Statistika, Vol. 24 No. 2, 181 – 190
November 2024

Pendekatan *Naïve Bayes* untuk Mengidentifikasi Perokok Tembakau Berdasarkan Faktor Sosio-Demografi dan Kesehatan

FAZA IZZATUL MUTTAQIN[1], ACHMAD FAUZAN[2]

[1, 2)] Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia, Jl. Kaliurang KM. 14,5, Sleman, 55584, Indonesia

e-mail: achmadfauzan@uii.ac.id

ABSTRAK

Merokok adalah kebiasaan yang berdampak negatif pada kesehatan masyarakat Indonesia. Untuk mengatasi masalah ini, diperlukan pemahaman tentang faktor-faktor demografi, pendidikan, dan kesehatan yang mempengaruhi perilaku merokok. Penggunaan metode klasifikasi *Naïve Bayes* relevan karena dapat memberikan wawasan menyeluruh mengenai faktor-faktor yang mempengaruhi perilaku perokok tembakau di Kota Tasikmalaya. Penelitian ini menggunakan data dari Survei Sosial Ekonomi Nasional (SUSENAS) Maret 2023 di Kota Tasikmalaya. Data dibagi menjadi data pelatihan dan data pengujian, dan kemudian dibagi lagi untuk validasi model. Model tersebut dibuat dan dievaluasi dengan mengacak sampel, membuat model, dan mengevaluasi sebanyak 100 kali. Hasil evaluasi kinerja model dalam mengklasifikasi data validasi menunjukkan akurasi rata-rata 87% dengan standar deviasi 0,0689 dan standar error 0,00689. *Confidence interval* antara 85.64% dan 88.35% menunjukkan model yang dapat diandalkan. Pengujian akhir model pada *data testing* mencapai akurasi 83.33%, menunjukkan model berkemampuan baik dalam mengklasifikasikan perokok tembakau di Kota Tasikmalaya pada Februari 2023.

Kata Kunci: Merokok, Klasifikasi, Naïve Bayes, Model, Akurasi.

ABSTRACT

Smoking is a habit that negatively impacts the health of the Indonesian population. To address this issue, it is essential to understand the demographic, educational, and health factors influencing smoking behavior. The use of the Naïve Bayes classification method is relevant as it provides comprehensive insights into the factors affecting tobacco smoking behavior in the city of Tasikmalaya. This study uses data from the National Socio-Economic Survey (SUSENAS) conducted in March 2023 in Tasikmalaya. The data is divided into training and testing datasets, with the training data further split for model validation. The model was built and evaluated by randomizing samples, creating models, and evaluating them 100 times. The model's performance evaluation in classifying the validation data showed an average accuracy of 87% with a standard deviation of 0.0689 and a standard error of 0.00689. The confidence interval between 85.64% and 88.35% indicates a reliable model. The final testing of the model on the testing data achieved an accuracy of 83.33%, model demonstrating a good capability in classifying tobacco smokers in Tasikmalaya in February 2023.

Keywords: Smoking, Classification, Naïve Bayes, Model, Accuration

1. PENDAHULUAN

Kesehatan merupakan elemen penting yang dipengaruhi oleh Berbagai faktor internal dan eksternal memengaruhi kesehatan seseorang. Faktor *internal* termasuk demografi seperti usia, jenis kelamin, tingkat pendidikan, dan pekerjaan, sementara faktor eksternal mencakup aspek lingkungan, sosial, budaya, serta riwayat kesehatan, kondisi fisik, dan status gizi (Ediana & Sari, 2022). Salah satu permasalahan kesehatan yang krusial di Indonesia adalah merokok, yang

menjadi penyebab utama beberapa penyakit kronis yang berpotensi fatal dan berujung pada kematian (Husein & Menga, 2019). Merokok kerap dikaitkan dengan dampak negatif terhadap kesehatan masyarakat Indonesia. Data dari *World Health Organization* (WHO) menunjukkan bahwa prevalensi merokok terus meningkat di Indonesia setiap tahunnya (World Health Organization, 2020). Hasil survei *Global Adult Tobacco Survey* (GATS) yang dilakukan oleh Kementerian Kesehatan (Kemkes) menunjukkan bahwa jumlah perokok dewasa telah meningkat secara signifikan, dari 60,3 juta pada tahun 2011 menjadi 69,1 juta pada tahun 2021 (Kementerian Kesehatan, 2022). Di Kota Tasikmalaya, data yang berasal dari BPS dan Dinas Kesehatan setempat mengindikasikan bahwa prevalensi perokok mencapai 28,5% dari total penduduk usia dewasa pada tahun 2022.

