Penerapan *Synthetic Minority Oversampling Technique* terhadap Data Perokok Anak di Nusa Tenggara Barat Tahun 2021

Rahma Mutiara Sari[1*], and Achmad Prasetyo[1]
[1]Politeknik Statistika STIS, Jakarta, Indonesia
*Corresponding author: rahmamutiara20@gmail.com

Received: 7 August 2023
Revised: 25 August 2023
Accepted: 15 September 2023

ABSTRAK – **Indonesia menduduki peringkat teratas sebagai negara dengan jumlah perokok usia muda paling banyak di Asia Tenggara. Situasi ini sangat mengkhawatirkan mengingat dampak negatif dari kebiasaan merokok dapat menyebabkan berbagai macam gangguan kesehatan bahkan mengakibatkan kematian. Provinsi Nusa Tenggara Barat merupakan provinsi dengan persentase tertinggi anak yang merokok di Indonesia tahun 2021 yaitu sebesar 2,28%. Data status merokok anak teridentifikasi sebagai data tidak seimbang karena perbandingan antara anak yang merokok dan tidak merokok sangatlah timpang. Oleh karena itu, diterapkan metode analisis regresi logistik biner pendekatan Synthetic Minority Oversampling Technique untuk menangani masalah tersebut. Penelitian ini bertujuan untuk mengetahui gambaran umum dan mengidentifikasi variabel-variabel yang memengaruhi perilaku merokok anak di Nusa Tenggara Barat tahun 2021 beserta kecenderungannya. Data yang digunakan pada penelitian ini adalah data sekunder hasil Survei Sosial Ekonomi Nasional tahun 2021 dengan unit analisis anak usia 5 sampai 17 tahun di Nusa Tenggara Barat tahun 2021. Hasil penelitian menunjukkan bahwa jenis kelamin, status ekonomi, usia, status wilayah tempat tinggal, tingkat pendidikan kepala rumah tangga, dan status bersekolah berpengaruh terhadap perilaku merokok pada anak di Nusa Tenggara Barat tahun 2021 dengan anak yang tidak bersekolah memiliki kecenderungan terbesar untuk merokok.**
Kata kunci– **imbalance data, perilaku merokok anak, regresi logistik, SMOTE.**

ABSTRACT – **Indonesia is ranked as the country with the highest number of young smokers in Southeast Asia. This situation is very worrying considering the negative impact of smoking can cause various health problems and even lead to death. West Nusa Tenggara Province has the highest percentage of children who smoke in Indonesia in 2021 at 2.28%. Data on children's smoking status is identified as unbalanced data because the ratio between children who smoke and do not smoke is very lame. Therefore, the binary logistic regression analysis method of the Synthetic Minority Oversampling Technique approach was applied to handle the problem. This study aims to determine an overview and identify variables that influence children's smoking behavior in West Nusa Tenggara in 2021 and their trends. The data used in this study are secondary data from the 2021 National Socio-Economic Survey with the unit analysis of children aged 5 to 17 years in West Nusa Tenggara in 2021. The results showed that gender, economic status, age, status of region of residence, education level of the head of household, and schooling status influenced children's smoking behavior in West Nusa Tenggara in 2021 with children who didnt attend school having the greatest tendency to smoke.**
Keywords– **imbalance data, child smoking behavior, logistic regression, SMOTE.**

I. PENDAHULUAN

Sustainable Development Goals (SDG's) merupakan suatu rencana aksi global yang disepakati oleh para pemimpin dunia guna untuk menyejahterakan kehidupan masyarakat [1]. SDG's mencakup 17 tujuan, salah satunya pada tujuan ketiga yaitu memastikan kehidupan yang sehat dan sejahtera bagi semua orang di segala kelompok usia. Target pada tujuan ketiga SDG's salah satunya yaitu mengurangi hingga 30 persen angka kematian dini akibat penyakit tidak menular, salah satunya adalah dengan melakukan pengendalian terhadap kebiasaan merokok di setiap negara.

Kebiasaan merokok sudah menjadi hal yang lumrah dilakukan masyarakat dan sudah menjadi salah satu ancaman di bidang kesehatan karena bahan kimia yang terdapat pada rokok akan memicu berbagai penyakit tidak menular bagi perokok aktif seperti gangguan saluran pernapasan, stroke, penyakit jantung, kanker, dan berbagai penyakit lainnya [2]. Tidak hanya perokok aktif, efek dari rokok juga ikut dirasakan oleh perokok pasif. Paparan asap rokok pada perokok pasif juga dapat meningkatkan risiko terjadinya berbagai masalah kesehatan, seperti *pneumonia*, penyakit jantung, dan berbagai penyakit lainnya [3]. *WHO Report on the Global Tobacco Epidemic* tahun 2021 menyebutkan bahwa terdapat sebanyak 8,7 juta kematian yang disebabkan oleh perilaku merokok. Lebih dari 7 juta kematian tersebut disebabkan oleh perokok aktif dan sisanya meninggal karena terpapar asap rokok orang lain (perokok pasif).

