Jurnal Teknologi Komputer dan Informasi (JUTEKINF)
Vol. 12, No. 02 Juli-Desember 2024, Hal. 152-157

Analisis Data Sampah Plastik Dunia pada Tahun 2023 dengan Metode Naive Bayes

Muhammad Satria Nugraha*[1], Imiel Ardhanenggar Tallane[2], Nabila Nur Fadhilah[3], Putri Citra Arrahma[4], Rifa Abdussalam[5], Anna Dina Kalifia[6]
[1,2,3,4,5,6]Universitas Teknologi Yogyakarta
[1,2,3,4,5,6]Program Studi Informatika, Universitas Teknologi Yogyakarta
*e-mail: msatrian76@gmail.com[1], imieltallane@gmail.com[2], nurfhadilahn@gmail.com[3], ptricitra15@gmail.com[4], sakaa@gmail.com[5], anna.dina.kalifia@staff.uty.ac.id[6].

Abstract

Plastic waste can have a significant impact on the environmental ecosystem, especially in coastal areas. This study aims to analyze and classify global plastic waste patterns using the Naïve Bayes algorithm. The dataset used is Global Plastic Waste 2023: Country-wise Data. The Naïve Bayes algorithm is applied to this dataset to assess its effectiveness in classifying the Global Plastic Waste 2023: Country-wise Data. The results of the study indicate that Naïve Bayes performs best in the High category, but in the Medium category, the model struggles with an F1-Score of 0.18 and Recall of only 0.11. Meanwhile, in the Very High category, the model completely fails to recognize data, with Precision, Recall, and F1-Score of 0.00 due to the very small amount of data. This study highlights the challenges of applying Naïve Bayes to datasets with uneven data distribution and suggests alternative methods to improve accuracy. These findings provide insights into the application of Naïve Bayes in global plastic waste analysis and can serve as a reference for developing more accurate models.

***Keywords**: Naïve Bayes, Classification, Data Balancing.*

Abstrak

Sampah pelastik dapat berdampak besar pada ekosistem lingkungan terutama lingkungan pesisir Pantai. Penelitian ini bertujuan untuk menganalisis dan mengklasifikasikan pola sampah plastik global dengan menerapkan algoritma Naïve Bayes. Dataset yang digunakan adalah *Global Plastic Waste 2023: Country-wise Data*. Algoritma Naïve Bayes diterapkan pada dataset ini untuk menilai efektivitasnya dalam mengklasifikasi dataset *Global Plastic Waste 2023: Country-wise Data.* Hasil penelitian menunjukkan bahwa Naïve Bayes memiliki performa terbaik pada kategori High, namun pada kategori Medium, model mengalami kesulitan dengan F1-Score 0.18 dan Recall hanya 0.11, sedangkan pada kategori Very High, model gagal sepenuhnya mengenali data Precision, Recall, F1-Score 0.00 akibat jumlah data yang sangat sedikit. Penelitian ini mengungkapkan tantangan dalam menerapkan Naïve Bayes pada dataset dengan distribusi data yang tidak merata dan menyarankan pendekatan metode lain untuk meningkatkan akurasi. Hasil ini memberikan wawasan tentang penerapan Naïve Bayes dalam analisis sampah plastik global dan dapat menjadi referensi untuk pengembangan model yang lebih akurat.

Kata Kunci: Naïve Bayes, Klassifikasi, Keseimbangan Data.

1. PENDAHULUAN

Data statistic menjadi aspesk yang berpengaruh pada proses pengambilan keputusan di berbagai aspek, termasuk dalam menganalisis data sampah plastik di setiap negara. Sampah plastik yang terus meningkat menjadi isu global yang mendesak, dengan berbagai dampak negative yang ditimbulkan. Menurut What a Waste 2.0 dari World Bank produksi limbah