Seiring dengan meningkatnya jumlah perokok, penting untuk memahami faktor-faktor yang memengaruhi perilaku merokok dalam masyarakat. Selain menjadi masalah kesehatan, merokok juga dianggap sebagai masalah sosial karena memiliki konsekuensi signifikan terhadap lingkungan keluarga dan kesehatan masyarakat secara keseluruhan. Kebiasaan ini lebih sering ditemukan pada masyarakat berpenghasilan rendah, yang berimplikasi pada ketidakstabilan ekonomi keluarga, akses pendidikan yang terhambat, pertumbuhan ekonomi yang lambat, serta ancaman terhadap nilai-nilai agama. Untuk mengurangi dampak negatifnya, pemerintah Indonesia telah melakukan berbagai upaya untuk mengontrol masalah merokok melalui berbagai kebijakan, termasuk Peraturan Pemerintah (PP) dan Peraturan Daerah. Salah satu tindakan konkret yang diambil adalah penerapan Peraturan Pemerintah Nomor 19 Tahun 2003 tentang Pengamanan Rokok bagi Kesehatan. PP ini bertujuan untuk mencegah penyakit yang disebabkan oleh rokok baik pada individu maupun masyarakat umum (Jannah & Purwanta, 2018). Di sisi lain, merokok dianggap sebagai masalah sosial karena memiliki konsekuensi yang signifikan terhadap lingkungan keluarga dan kesehatan masyarakat secara keseluruhan. Kebiasaan ini cenderung tersebar di kalangan masyarakat yang berpenghasilan rendah atau kurang mampu secara ekonomi. Dampaknya sangat merugikan dalam hal meningkatkan risiko ketidakstabilan ekonomi keluarga, merintangi akses terhadap pendidikan, menghambat pertumbuhan ekonomi, serta mengancam nilai-nilai agama. Selain itu, praktik merokok juga dibatasi oleh hukum untuk meminimalkan efek negatifnya (Gule, 2022).

Namun, untuk mengatasi masalah merokok secara lebih efektif, diperlukan pemahaman mendalam mengenai perilaku merokok berdasarkan faktor-faktor demografi, pendidikan, dan kesehatan. Oleh karena itu, penelitian tentang penerapan Metode Naïve Bayes untuk klasifikasi perilaku merokok di Kota Tasikmalaya menjadi sangat relevan. Salah satu penelitian yang mengkaji tentang klasifikasi perokok menggunakan metode Naïve Bayes. Penelitian oleh Arindi & Lumbanbatu (2022) menunjukkan penerapan Naïve Bayes dalam klasifikasi kecanduan rokok berdasarkan atribut-atribut yang relevan seperti umur, tingkat kecanduan, waktu yang dibutuhkan merokok, alasan merokok, dan tempat merokok. Hasil penelitian tersebut yaitu mendapatkan model Naïve Bayes untuk mengklasifikasi tingkat kecanduan perokok. Masih banyak lagi penelitian yang dilakukan dengan menggunakan metode Naïve Bayes. Namun, masih banyak ruang untuk mengeksplorasi penggunaan Naïve Bayes untuk mengindentifikasi perokok berdasarkan faktor sosio-demografi dan kesehatan.

Penelitian ini berbeda dari penelitian-penelitian sebelumnya karena lebih berfokus pada penggunaan metode Naïve Bayes untuk klasifikasi perokok dengan mempertimbangkan atribut sosio-demografi dan kesehatan. Selain itu, penelitian ini juga akan mengevaluasi efektivitas Naïve Bayes dalam mengklasifikasi data perokok di berbagai kelompok umur dan latar belakang sosial, yang belum banyak dieksplorasi dalam penelitian-penelitian sebelumnya. Dalam hal ini, keunggulan utama Naïve Bayes adalah kemampuan untuk memberikan perkiraan kemungkinan dari kelas target, baik perokok maupun bukan perokok, berdasarkan kombinasi atribut input. Dengan menggunakan model ini, strategi intervensi yang lebih tepat sasaran dapat dibuat dengan membagi individu ke dalam kategori yang lebih spesifik berdasarkan perilaku merokok mereka. Penelitian ini diharapkan dapat memberikan wawasan komprehensif mengenai dinamika dan faktor-faktor yang memengaruhi kebiasaan merokok di wilayah tersebut, serta berkontribusi dalam upaya pengendalian perilaku merokok melalui pendekatan yang lebih terfokus dan berbasis data.

2. METODE PENELITIAN

Perilaku merokok berarti membakar, menghisap, dan/atau menghirup produk tembakau yang mengandung nikotin dan tar, dengan atau tanpa bahan tambahan (Alamsyah & Nopianto, 2022).

Rokok mengandung substansi adiktif yang memiliki potensi membahayakan kesehatan individu maupun masyarakat jika dikonsumsi (Patandung & Feriyanto, 2022). Dalam satu batang rokok terdapat sekitar 4000 bahan kimia yang dapat mengancam kesehatan. Di antara berbagai bahan tersebut, tiga di antaranya yang paling berpotensi membahayakan adalah karbon monoksida, nikotin, dan tar (Oroh et al., 2022).

Pre-processing adalah langkah penting dalam proses ekstraksi data, dan terkadang data yang digunakan pada tahap ini tidak cocok untuk diproses secara langsung (Purbolaksono et al., 2021). Menurut Gunawan, proses mempersiapkan data, seperti membersihkan suara atau mengubah formatnya, disebut data *pre-processing (Gunawan, 2016).* Oleh karena itu, *pre-processing* merupakan langkah penting dalam mengatasi masalah yang dapat mempengaruhi hasil klasifikasi data. Proses *pre-processing* dalam penelitian ini mencakup beberapa tahapan, seperti filtering, penyesuaian kelas, dan pemilihan sampel dari setiap kelas.