Indonesia menduduki peringkat teratas sebagai negara dengan jumlah perokok paling tinggi di Asia Tenggara [4]. Terdapat sebanyak 124,3 juta jiwa perokok dewasa di kawasan Asia Tenggara dan lebih dari setengahnya tinggal di Indonesia yaitu sekitar 65,7 juta jiwa. Tidak hanya perokok dewasa, perokok usia muda penduduk Indonesia juga menempati urutan pertama sebagai negara dengan jumlah perokok paling banyak, yaitu sebanyak 19,2 persen diikuti oleh Malaysia sebanyak 13,2 persen dan Thailand yaitu sebanyak 11,3 persen. Jumlah perokok yang tinggi pada wilayah yang padat penduduk tentunya akan menjadi penyebab utama morbiditas dan mortalitas.

Berdasarkan Profil Statistik Kesehatan yang dipublikasikan oleh Badan Pusat Statistik dengan menggunakan data SUSENAS, di Indonesia terdapat sebanyak 23,78 persen penduduk berusia lima tahun keatas yang merokok tahun 2021. Artinya, ada sekitar 23 sampai 24 orang dari 100 penduduk Indonesia berusia lima tahun ke atas yang merokok pada tahun 2021. Pada tahun 2020 terdapat penurunan jumlah perokok usia 5 tahun ke atas di Indonesia, di mana ini

merupakan suatu hal yang bagus dan harus dipertahankan. Namun, pada tahun 2021 persentase jumlah perokok usia 5 tahun ke atas di Indonesia meningkat sebesar 0,57 persen dari tahun sebelumnya menjadi 23,78 persen. Meskipun sudah banyak yang mengetahui efek berbahaya dari merokok, namun faktanya jumlah perokok masih terus bertambah [5].

Peningkatan jumlah perokok di kalangan masyarakat Indonesia sangat dipengaruhi oleh usia seseorang saat pertama kali mengonsumsi rokok yang semakin dini. Indonesia dikenal sebagai *baby smoker country* sebab Indonesia merupakan salah satu negara dengan prevalensi merokok anak tertinggi dan kasus perokok anak di bawah usia 10 tahun di Indonesia akan cenderung terus bertambah [6]. Selain berdampak buruk bagi tumbuh kembang anak, mengonsumsi rokok pada usia dini juga dapat menyebabkan perokok seumur hidup karena rokok bersifat adiktif. Hal ini didukung oleh Nasution yang menjelaskan bahwa perilaku merokok ketika masih anak-anak sangat berbahaya untuk kesehatan karena dapat membuat kecanduan sejak dini [7]. Anak yang dimaksudkan penelitian ini mengacu pada definisi Undang-Undang Nomor 35 Tahun 2014 Pasal 1 yang menjelaskan bahwa anak adalah seseorang yang belum berusia 18 tahun, termasuk anak yang masih dalam kandungan.

Pasal 46 Peraturan Pemerintah Republik Indonesia Nomor 109 Tahun 2012 tentang Pengamanan Bahan melarang memerintah anak-anak di bawah usia 18 tahun untuk menjual, membeli, atau mengonsumsi produk tembakau. Tidak hanya itu, Kementerian kesehatan juga menyatakan bahwa pemberitahuan mengenai bahaya merokok sudah dilakukan kepada masyarakat secara luas melalui iklan rokok dan bahkan pada bungkus rokok itu sendiri. Akan tetapi, dengan adanya peraturan dan pemberitahuan tersebut masih ditemukan anak di bawah usia 18 tahun yang mengonsumsi rokok.

Perilaku merokok pada anak terjadi di seluruh wilayah di Indonesia. Berdasarkan Profil Statistik Kesehatan 2021, terdapat sekitar 1,51 persen anak usia 5-17 tahun yang merokok di Indonesia. Nusa Tenggara Barat merupakan provinsi yang memiliki persentase tertinggi anak usia 5-17 tahun yang merokok selama tiga tahun berturut-turut dibandingkan dengan provinsi lainnya walaupun persentase kejadiannya tergolong kecil. Pada tahun 2021, persentase anak yang merokok di Nusa Tenggara Barat yaitu sebesar 2,28 persen [8].