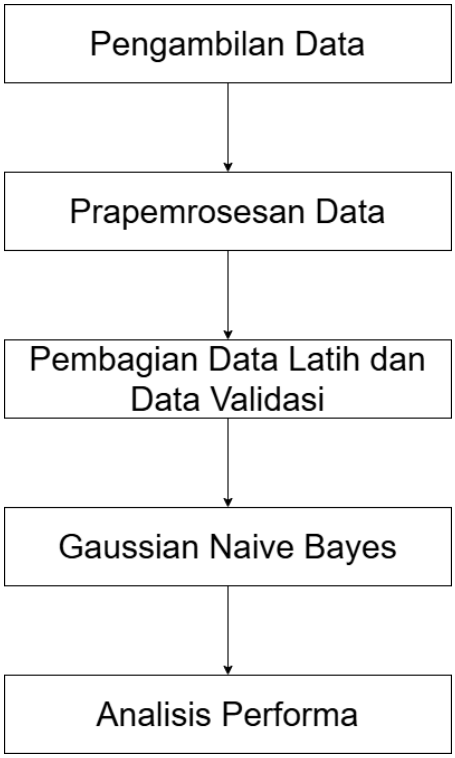
diprediksi meningkat menjadi 3,40 miliar ton pada tahun 2050, mengalami peningkatan dari 2,01 miliar ton pada tahun 2016. Di seluruh dunia, setidaknya 33% limbah ini dikelola secara tidak tepat melalui pembuangan terbuka atau pembakaran(Kaza dkk., 2018). Naive Bayes adalah salah satu dari metode Klasifikasi yang menggunakan perhitungan probabilitas dan statistic, menurut ilmuwan yang berasal dari Inggris yang bernama Thomas Bayes, naïve bayes adalah memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya(Arfanda dkk., 2021).

Algoritma Naïve Bayes memungkinkan analisis data statistik untuk memahami pola distribusi dan faktor-faktor yang memengaruhi produksi dan pengelolaan sampah plastik di setiap negara. Metode ini memungkinkan pengelompokan data berdasarkan probabilitas, yang menghasilkan hasil yang akurat meskipun data bersifat tidak seimbang. Penelitian ini bertujuan untuk mengisi celah tersebut dan menganalisis dan mengklasifikasikan pola produksi sampah plastik di seluruh dunia pada tahun 2023. Beberapa penelitian yang berkaitan dengan naïve bayes sudah banyak di lakukan diantaranya adalah penelitian dari Kamel dkk dengan judul "*Cancer Classification Using Gaussian Naive Bayes Algorithm*" yang menerapkan naïve bayes pada dataset *Wisconsin Breast Cancer Dataset (WBCD)* dan dataset kanker paru-paru, dengan hasil evaluasi menunjukkan tingkat akurasi persentasenya 98%, sedangkan pada pemprediksian kanker payudara terdapat hasil akurasi persentasenya 90% dalam memprediksi kanker paru-paru(Kamel dkk., 2019). Adapun penelitian lain yang berkaitan dengan Naïve Bayes dari Saputra J dengan judul "The Naive Bayes Algorithm in Predicting the Spread of the Omicron Variant of Covid-19 in Indonesia: Implementation and Analysis" hasil dari penelitian tersebut menunjukkan 16 data dari 33 data yang diuji untuk kasus Covid-19 pada daerah tingkat I terklasifikasi dengan benar, dengan akurasi 46,43%, sementara 16 data lainnya salah klasifikasi dengan akurasi yang sama(Saputra, 2022).

Pada penelitian dari Alvina dkk "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia" dengan tujuan penelitian memprediksi penyebaran pandemi COVID-19, khususnya di Indonesia. Hasil dari penelitian tersebut menunjukkan 16 data dari 33 data yang di uji dalam kasus tersebut, dengan keakuratan sebesar 48,49% (Felicia Watratan dkk., 2020). Dengan hasil penelitian ini, diharapkan dipelorehnya wawasan tentang penerapan Naive Bayes dalam sebuah dataset. Menjadi acuan tingkat kecocokan dataset Global Plastic Waste 2023: Country-wise dengan penerapan algoritma Naive Bayes sebagai bahan pertimbangannya adalah tingkat akurasi, presisi, recall, dan skor F1. Serta diharapkan hasil dari penelitian ini dapat memberikan gambaran bagaimana algoritma Naive Bayes dapat dioptimalkan untuk analisis dataset bertema lingkungan, khususnya dalam konteks Global Plastic Waste.