Klasifikasi merupakan langkah dimana objek-objek dikelompokkan berdasarkan kemiripan karakteristik atau ciri-ciri tertentu ke dalam kelas-kelas yang berbeda (Indriani et al., 2017). Ini melibatkan mengalokasikan data atau objek baru ke kelas atau label yang sesuai, didasarkan pada atribut khusus. Teknik klasifikasi melibatkan pengamatan variabel dalam dataset yang ada untuk memprediksi kelas dari objek yang belum dikenal sebelumnya. Proses klasifikasi meliputi tiga tahap utama: pembuatan model, penggunaan model, dan evaluasi (Nasution et al., 2019). Pembuatan model melibatkan konstruksi model menggunakan data latih yang memiliki atribut dan label, yang kemudian diterapkan pada data baru untuk mengklasifikasikannya. Evaluasi dilakukan untuk mengevaluasi sejauh mana model mampu memprediksi dengan benar menggunakan data yang belum dilihat sebelumnya. Proses klasifikasi juga dapat dibagi menjadi tahap pelatihan dan tahap pengujian. Pembagian dataset adalah cara yang dianggap sangat penting dan diperlukan untuk menghilangkan atau mengurangi bias pada data *training* dalam model *machine learning* (Muraina, 2022). Tahap pelatihan melibatkan penggunaan data untuk membentuk model, sementara tahap pengujian melibatkan pengujian model dengan data terpisah untuk menilai performanya.

Algoritma Naïve Bayes mengaplikasikan metode probabilitas dan statistik yang dikembangkan oleh ilmuwan Inggris, Thomas Bayes. Dalam proses klasifikasi teks, algoritma ini melalui dua tahap utama: pelatihan dan pengujian(Wibisono et al., 2020). Sampel dokumen dianalisis selama tahap pelatihan untuk menentukan kata-kata yang kemungkinan besar muncul dalam kumpulan dokumen tersebut, yang kemudian digunakan sebagai representasi dokumen. Selain itu, pada tahap klasifikasi, probabilitas awal untuk setiap kategori ditentukan berdasarkan sampel dokumen. Pola kemunculan kata-kata dalam dokumen menentukan kategori dokumen tersebut. Asumsi yang kuat bahwa itu independen dari setiap situasi atau kejadian adalah ciri utama dari Naïve Bayes *Classifier* (Watratan et al., 2020). Persamaan metode naïve bayes (Alfa, 2015).

$$P(H|X) = \frac{P(X|H)\,.\,P(H)}{P(X)} \quad \ldots (1)$$

Teorema ini digunakan untuk menghitung probabilitas dari suatu hipotesis H yang diberikan data X, atau lebih dikenal sebagai probabilitas posterior P(H|X). Teorema Bayes menggabungkan informasi dari probabilitas prior P(H) dan *likelihood* P(X|H) untuk memperbarui keyakinan kita terhadap hipotesis H setelah mengamati data X. Dengan kata lain, teorema ini memungkinkan kita untuk memperbaiki atau memperbarui estimasi probabilitas suatu hipotesis berdasarkan data baru yang diperoleh.

Terdapat beberapa cara dalam mengevaluasi kinerja model Naïve Bayes dalam melakukan klasifikasi. Evaluasi kinerja model penelitian ini, yaitu dengan menggunakan nilai *accuracy*, standar deviasi, standar *error*, dan interval kepercayaan.

Akurasi adalah metrik evaluasi yang menilai seberapa baik model memiliki kemampuan untuk membuat prediksi yang akurat berdasarkan data yang belum pernah dilihat sebelumnya (Sanhaji et al., 2024). Akurasi mengukur persentase prediksi yang tepat dibuat oleh model dari total prediksi yang dibuat. Secara matematis, akurasi dihitung dengan menggunakan rumus sebagai berikut (Ayudhitama & Pujianto, 2020).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad \ldots (2)$$

Jumlah prediksi benar *True Positive* (TP) dan *True Negative* (TN) dibagi dengan total prediksi, yang terdiri dari *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN), untuk menentukan akurasi. Hasil pembagian ini kemudian dikalikan dengan 100% untuk mendapatkan nilai akurasi dalam bentuk persentase.

Standar deviasi, juga dikenal sebagai simpangan baku, adalah ukuran yang menunjukkan tingkat variasi dalam suatu kelompok atau standar penyimpangan dari rata-rata (Febriani, 2020). Ini memberikan gambaran tentang seberapa variabel data tersebut. Standar deviasi menunjukkan seberapa dekat atau jauh setiap titik data dari rata-rata. Standar deviasi yang lebih tinggi menunjukkan lebih banyak variabilitas dalam data, dan standar deviasi yang lebih rendah menunjukkan bahwa data lebih homogen atau konsisten. Rumus umum untuk menghitung standar deviasi adalah sebagai berikut (Pratama et al., 2023).

$$s = \sqrt{\frac{\sum_{n-1}^{n}(x_i - \bar{x})^2}{n-1}} \quad \ldots (3)$$

Rumus ini digunakan untuk menghitung standar deviasi sampel, yang memberikan perkiraan tentang variasi atau dispersi dari nilai-nilai dalam sampel data. Standar deviasi dihitung dengan terlebih dahulu menghitung deviasi (selisih) setiap nilai data x_i dari rata-rata sampel $\bar{x}$, kemudian mengkuadratkan deviasi tersebut, menjumlahkan hasilnya, membaginya dengan $n-1$ (di mana $n-1$ adalah derajat kebebasan), dan akhirnya mengambil akar kuadrat dari hasil tersebut.

Standar deviasi dari distribusi sampel dalam statistik disebut standar *error* (SE). Standar *error* merujuk pada perkiraan standar deviasi dari sampel tertentu yang digunakan untuk menghitung nilai estimator (Arieska & Pusponegoro, 2016). Ini memberikan perkiraan seberapa jauh rata-rata sampel dapat bervariasi dari rata-rata populasi. Rumus umum menghitung standar *error* adalah sebagai berikut (Turot et al., 2016).

$$SE = \frac{s}{\sqrt{n}} \quad \ldots (4)$$

Standar deviasi (s) menunjukkan seberapa jauh data dalam sampel tersebar dari nilai rata-rata sampel. Jumlah total data dalam sampel (n) digunakan untuk mengukur ukuran sampel yang diambil dari populasi. Semakin besar ukuran sampel, semakin kecil standar *error*, yang menunjukkan bahwa sampel lebih mungkin mewakili populasi secara akurat.

