hanya menjadi kumpulan informasi yang sulit dimanfaatkan dalam pengambilan keputusan.

Data mining merupakan salah satu teknik yang digunakan untuk menemukan pola, hubungan, dan informasi penting dari data dalam jumlah besar menggunakan pendekatan statistik, matematika, dan kecerdasan buatan [2]. Salah satu teknik utama dalam data mining adalah klasifikasi, yaitu proses pengelompokan data berdasarkan karakteristik tertentu [3].

Permasalahan yang sering terjadi pada proses klasifikasi adalah rendahnya akurasi model serta tingginya kompleksitas komputasi. Oleh karena itu, diperlukan algoritma yang mampu melakukan klasifikasi dengan cepat dan efisien. Algoritma Naive Bayes menjadi salah satu metode yang banyak digunakan karena memiliki performa klasifikasi yang baik dengan pendekatan probabilitas matematis [4].

Penelitian sebelumnya oleh Gholib dan Hamizan [5] menunjukkan bahwa Naive Bayes mampu meningkatkan akurasi klasifikasi data penjualan. Penelitian Saputra dkk. [6] berhasil menerapkan Naive Bayes untuk prediksi kelulusan mahasiswa dengan tingkat akurasi yang tinggi. Iwandini dan Agung [7] juga menyatakan bahwa algoritma Naive Bayes efektif digunakan pada klasifikasi dataset pendidikan. Selain itu, penelitian Rahman dan Maulan [8] menunjukkan bahwa optimasi feature selection dapat meningkatkan performa algoritma Naive Bayes.

Berdasarkan penelitian terdahulu, masih diperlukan pembahasan yang lebih detail terkait penerapan pendekatan matematis pada algoritma Naive Bayes dalam proses klasifikasi data[9]. Oleh karena itu, penelitian ini bertujuan untuk mengimplementasikan algoritma Naive Bayes berbasis pendekatan probabilistik matematis dan mengevaluasi performanya menggunakan confusion matrix, accuracy, precision, dan recall [10] .

II. METODE PENELITIAN

Penelitian ini menggunakan metode kuantitatif dengan pendekatan eksperimen. Dataset yang digunakan berasal dari dataset publik Kaggle mengenai klasifikasi pelanggan berdasarkan keputusan pembelian produk. Dataset terdiri dari 100 data dengan atribut usia, pendapatan, status mahasiswa, dan keputusan pembelian.

A. Pengumpulan Data

Data diperoleh dari platform Kaggle kemudian diolah menggunakan Google Colab. Dataset dibagi menjadi data training sebanyak 80 data dan data testing sebanyak 20 data menggunakan metode split 80:20.

B. Preparasi Data

Proses preprocessing dilakukan untuk meningkatkan kualitas data sebelum proses klasifikasi. Tahapan preprocessing meliputi:

1. Data Cleaning: menghapus data kosong dan data duplikat.
2. Data Transformation: mengubah atribut kategorikal menjadi bentuk numerik
3. Data Normalization: menyetarakan skala data untuk mengurangi bias perhitungan.

Sebelum preprocessing ditemukan beberapa data kosong pada atribut pendapatan dan status mahasiswa. Setelah preprocessing seluruh data berhasil dibersihkan dan siap digunakan dalam proses klasifikasi.

C. Implementasi Algoritma

Algoritma Naive Bayes menggunakan pendekatan probabilitas berdasarkan Teorema Bayes dengan rumus [10]:

$$P(C|X) = (P(X|C) \times P(C))/P(X)$$

Keterangan:
P(C|X) : Probabilitas posterior
P(X|C) : Probabilitas likelihood
P(C) : Probabilitas prior
P(X) : Probabilitas evidence

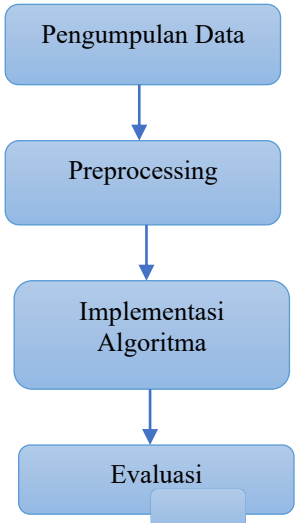
Pendekatan matematis diterapkan pada proses perhitungan probabilitas posterior untuk menentukan kemungkinan suatu data termasuk ke dalam kelas tertentu berdasarkan atribut yang dimiliki.