2. METODE

Metode yang terapkan pada penelitian berhubungan dengan penerapan algoritma naïve bayes. Algoritma naïve bayes mengklasifikasikan data *Global Plastic Waste 2023: Country-wise* agar dapat diketahui efektifitas naïve bayes dalam mengklasifikasikan data tersebut. Dalam metode ini terdapat langkah yang di lakukan dalam proses penelitian yang dipaparkan pada Gambar 1.



Gambar 1. Langkah penelitian

Pengambilan Data

Data yang diambil untuk keperluan penelitian ini menggunakan dataset yang sudah di sediakan didalam platform Kaggle. Data yang di ambil dari Kaggle adalah dataset *Global Plastic Waste 2023: Country-wise Data* yang berisi informasi terkait sampah plastik di berbagai negara pada tahun 2023. Data tersebut memfokuskan pada seberapa besar dampak sampah pelastik mempengaruhi lingkungan pesisir. Link pengambilan data dapat melalui url berikut *Global Plastic Waste 2023: Country-wise Data*.

Prapemrosesan Data

Prapemrosesan data yang dilakukan pada penelitian ini yaitu transformasi dataset kategorical menjadi numerik agar data dapat sesuai dengan format yang di perlukan untuk data maining. Data mining berfungsi sebagai menentukan pola atau informasi penting dalam data yang dipilih dengan menggunakan metode yang telah ditentukan (A'ayunnisa dkk., 2022). Selanjutnya, dilakukan pememilihan fitur dan label yang relevan agar analisis yang dilakukan menggunakan algoritma *naïve bayes* dapat berjalan lancar dan maksimal, terkait pemilihan fitur dan label telah diambil beberapa fitur dan tabel yang relevan.

Membagi Data

Membagi data atau sering di kenal dengan Cross-Validation adalah metode yang digunakan dalam pembelajaran mesin dan statistik untuk menilai performa model pada data yang belum pernah digunakan sebelumnya. Proses ini melibatkan pemisahan dataset menjadi 2 jenis data, data latih, dan data validasi. Data latih untuk melatih model dan Data validasi bertujuan untuk menguji kinerja model (Julkarnain, 2024).

Gausian Naïve Bayes

Naïve Bayes sendiri adalah metode pengelompokan berdasarkan pada Prinsip Bayes. Dalam proses klasifikasinya, Prinsip Bayes digunakan untuk mengidentifikasi keterkaitan antara probabilitas dua kejadian, yaitu A dan B, P(A) dan P(B), serta peluang terjadinya A yang dipengaruhi oleh B, dan peluang B yang dipengaruhi oleh A, yaitu P (A | B) dan P (B | A). (Afif, 2023). Untuk perumusan naïve bayes sendiri akan dipaparkan pada rumus persamaan 1.

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \quad [1]$$

Gaussian Naïve Bayes adalah sebuah algoritma pengelompokan numerik atau bisa disimpulkan klasifikasi yang bekerja dengan target data yang berjenis nilai atau nominal. Gaussian Naïve Bayes adalah klasifikasi sederhana yang didasarkan pada prinsip teorem bayes $(Naufal\ dkk., 2024)$. Gaussian Naïve Bayes untuk klasifikasi adalah metode paling penting dalam menghitung probabilitas dan statistik. Untuk perumusan dari Gaussian Naïve bayes akan dipaparkan pada persamaan 2.

$$P(X_i) = \frac{1}{\delta\sqrt{2\pi}} \exp\left(-\frac{(X_i - \mu)^2}{2\delta^2}\right) \quad [2]$$