Interval kepercayaan, yang didasarkan pada observasi sampel dan memiliki probabilitas tertentu yang telah ditentukan, disebut interval kepercayaan (Kamulyan et al., 2017). Interval kepercayaan memberikan perkiraan berdasarkan data sampel tentang di mana parameter populasi sebenarnya mungkin berada. Rumus umum menghitung *Confidence Interval* adalah sebagai berikut (Prastiwi, 2018).

$$CI = \bar{x} \pm z \times \left(\frac{s}{\sqrt{n}}\right) \quad \ldots (5)$$

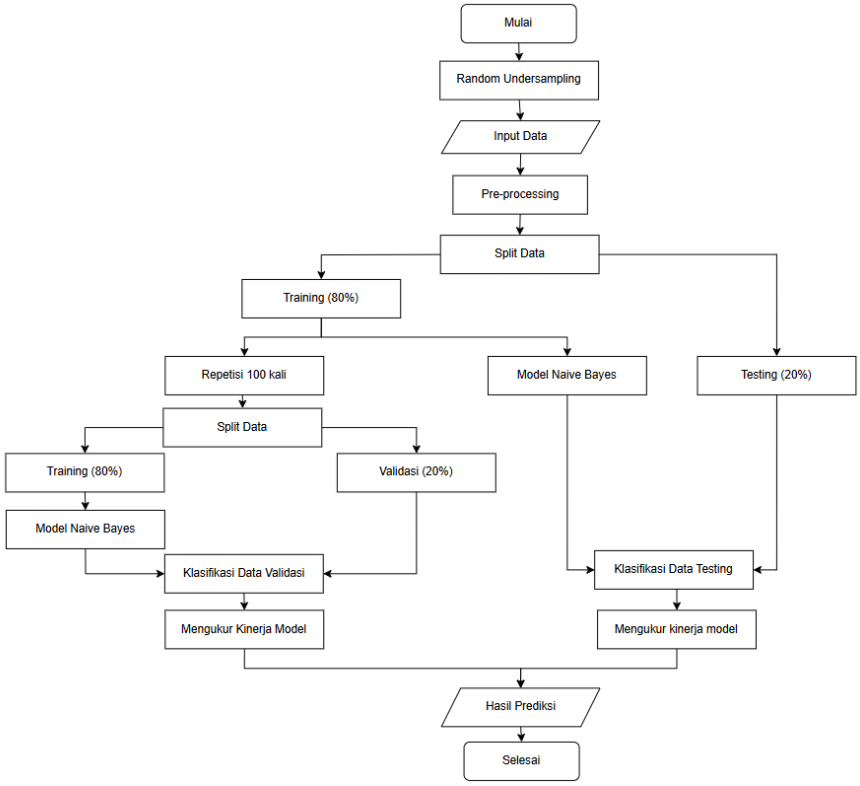
Dalam rumus ini, $\bar{x}$ merupakan rata-rata dari sampel yang telah diambil. Nilai z adalah skor z yang sesuai dengan tingkat kepercayaan yang diinginkan, misalnya untuk tingkat kepercayaan 95%, nilai z biasanya adalah 1.96. Parameter s merupakan standar deviasi dari sampel, yang menggambarkan seberapa tersebar data sampel tersebut. Sementara itu, n adalah jumlah total data dalam sampel yang diambil. Interval kepercayaan ini memberikan rentang nilai yang diyakini, dengan tingkat keyakinan tertentu, mencakup rata-rata sebenarnya dari populasi yang lebih besar. Dalam interpretasinya, semakin kecil standar deviasi atau semakin besar ukuran sampel, semakin sempit interval kepercayaan, yang menunjukkan estimasi rata-rata populasi yang lebih akurat.

Data yang digunakan pada penelitian ini merupakan data sukender yang diperoleh dari Survei Sosial Ekonomi Nasional (SUSENAS) 2023 bulan Maret di Kota Tasikmalaya yang diperoleh dengan membelinya di website silastik.bps.go.id. Terdapat 6 variabel pada penelitian ini yang terdiri dari variabel dependen dan independen. Variabel dependen adalah variabel yang dipengaruhi, yaitu Apalah merokok tembakau. Variabel indepeden adalah varaibel yang mempengaruhi, yaitu Jenis Kelamin, Umur, Apakah bersekolah, Apakah mempunyai keluhan kesehatan, dan Apakah dulu pernah merokok tembakau.

Tabel 1. Keterangan Variabel dan Kelasnya

No	Variabel	Simbol	Keterangan
1	Apakah merokok tembakau	Y_1	Apakah selama sebulan terakhir merokok tembakau? o Ya, setiap hari o Ya, tidak setiap hari o Tidak
2	Umur	X_1	Umur (nama) o GEN Z (11-27 tahun) o GEN Y (28-43 tahun) o GEN X (44-59 tahun) o Baby_Boomer (60-78 tahun)
3	Jenis Kelamin	X_2	Jenis kelamin seseorang o Laki-laki o Perempuan
4	Apakah bersekolah	X_3	Apakah (nama) bersekolah? o Tidak/belum pernah bersekolah o Masih bersekolah o Tidak bersekolah lagi
5	Apakah mempunyai keluhan kesehatan	X_4	Dalam sebulan terakhir, apakah (nama) mempunyai keluhan kesehatan (panas, batuk, pilek, diare, pusing, penyakit kronis, dsb.)? o Ya o Tidak
6	Apakah dulu pernah merokok tembakau	X_5	Apakah dulu, sebelum sebulan terakhir merokok tembakau? o Ya, setiap hari o Ya, tidak setiap hari o Tidak

Alur penelitian ini dijelaskan melalui diagram alir berikut:



Gambar 1. Diagram Alir Penelitian

Berdasarkan diagram alir penelitian diatas, uraian tahapan yang dilakukan adalah sebagai berikut.