D. Evaluasi Model

Evaluasi model dilakukan menggunakan confusion matrix, accuracy, precision, dan recall dengan bantuan library Scikit-learn, Pandas, dan NumPy pada Google Colab.

E. Alur Penelitian

Adapun alur penelitian dapat di lihat pada gambar 1 di bawah ini:



Gbr 1. Alur Penelitian

Alur ini dirancang agar proses penelitian berjalan terstruktur dan menghasilkan model klasifikasi yang optimal berbasis pendekatan matematis.

III. HASIL DAN PEMBAHASAN

A. Deskripsi Dataset

Dataset penelitian terdiri dari 100 data pelanggan dengan atribut usia, pendapatan, status mahasiswa, dan keputusan pembelian. Contoh dataset ditampilkan pada Tabel 1.

TABEL 1. ATRIBUT DATA

Usia	Pendapatan	Mahasiswa	Keputusan
Muda	Tinggi	Ya	Layak
Dewasa	Sedang	Tidak	Tidak Layak
Tua	Rendah	Ya	Layak

B. Hasil Pengujian Model

Implementasi algoritma Naïve Bayes dilakukan menggunakan bahasa pemrograman Python pada platform Google Colab dengan bantuan library Scikit-learn, Pandas, dan NumPy. Setelah proses training dan testing dilakukan menggunakan pembagian data 80:20, diperoleh hasil confusion matrix seperti pada Gambar 2.

```
Confusion Matrix Output (Google Colab)

[[10  2]
 [ 1  7]]
```

Gbr2. hasil confusion matrix

Keterangan:

- True Positive (TP) = 10
- False Negative (FN) = 2
- False Positive (FP) = 1
- True Negative (TN) = 7

Berdasarkan hasil confusion matrix tersebut, diperoleh nilai evaluasi model sebagai berikut:

```
</> Python

Accuracy  : 85.0%
Precision : 90.9%
Recall    : 83.3%
```

Gbr3. Nilai evaluasi model

Hasil pengujian menunjukkan bahwa algoritma Naive Bayes memiliki performa klasifikasi yang cukup baik. Nilai accuracy sebesar 85% menunjukkan bahwa model mampu melakukan klasifikasi data dengan tingkat ketepatan yang tinggi. Selain itu, nilai precision sebesar 90,9% menunjukkan bahwa model mampu memberikan prediksi positif yang akurat, sedangkan recall sebesar 83,3% menunjukkan bahwa model cukup baik dalam mengidentifikasi data yang termasuk ke dalam kelas positif.

C. Pembahasan Pendekatan Matematis

Pendekatan matematis pada algoritma Naive Bayes diterapkan melalui perhitungan probabilitas bersyarat menggunakan Teorema Bayes. Setiap atribut pada data dihitung probabilitas kemunculannya terhadap masing-masing kelas. Kelas dengan probabilitas terbesar akan menjadi hasil prediksi akhir.

Apabila data memiliki atribut usia muda, pendapatan tinggi, dan status mahasiswa aktif, maka probabilitas setiap atribut terhadap kelas Layak dan Tidak Layak dihitung menggunakan probabilitas posterior. Pendekatan ini memungkinkan proses klasifikasi dilakukan secara sistematis dan terukur. Selain itu, penggunaan probabilitas matematis membuat algoritma Naive Bayes lebih efisien dalam proses komputasi dibandingkan algoritma klasifikasi lainnya. Hal ini sesuai dengan penelitian sebelumnya yang menyatakan bahwa Naive Bayes memiliki performa yang baik pada data berukuran kecil hingga menengah [6][8].

IV. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil penelitian, algoritma Naive Bayes berhasil diimplementasikan untuk proses klasifikasi data berbasis pendekatan probabilistik matematis. Hasil pengujian menunjukkan nilai accuracy sebesar 85%, precision sebesar 90,9%, dan recall sebesar 83,3%. Pendekatan matematis melalui Teorema Bayes terbukti mampu meningkatkan ketepatan klasifikasi data.