Analisis Performa

Pada bagian Analisis performa diharapkan dapat menghasilkan evaluasi kinerja model yang dibuat. Evaluasi dapat dilakukan dengan mengetahui akurasi, presisi, recall, dan skor F1. Dimana Akurasi dapat di peroleh dari menghitung kedekatan nilai prediksi dengan fakta. Rumus akurasi disajikan pada Persamaan 3. Presisi merupakan rasio antara sampel relevan yang terpilih terhadap seluruh sampel yang dipilih. Presisi juga dikenal sebagai tingkat kesesuaian antara permintaan informasi dengan jawabannya. Rumus untuk presisi ditunjukkan pada persamaan. 4. Hasil analisis kinerja memberikan gambaran seberapa baik model melakukan klasifikasi pada dataset pengujian. Selain itu, matriks konfusi dapat digunakan untuk menampilkan distribusi prediksi model terhadap data aktual, sehingga memudahkan identifiksi kesalahan prediksi untuk setiap kategori risiko(Baharuddin dkk., 2019).

$$\text{AKURASI} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad [3]$$

$$\text{PRESISI} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad [4]$$

$$\text{RECALL} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad [5]$$

$$F\text{-Measure} = 2 \cdot \frac{\text{Presisi} \times \text{Recall}}{\text{Presisi} + \text{Recall}} \quad [6]$$

Dimana:

- **TP** : Jumlah prediksi positif yang *true*.
- **FP** : Jumlah prediksi positif yang *false*.
- **FN** : Jumlah data positif yang tidak terbaca oleh model.
- **TN** : Jumlah prediksi negatif yang *true*.

3. HASIL DAN PEMBAHASAN

Pengambilan Data

Pengambilan data akan menghasilkan dataset dari *Global Plastic Waste 2023: Country-wise Data.* Data ini memiliki beberapa atribut diantaranya. Data ini juga memiliki validasi data yaitu Coastal_Waste_Risk. Untuk bentuk dataset akan dipaparkan pada Tabel 1. Data akan dilakukan prapemrosesan untuk melihat apakah data sudah siap untuk diterapkan pada metode Naïve Bayes.

Tabel 1. Dataset *global plastic waste 2023: country-wise*

Country	Total_Plastic_Waste_MT	Main_Sources	Recycling_Rate	Per_Capita_Waste_KG	Coastal_Waste_Risk
China	59.08	Packaging_Industrial	29.8	41.2	High
United States	42.02	Packaging_Consumer	32.1	127.5	Medium
India	26.33	Consumer_Goods	11.5	19.3	High
Japan	7.99	Packaging_Electronics	84.8	63.2	Medium
Germany	6.28	Automotive_Packaging	56.1	75.6	Low
Brazil	5.96	Consumer_Packaging	1.2	28.1	Medium
Indonesia	5.85	Food_Packaging	11.8	21.3	Very_High
Russia	5.84	Industrial_Consumer	5.6	40.2	Medium
United Kingdom	5.03	Packaging_Consumer	46.2	74.3	Low

Analisis Performa

Hasil akhir dari penelitian ini adalah Hasil evaluasi model berupa akurasi, presisi, recall, dan skor F1. Evaluasi model di perlukan untuk memberikan gambaran performa model algoritma Naïve Bayes, terutama dalam konteks data yang mungkin memiliki distribusi kelas tidak seimbang. Hasilnya data *Global Plastic Waste 2023: Country-wise Data* kurang cocok dengan naïve bayes karena Tingkat akurasi pada yang hanya 48.48%. Tingkat akurasi data uji bisa rendah adalah karena adanya data yang tidak seimbang dalam jumlah sampel label. Analisis performa akan dipaparkan pada Tabel 2.

Tabel 2. Analisis Performa

Kategori Risiko	Precision	Recall	F1-Score	Support
High	0.61	0.92	0.73	12
Low	0.36	0.36	0.36	11
Medium	0.50	0.11	0.18	9
Very High	0.00	0.00	0.00	1

Hasil analisis performa memberikan pemaparan bahwa model algoritma Naïve Bayes mempunyai performa terbaik pada kategori risiko High dengan F1-Score 0.73, berarti kemampuan tinggi dalam mengidentifikasi kategori dengan Recall 0.92. Pada kategori Low, model menunjukkan performa sedang dengan F1-Score 0.36, berarti tingkat prediksi dan identifikasi yang seimbang. Namun, pada kategori Medium, model mengalami kesulitan dengan F1-Score 0.18 dan Recall hanya 0.11, sedangkan pada kategori Very High, model gagal sepenuhnya mengenali data Precision, Recall, F1-Score 0.00 akibat jumlah data yang sangat

sedikit (Support 1). Secara keseluruhan, model bekerja lebih baik pada kategori dengan jumlah data lebih banyak atau pada kategori High tetapi kesulitan pada kategori dengan data terbatas seperti Medium dan Very High.