1. Mulai.
2. Data yang digunakan pada penelitian ini merupakan dari Survei Sosial Ekonomi Nasional (SUSENAS) 2023 bulan Maret di Kota Tasikmalaya yang diperoleh dengan membelinya di website silastik.bps.go.id. Pada laporan ini, penulis melakukan klasifikasi dengan menggunakan metode *Naïve Bayes* apakah seseorang merokok tembakau berdasarkan faktor pendidikan, kesehatan, dan demografi. Sebelum melakukan klasifikasi, penulis melakukan metode random undersampling untuk mengurangi jumlah data dari kelas mayoritas (kelas: ya, setiap hari dan tidak) secara acak sampai mencapai jumlah yang seimbang dengan kelas minoritas (ya, tidak setiap hari).
3. Melakukan input data untuk di analisis menggunakan *software* Rstudio.
4. *Pre-processing data* seperti perubahan tipe data dan pengecekkan *missing value.*
5. Melakukan *split data* dengan menggunakan *hold out method* yaitu dengan membagi data menjadi *data training* dan *data testing* dengan proporsi (80:20).
6. Melakukan pengacakkan sampel, pemodelan, dan *confusion matrix* sebanyak 100 kali.
7. *Data training* yang sudah dibagi kemudian dibagi kembali menjadi *data training* dan data validasi dengan proporsi (80:20).
8. Pembuatan model dengan menggunakan *data training*. Setelah model dibuat, dilakukan prediksi dan klasifikasi data validasi dengan menggunakan model yang sudah dibuat.
9. Mengukur kinerja model dalam mengklasifikasikan data validasi dengan mengetahui rata-rata akurasi, standar deviasi, standar *error*, dan *confidence interval.*
10. Melakukan evaluasi akhir model dengan melakukan klasifikasi *data testing* menggunakan model yang dilatih menggunakan *data training.*
11. Mengukur kinerja model dengan mengetahui akurasi kinerja model dalam mengklasifikasikan *data testing.*
12. Dari pengujian data testing dan validasi didapatkan kemampuan model dalam mengklasifikasikan data.
13. Selesai.

3. HASIL DAN PEMBAHASAN

3.1 *Pre-Processing Data*

Data dalam penelitian ini merupakan data Survei Sosial Ekonomi Nasional (SUSENAS) 2023 bulan Maret di Kota Tasikmalaya yang diperoleh dengan membelinya di *website* silastik.bps.go.id. Penulis menentukan variabel yang ingin digunakan dan membelinya di *website* silastik.bps.go.id. Data yang diperoleh dari silastik diubah menjadi pertanyaan sesuai pada kuisioner sensus nya, sebelum melakukan analisis peneliti melakukan *random undersampling* untuk mengurangi jumlah data dari kelas mayoritas (kelas: ya, setiap hari dan tidak) secara acak sampai mencapai jumlah yang seimbang dengan kelas minoritas (ya, tidak setiap hari) agar menghindari *overfitting*, kemudian untuk variabel umur dilakukan penyesuaian kelas menurut generasi untuk memudahkan proses pengklasifikasian, dan dilakukan *pre-processing* seperti mengubah tipe data dan pengecekkan *missing values* sehingga data dapat di analisis.

3.2 *Split Data*

Data yang digunakan dalam penelitian ini dibagi menjadi dua yaitu *data training* dan *data testing* dengan proporsi (95:5). Data yang digunakan berjumlah 102 data, setelah dilakukan pembagian diketahui *data training* sebanyak 96 data dan *data testing* sebanyak 6 data. Alasan dari pembagian ini yaitu keterbatasan data yang dimiliki untuk menguji kinerja model yang sebenarnya. Oleh karena itu, pembagian ini dilakukan agar model dilatih dengan maksimal dan mendapatkan kinerja model yang maksimal.

3.3 Klasifikasi *Naïve Bayes* dengan Menggunakan Data Validasi

Data training dari keseluruhan data dipecah kembali menjadi dua yang terdiri dari *data training* dan data validasi dengan proporsi (80:20) untuk mengevaluasi model selama proses pelatihan untuk memilih model terbaik. Ini membantu dalam mendeteksi *overfitting* karena data validasi tidak digunakan untuk melatih model. Dari *data training* yang berjumlah 96 data dipecah kembali menjadi 2, sehingga diketahui *data training* sebanyak 76 data dan data validasi sebanyak 20 data.