B. Saran

Berdasarkan hasil penelitian yang diperoleh, beberapa saran yang dapat diberikan untuk pengembangan penelitian selanjutnya adalah sebagai berikut:

1. Perlu dilakukan perbandingan dengan algoritma klasifikasi lainnya, seperti Support Vector Machine atau Decision Tree, untuk mengetahui metode yang paling optimal.
2. Penggunaan teknik feature selection dan optimasi parameter dapat dilakukan untuk meningkatkan akurasi model.
3. Penelitian selanjutnya juga dapat mengembangkan pendekatan hybrid atau kombinasi beberapa algoritma untuk memperoleh hasil klasifikasi yang lebih baik.

DAFTAR PUSTAKA

[1] F. D. Pratama, I. Zufria and Triase, “Implementasi data mining menggunakan algoritma naïve bayes untuk klasifikasi penerima program indonesia pintar”, *RABIT: Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 7, no. 1, pp. 77–84, 2022.

[2] S. D. R. Sitompul, “Optimasi Hyperparameter Naïve Bayes dalam Memprediksi Kanker Payudara Menggunakan GridSearch”, *Jurnal Penelitian Ilmu Komputer (JPILKOM)*, vol. 3, no. 1, pp. 17-26, 2025.

[3] R. Fajriah and D. Kurniawan, “Optimalisasi Model Klasifikasi Naive Bayes dan Support Vector Machine Dengan Fast Text dan Chi Square,” *Faktor Exacta*, vol. 17, no. 4, pp. 334–345, 2025, doi: 10.30998/faktorexacta.v17i4.24751.

[4] S. F. Tahir and C. A. Sugianto, “Optimasi Naive Bayes Menggunakan Algoritma Genetika Pada Klasifikasi Komentar Cyberbullying Pada Media Sosial X,” *Jurnal Informatika Dan Teknik Elektro Terapan*, vol. 12, no. 3, pp. 3350–3356, 2024. doi:

10.23960/jitet.v12i3.4834.

[5] H. Gholib, M. Yasa, A. Barana, S. A. Girsang, and S. A. Sobri, "Model Prediksi Harga Cabai Merah Besar Di Tingkat Pasar Tradisional Tahun 2017 - 2024 : Pendekatan Supervised Learning Berbasis Orange Data Mining," *Integrative Perspectives of Social and Science Journal*, vol. 2, no. 3, pp. 6823–6839, 2025.

[6] S. N. S. Sitinur and Z. Sitorus, "Implementasi Data Mining Untuk Clustering Produktivitas Bawang Merah Menggunakan Metode K-Means," *Jurnal Multimedia dan Teknologi Informasi (JATILIMA)*, vol. 7, no. 2, pp. 109–121, 2025, doi: 10.54209/jatilima.v7i02.1442..

[7] I. Iwandini, A. Triayudi, and G. Soepriyono, "Analisa Sentimen Pengguna Transportasi Jakarta Terhadap Transjakarta Menggunakan Metode Naives Bayes dan K-Nearest Neighbor," *JOSH: Journal of Science and Technology*, vol. 4, no. 2, pp. 543–550, 2023, doi: 10.47065/josh.v4i2.2937.

[8] M. A. Rahman, N. Hidayat, and A. A. Supianto, "Komparasi Metode Data Mining K-Nearest Neighbor Dengan Naïve Bayes Untuk Klasifikasi Kualitas Air Bersih (Studi Kasus PDAM Tirta Kencana Kabupaten Jombang)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK)*, vol. 2, no. 12, pp. 6346–6353, Dec. 2018.

[9] M. Y. Putra and D. I. Putri, "Pemanfaatan Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Kelas XI," *Jurnal Tekno Kompak*, vol. 16, no. 2, pp. 176–187, Aug. 2022, doi: 10.33365/jtk.v16i2.2002.

[10] R. D. Dongoran, S. F. Harahap, M. S. Tanjung, D. T. Safitri, and M. Siregar, "Aanalisa Probalitas dan Statistika dalam Pengambilan Keputusan Berbasis Data," *J. Comput. Sci. Inf. Technol. (JCoInT)*, vol. 6, no. 2, pp. 187–197, 2025.