Berdasarkan perilaku merokok tersebut, status merokok diklasifikasikan menjadi dua kategori yaitu merokok dan tidak merokok. Regresi logistik biner merupakan salah satu metode yang dapat digunakan ketika variabel respon (*dependen*) berupa variabel kualitatif (kategorik) yang terdiri dari dua kategori [9]. Regresi logistik sangat populer digunakan karena memiliki beberapa kelebihan, diantaranya yaitu sangat sederhana, mudah diinterpretasikan, dan terbukti dapat menghasilkan hasil yang baik dan akurat. Namun, metode ini masih memiliki kelemahan salah satunya yaitu rentan terhadap *underfitting* dan *overfitting* yang disebabkan oleh rasio yang tidak seimbang antara satu kelas dan kelas lainnya (*imbalanced*) [10].

Status merokok pada anak di Nusa Tenggara Barat tahun 2021 menggambarkan keadaan yang *imbalanced* atau tidak seimbang. *Imbalanced data* terjadi jika jumlah objek pada suatu kelas jauh lebih banyak dibandingkan objek pada kelas lainnya [11]. Penggunaan data yang tidak seimbang *(imbalanced)* akan sangat memengaruhi hasil dalam pembentukan model. Sehingga, penggunaan regresi logistik biner dalam mengklasifikasikan kasus *imbalanced data* akan menjadi kurang akurat dikarenakan kasus *imbalanced* akan cenderung mengklasifikasikan kelas mayoritas dan mengabaikan kelas minoritas. Metode regresi logistik seringkali akan menghasilkan performa yang buruk ketika terdapat perbandingan kelas yang tidak seimbang [12].

Cara paling populer untuk menangani permasalahan data yang tidak seimbang adalah dengan metode *resampling* yaitu berupa *undersampling* atau *oversampling*. Penelitian Batista menyatakan bahwa metode *oversampling* pada umumnya memiliki hasil yang lebih baik jika dibandingkan dengan metode *undersampling* [13]. Namun, penelitian Komori & Eguchi menyatakan bahwa penggunaan metode *oversampling* memiliki kelemahan yaitu mengakibatkan *overfitting* karena metode ini menduplikasi data yang sudah ada sebelumnya, akibatnya terdapat klasifikasi yang informasinya duplikat atau saling tumpang tindih [14].

Untuk menangani kelemahan tersebut, Chawla et al memperkenalkan *Synthetic Minority Oversampling Technique* (SMOTE) yang merupakan pengembangan dari metode *oversampling* [15]. SMOTE merupakan metode paling populer digunakan karena dinilai efektif dan baik untuk mengatasi permasalahan *overfitting* dalam menangani ketidakseimbangan di kelas mayoritas dan kelas minoritas.

Penelitian sebelumnya menunjukkan bahwa terdapat beberapa variabel yang dapat memengaruhi anak untuk merokok. Zahrani & Arcana menyebutkan bahwa status wilayah tempat tinggal, jenis kelamin, jenjang pendidikan, status bekerja, status perkawinan, dan usia pertama kali merokok memengaruhi perilaku merokok remaja tiap hari. Remaja laki-laki yang bekerja dan merokok pertama kali pada usia di bawah 18 tahun memiliki kemungkinan terbesar untuk merokok [16]. Hal ini didukung oleh Kusumawardhani et al. yang pada penelitiannya menjelaskan bahwa jenis kelamin, umur, pendidikan signifikan mempengaruhi perilaku merokok remaja di Indonesia. Remaja laki-laki cenderung untuk berperilaku merokok dibandingkan perempuan. Usia diduga berhubungan positif dengan perilaku merokok remaja. Sedangkan, pendidikan diduga berhubungan negatif dengan perilaku merokok pada remaja [17]. Penelitian Wang et al. meunjukkan bahwa orang dengan tingkat pendidikan yang lebih rendah cenderung untuk merokok setiap hari dibandingkan dengan tingkat pendidikan yang lebih tinggi [18].

Penelitian mengenai variabel yang memengaruhi perilaku merokok yang sudah pernah dilakukan pada penelitian terdahulu tidak mempertimbangkan adanya ketidakseimbangan data. Oleh sebab itu, penelitian ini bertujuan untuk mengetahui gambaran umum anak yang merokok di Nusa Tenggara Barat tahun 2021, serta mengidentifikasi variabel-variabel yang memengaruhi beserta kecenderungannya dengan menggunakan regresi logistik biner pendekatan SMOTE.