Pengacakkan sample yang digunakan dalam bentuk *data training* dan validasi dapat mempengaruhi hasil akurasi dari model pengklasifikasian. Oleh karena itu, untuk mengetahui hasil akurasi yang sebenarnya dari model pengklasifikasian dilakukan lah pengacakkan sampel, pemodelan, *confusion matrix* sebanyak 100 kali untuk mengetahui rata-rata dari akurasi kinerja model dalam melakukan klasifikasi apakah seseorang merokok tembakau dengan menggunakan metode *Naïve Bayes*. Menentukan akurasi sebuah model dapat menggunakan persamaan (2). Berikut merupakan tampilan *confusion matrix* dari simulasi pertama:

Tabel 2. C*onfusion Matrix* Simulasi 1

Nilai Prediksi	Nilai Sebenarnya		
	Tidak	Ya, setiap hari	Ya, tidak setiap hari
Tidak	8	0	0
Ya, setiap hari	3	4	1
Ya, tidak setiap hari	0	1	3

Dari tabel diatas, didapatkan hasil klasifikasi sebagai berikut:

1. Kelas tidak terklasifikasi tidak sebanyak 6 dan terklasifikasi ya, setiap hari sebanyak 1
2. Kelas ya, setiap hari terklasifikasi tidak sebanyak 1, terklasifikasi ya, setiap hari sebanyak 1, dan terklasifikasi ya, tidak setiap hari sebanyak 2
3. Semua kelas ya, tidak setiap hari terklasifikasi dengan benar sebanyak 6

Berdasarkan tabel diatas, didapatkan juga hasil perhitungan nilai akurasinya:

$$Accuracy = \frac{8+4+3}{8+3+4+1+1+3} \times 100\% = 75\%$$

Accuracy kinerja model data training simulasi 1 dalam mengklasifikasikan data validasi yaitu sebesar 75% yang berarti model mengklasifikasikan data dengan cukup baik. Untuk mendapatkan akurasi yang mendekati "sebenarnya" dari model *data training*, maka dilakukan lah pengacakkan sampel, pemodelan, dan *confusion matrix* sebanyak 100 kali untuk mengetahui rata-rata dari akurasi kinerja model dalam melakukan klasifikasi apakah seseorang merokok tembakau. Berikut merupakan tabel akurasi 100 simulasi:

Tabel 3. Akurasi Model *Training*

Simulasi	**Akurasi**
1	0.75
2	0.90
⋮	⋮
99	0.70
100	0.85
Rata-rata	**0.87**

Dari Tabel 3, diperoleh rerata akurasi kinerja model yang dilakukan sebanyak 100 kali yang digunakan dalam mengklasifikasikan apakah seseorang merokok tembakau adalah sebesar 87%. Dalam mengevaluasi kinerja model selain mengetahui rata-rata akurasi model, dapat dilakukan juga melalui nilai standar deviasi, standar *error*, dan *confidence interval*. Berikut merupakan hasil evaluasi kinerja model:

Tabel 4. Hasil Evaluasi Kinerja Model

Metode Evaluasi Kinerja Model	Hasil
Rata-rata Akurasi	0.87
Standar Deviasi	0.06890192
Standar *Error*	0.006890192
Confidence Interval	85.64% dan 88.35%.

Pada tabel diatas, didapatkan hasil evaluasi kinerja model. Pada tabel diatas didapatkan hasil sebagai berikut:

1. Standar deviasi digunakan untuk mengukur seberapa tersebar data dari nilai rata-ratanya. Standar deviasi dihitung dengan menggunakan rumus (3). Berikut perhitungannya:

$$s = \sqrt{\frac{(0.75-0.87)^2 + (0.90-0.87)^2 + (0.85-0.87)^2 + \cdots + (0.85-0.87)^2}{100-1}} = 0.06890192$$

   Standar deviasi dari semua simulasi pelatihan model menunjukkan variasi atau penyebaran dari nilai akurasi di antara iterasi pelatihan model. Semakin rendah standar deviasi, semakin konsisten hasilnya dari iterasi ke iterasi. Pada pelatihan model ini didapatkan standar deviasi sekitar 0.06890192 yang berarti variasi dari nilai akurasi konsisten dari iterasi ke iterasi.

2. Standar *error* mengukur ketidakpastian dari rata-rata sampel sebagai estimasi rata-rata populasi. Standar *error* dihitung dengan menggunakan rumus (4). Berikut perhitungannya:

$$SE = \frac{s}{\sqrt{n}} = \frac{0.06890192}{\sqrt{100}} = 0.006890192$$

   Standar *error* dalam evaluasi kinerja model digunakan guna mengukur seberapa akurat rata-rata sampel (rata-rata akurasi) merepresentasikan populasi (akurasi sebenarnya). Semakin rendah standar *error*, semakin akurat rata-rata sampel dalam merepresentasikan populasi. Pada pelatihan model ini didapatkan standar *error* sekitar 0.006890192 yang berarti akurasi rata-rata sampel cukup akurat dalam merepresentasikan populasi.

3. *Confidence Interval* memberikan rentang nilai di mana rata-rata akurasi populasi yang sebenarnya berada pada tingkat keyakinan tertentu. *Confidence Interval* dihitung dengan menggunakan rumus (5). Berikut perhitungannya:

$$CI = 0.87 \pm 1.96 \times (0.006890192) = 0.8564952 \text{ dan } 0.8835048$$

   Dengan tingkat kepercayaan 95%, penulis yakin bahwa rata-rata akurasi populasi sebenarnya berada di antara 85.64% dan 88.35%. Artinya, penulis yakin bahwa rata-rata akurasi hasil sampel (87%) cukup mewakili akurasi sebenarnya dari model Naïve Bayes.