4. PENUTUP

Penelitian ini berfokus pada menganalisa performa algoritma Naïve Bayes dalam mengklasifikasikan dataset *Global Plastic Waste 2023: Country-wise Data*. Meskipun algoritma Naïve Bayes menunjukkan performa terbaik pada kategori High, namun hasil evaluasi secara keseluruhan mengungkapkan bahwa model ini kurang optimal untuk dataset yang digunakan. Alasan utamanya karena ketidakseimbangan jumlah data antar kategori. Tingkat akurasi keseluruhan sebesar 48.48% menggambarkan bahwa algoritma ini memiliki batasan dalam menangani data yang tidak merata. Maka diperlukan untuk mencari metode lain yang lebih akurat dan relevan untuk pengambilan keputusandalam kasus klasifikasi dataset *Global Plastic Waste 2023: Country-wise*.

DAFTAR PUSTAKA

A'ayunnisa, N., Salim, Y., & Azis, H. (2022). Analisis performa metode Gaussian Naïve Bayes untuk klasifikasi citra tulisan tangan karakter arab. *Indonesian Journal of Data and Science (IJODAS)*, *3*(3), 115–121.

Afif, A. (2023). Analisis Metode Klasifikasi Data Naïve Bayes dan SVM Dalam Menentukan Keunikan Hotel. Dalam *Jurnal Teknologi Komputer dan Informasi (JUTEKINF)* (Vol. 11, Nomor 1).

Arfanda, I., Ramdhan, W., Yusda, R. A., & Artikel, H. (2021). *Naive Bayes Dalam Menentukan Penerima Bantuan Langsung Tunai*. *1*(1). https://doi.org/10.47709/briliance.vxix.xxxx.

Baharuddin, M. M., Azis, H., & Hasanuddin, T. (2019). analisis performa metode k-nearest neighbor untuk identifikasi jenis kaca. *ILKOM Jurnal Ilmiah*, *11*(3), 269–274. https://doi.org/10.33096/ilkom.v11i3.489.269-274.

Felicia Watratan, A., Puspita, A. B., Moeis, D., Informasi, S., & Profesional Makassar, S. (2020). Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia. Dalam *Journal Of Applied Computer Science And Technology (JACOST)* (Vol. 1, Nomor 1). http://journal.isas.or.id/index.php/JACOST.

Julkarnain, M. (2024). *Penerapan Algoritma Naive Bayes dalam Memprediksi Lulus Tepat Waktu Mahasiswa*. *4*(2). https://doi.org/10.47709/digitech.v4i2.4963

Kamel, H., Abdulah, D., & Al-Tuwaijari, J. M. (2019). Cancer Classification Using Gaussian Naive Bayes Algorithm. *2019 International Engineering Conference (IEC)*, 165–170. https://doi.org/10.1109/IEC47844.2019.8950650.

Kaza, S., Yao, L. C., Bhada-Tata, P., & Van Woerden, F. (2018). *What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050*. Washington, DC: World Bank. https://doi.org/10.1596/978-1-4648-1329-0.

Naufal, Y., Putro, R., Afriansyah, A., & Bagaskara, R. (2024). Penggunaan Algoritma Gaussian Naïve Bayes & Decision Tree Untuk Klasifikasi Tingkat Kemenangan Pada Game Mobile Legends. *JUKI : Jurnal Komputer dan Informatika*, *6*.

Saputra, J. (2022). The Naive Bayes Algorithm in Predicting the Spread of the Omicron Variant of Covid-19 in Indonesia: Implementation and Analysis. *IJIIS: International Journal of Informatics and Information Systems*, *5*(2), 84–91. https://doi.org/10.47738/ijiis.v5i2.131.