3.4 Klasifikasi Naïve Bayes dengan Menggunakan *Data Testing*

Data testing digunakan untuk mengevaluasi model akhir, dilakukan klasifikasi dengan menggunakan model *naïve bayes* dari *data training* keseluruhan data. untuk memastikan generalisasi yang baik dan mengukur kinerja model yang sebenarnya pada data yang belum pernah dilihat *data training* dan data validasi. Setelah dilakukan prediksi dan klasifikasi diperlukan evaluasi kinerja model dengan menggunakan akurasi untuk mengetahui berapa tingkat akurasi akhir dari model dalam mengklasifikasi. Berikut merupakan *confusion matrix* yang berupa hasil kinerja model data training dalam mengklasifikasikan *data testing*:

Tabel 5. *Confusion Matrix Data Testing*

Nilai Prediksi	Nilai Sebenarnya		
	Tidak	Ya, setiap hari	Ya, tidak setiap hari
Tidak	2	0	0
Ya, setiap hari	0	2	1
Ya, tidak setiap hari	0	0	1

Dari tabel diatas, didapatkan hasil klasifikasi sebagai berikut:

1. Semua kelas “tidak” terklasifikasi dengan benar sebanyak 2

2. Semua kelas "ya, setiap hari" terklasifikasi dengan benar sebanyak 2
3. Kelas "ya, tidak setiap hari" terklasifikasi dengan benar sebanyak 1, terklasifikasi sebagai "ya, setiap hari" sebanyak 1

Berdasarkan tabel diatas, didapatkan juga hasil perhitungan nilai akurasinya:

$$Accuracy = \frac{2+2+1}{2+2+1+1} \times 100\% = 83.33\%$$

Berdasarkan tabel *confusion matrix*, kinerja metode *Naïve Bayes* dalam mengklasifikasikan *data testing* menunjukkan akurasi sebesar 83.33%. Hal ini mengindikasikan bahwa model ini mampu mengklasifikasikan data dengan baik. Dengan akurasi ini, model *Naïve Bayes* dapat digunakan secara efektif untuk mengidentifikasi perokok tembakau di Kota Tasikmalaya. Selain itu, model ini dapat membantu dalam menentukan faktor-faktor yang memiliki risiko tinggi untuk merokok, yang bisa menjadi dasar untuk intervensi dan kebijakan kesehatan masyarakat yang lebih tepat sasaran.

4. SIMPULAN DAN SARAN

Berdasarkan hasil dan pembahasan yang sudah dipaparkan, kemampuan model *Naïve Bayes* dalam mengklasifikasikan apakah seseorang merokok tembakau berdasarkan faktor kesehatan, pendidikan, dan demografi sudah sangat baik. Kinerja model *Naïve Bayes* yang dilatih dengan menggunakan *data training* dalam mengklasifikasikan data validasi sudah sangat baik dengan didapatkan hasil rata-rata akurasi sebesar 87% yang berarti akurasi model dalam mengklasifikasikan sudah baik, kemudian standar deviasi sebesar 0.06890192 yang berarti variasi dari nilai akurasi konsisten dari iterasi ke iterasi, standar *error* sebesar 0.006890192 yang berarti akurasi rata-rata sampel cukup akurat dalam merepresentasikan populasi, dan interval kepercayaan berada diantara 85.64 % dan 88.35% yang berarti akurasi 87% cukup mewakili akurasi yang sebenarnya. Untuk mengevaluasi akhir model dilakukan klasifikasi dengan menggunakan model *data training* terhadap data testing dan didapatkan hasil akurasi sebesar 83.33% yang berarti model mengklasifikasikan data testing dengan baik. Sehingga dapat disimpulkan bahwa model *naïve bayes* sudah baik dalam mengklasifikasikan apakah seseorang merokok tembakau pada bulan Februari 2023 di Kota Tasikmalaya.

Saran untuk penelitian selanjutnya yaitu dapat melakukan perbandingan metode klasifikasi lainnya agar dapat mengetahui perbandingan tingkat akurasi yang baik dalam mengklasifikasikan apakah seseorang merokok tembakau.

DAFTAR PUSTAKA

Alamsyah, A., & Nopianto. (2022). Determinan Perilaku Merokok pada Remaja. *Jurnal Endurance*, *2*(1), 25–30. https://doi.org/10.22216/jen.v2i1.1014

Alfa, S. (2015). Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. *Citec Journal*, *2*(3).

Arieska, D. I., & Pusponegoro, N. H. (2016). Pendugaan Standard Errordan Confidence Intervalkoefisien Gini dengan Metode Bootstrap: Terapan pada Data Susenas Provinsi Papua Barat Tahun 2013. *Jurnal Aplikasi Statistika & Komputasi Statistik*, *8*(2), 57–66.

Arindi, P., & Lumbanbatu, K. (2022). Klasifikasi Kecanduan Rokok Dengan Naive Bayes. *Pelita Informatika: Informasi Dan Informatika*, *11*(1).

Ayudhitama, A. P., & Pujianto, U. (2020). Analisa 4 Algoritma dalam Klasifikasi Penyakit Liver Menggunakan Rapidminer. *Jurnal Informatika Polinema*, *6*(2).

Ediana, D., & Sari, N. (2022). Faktor-Faktor yang Berhubungan dengan Kebiasaan Merokok Dalam Rumah. *Jurnal Endurance*, *6*(1), 150–161. https://doi.org/10.22216/jen.v6i1.152

Febriani, S. (2020). Analisis Deskriptif Standar Deviasi. *Jurnal Pendidikan Tambusai*, *6*(1).

Gule, Y. (2022). Edukasi Bahaya Merokok dalam Perspektif Kristen. *Jurnal Abdidas*, *3*(4), 637–643. https://doi.org/10.31004/abdidas.v3i4.635

Gunawan, D. (2016). Evaluasi Performa Pemecahan Database dengan Metode Klasifikasi Pada Data Preprocessing Data mining. *Khazanah Informatika : Jurnal Ilmu Komputer Dan Informatika*, *2*(1), 10–13. https://doi.org/10.23917/khif.v2i1.1749

Husein, H., & Menga, M. K. (2019). Pengetahuan dengan Perilaku Merokok Remaja Knowledge with Adolescent Smoking Behavior. *Jurnal Ilmiah Kesehatan*, *1*(1), 45–50. https://doi.org/10.36590/jika

Indriani, N., Rainarli, E., & Dewi, K. E. (2017). Peringkasan dan Support Vector Machine pada Klasifikasi Dokumen. *Jurnal Infotel*, *9*(4). https://doi.org/10.20895/infotel.v9i4

Jannah, Q. M., & Purwanta, P. (2018). Hubungan Pengetahuan dan Sikap Tentang Rokok dengan Kepatuhan Masyarakat pada Program Rumah Bebas Asap Rokok di Kota Yogyakarta. *Jurnal Keperawatan Klinis Dan Komunitas*, *2*(2), 94. https://doi.org/10.22146/jkkk.44293

Kamulyan, P., Wiguna, P. A., & Slamet, D. A. (2017). Penilaian Keberlanjutan Pengelolaan Sistem Penyediaan Air Minum Berbasis Masyarakat di Kota Blitar. *ITS Journal of Civic Engineering*, *32*(2).

Kementerian Kesehatan. (2022, October). *Temuan Survei GATS : Perokok Dewasa di Indonesia Naik 10 Tahun Terakhir*. Https://Sehatnegeriku.Kemkes.Go.Id/Baca/Umum/20220601/4440021/Temuan-Survei-Gats-Perokok-Dewasa-Di-Indonesia-Naik-10-Tahun-Terakhir/#:~:Text=Dalam%20temuannya%2C%20selama%20kurun%20waktu,Juta%20perokok%20pada%20tahun%202021.

Muraina, I. O. (2022). Ideal Dataset Splitting Ratios in Machine Learning Algorithms: General Concerns For Data Scientists And Data Analysts. *7th International Mardin Artuklu Scientific Researches Conference*. https://www.researchgate.net/publication/358284895

Nasution, D. A., Khotimah, H. H., & Chamidah, N. (2019). Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN. *Computer Engineering, Science and System Journal*, *4*(1), 78. https://doi.org/10.24114/cess.v4i1.11458

Oroh, A. G., Oktamianti, P., Wardiah, R., & Hidayati, F. (2022). Determinan Perilaku Merokok Pada Remaja di Kecamatan Alam Barajo Kota Jambi. *Jurnal Endurance*, *7*(3), 654–660. https://doi.org/10.22216/jen.v7i3.1741

Patandung, Y., & Feriyanto. (2022). Modifikasi Perilaku Merokok Menggunakan Strategi Pengendalian Diri. *Jurnal Sinestesia*, *12*(1). https://sinestesia.pustaka.my.id/journal/article/view/152

Prastiwi, A. (2018). *Estimasi Cadangan Klaim Incurred But Not Reported (IBNR) Menggunakan Metode Chain Ladder dan Pendekatan Over-Dispersed Poisson*. Universitas Islam Indonesia.

Pratama, Y., Rasywir, E., Fachruddin, F., Kisbianty, D., & Irawan, B. (2023). Eksperimen Layer Pooling menggunakan Standar Deviasi untuk Klasifikasi Dataset Citra Wajah dengan Metode CNN. *Building of Informatics, Technology and Science (BITS)*, *5*(1). https://doi.org/10.47065/bits.v5i1.3604

Purbolaksono, M. D., Tantowi, M. I., Hidayat, A. I., & Adiwijaya, A. (2021). Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(2), 393–399. https://doi.org/10.29207/resti.v5i2.3008

Sanhaji, G., Febrianti, A., & Hidayat, H. (2024). Aplikasi DIATECT Untuk Prediksi Penyakit Diabetes Menggunakan SVM Berbasis Web. *Jurnal Tekno Kompak*, *18*(1), 150. https://doi.org/10.33365/jtk.v18i1.3643

Turot, M. . , Polii, B. . , & Walangitan, H. D. (2016). Potensi Pemanfaatan Tumbuhan Paku Diplazium Esculentum Swartz (Studi Kasus) di Distrik Aifat Utara Kabupaten Maybrat Provinsi Papua Barat. *AGRI-SOSIOEKONOMI*, *12*(3A), 1. https://doi.org/10.35791/agrsosek.12.3A.2016.14232

Watratan, A. F., B, A. Puspita., & Moeis, D. (2020). Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia. *Journal of Applied Computer Science and Technology*, *1*(1), 7–14. https://doi.org/10.52158/jacost.v1i1.9

Wibisono, A. D., Dadi Rizkiono, S., & Wantoro, A. (2020). Filtering Spam Email Menggunakan Metode Naive Bayes. *TELEFORTECH : Journal of Telematics and Information Technology*, *1*(1). https://doi.org/10.33365/tft.v1i1.685

World Health Organization. (2020, October). *Pernyataan: Hari Tanpa Tembakau Sedunia 2020*. https://www.who.int/indonesia/news/detail/30-05-2020-pernyataan-hari-tanpa-tembakau-sedunia-2020#:~:text=Setiap%20tahun%2C%20sekitar%20225.700%20orang,mitra%20setiap%20tanggal%2031%20Mei.