II. TINJAUAN PUSTAKA

A. Perilaku Merokok

Rokok adalah hasil olahan daun tembakau yang diolah menjadi silinder dari kertas dengan panjang antara 70 mm sampai 120 mm berdiameter sekitar 10 mm [19]. Unsur utama dari rokok yaitu *nikotin, tar*, dan *karbon monoksida* (CO). Kementerian Kesehatan menyatakan bahwa dalam satu batang rokok terdapat sekitar 4000 bahan kimia beracun yang sangat berbahaya untuk tubuh, 43 diantaranya bersifat karsinogenik atau penyebab kanker ganas. Definisi merokok yang digunakan dalam penelitian ini merujuk pada Survei Sosial Ekonomi Nasional (SUSENAS), yaitu aktivitas membakar tembakau lalu menghirup asapnya dengan menggunakan sebatang rokok atau pipa dalam sebulan terakhir sampai saat pencacahan. Perilaku merokok pada anak disebabkan oleh beberapa faktor.

Teori yang digunakan pada penelitian ini adalah *Problem Behavior Theory* (PBT) yang dikemukakan oleh Jessor. PBT mengacu pada sikap yang secara sosial didefinisikan sebagai masalah, sebagai sumber perhatian, atau sebagai suatu sikap yang menyimpang dari norma sosial dan hukum yang tidak disetujui dan tidak diinginkan oleh hukum masyarakat atau perilaku yang biasanya memunculkan beberapa respon kontrol sosial seperti penolakan [20]. Fokus utama dalam teori ini terletak pada struktur konseptual teorinya yang bersifat kompleks dan komprehensif yang mencakup tiga sistem utama yaitu *the personality system, the perceived-environment system,* dan *the behaviour system*. Setiap sistem terdiri dari variabel yang berfungsi baik sebagai dorongan untuk terlibat dalam perilaku bermasalah maupun sebagai kontrol terhadap keterlibatan dalam *problem behavior*.

B. Regresi Logistik Biner

Regresi logistik biner dapat diartikan sebagai regresi logistik yang memiliki variabel respons dengan dua kategori yaitu kategori sukses dan gagal [9]. Secara matematis, Model probabilitas regresi logistik dengan p variabel bebas adalah sebagai berikut [21]:

$$\pi(x) = \frac{(\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p))}{(1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p))} \quad (1)$$

Peluang kejadian sukses π(x) bersifat non linier sehingga harus dilakukan transformasi agar π(x) menjadi fungsi linier. Transformasi ini dikenal dengan transformasi logit. Ketika beberapa variabel bebas bersifat diskrit dengan skala data nominal, maka variabel bebas yang digunakan disebut dengan variabel dummy. Jika variabel dengan skala data nominal memiliki kemungkinan nilai sebanyak k, maka dibutuhkan sebanyak k - 1 variabel dummy. Sehingga, model logit dari sebanyak p variabel bebas dan variabel ke-j adalah dummy diformulasikan sebagai berikut:

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{l=1}^{k_j - 1} \beta_{jl} D_{jl} + \cdots + \beta_p x_p \quad (2)$$

Keterangan:

$\pi(x)$: peluang kejadian sukses
$g(x)$: tranformasi logit
β_0 : nilai parameter intersep
β_1 : nilai parameter ke-1
x_p : nilai variabel bebas ke-p
p : banyaknya variabel bebas yang diamati

C. Synthetic Minority Oversampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) pertama kali diperkenalkan oleh Nithes V. Chawla pada tahun 2002 yang digunakan untuk mengatasi masalah ketidakseimbangan data. Prinsip yang digunakan pada metode ini adalah memperbanyak jumlah data kelas minoritas agar sebanding dengan kelas mayoritas dengan membangkitkan *synthesis data* atau data buatan yang berasal dari k-tetangga terdekat (*k-nearest neighbour*) [15]. Berikut merupakan langkah-langkah SMOTE dalam menghasilkan data buatan:

a. Temukan k-tetangga terdekat untuk setiap sampel $x_i \in S_{min}$. Tetangga terdekat dapat dihitung menggunakan jarak *Euclidean* pada data numerik atau VDM (*Value Difference Metric*) pada data kategorik.
b. Untuk membuat sampel baru, pilih salah satu tetangga terdekat secara acak.
c. Lalu, temukan perbedaan antara sampel yang dipilih dengan tetangga terdekatnya.
d. Kalikan perbedaan ini dengan angka acak antara nol dan satu.
e. Tambahkan vektor ini ke dalam sampel yang dipilih (x_i).

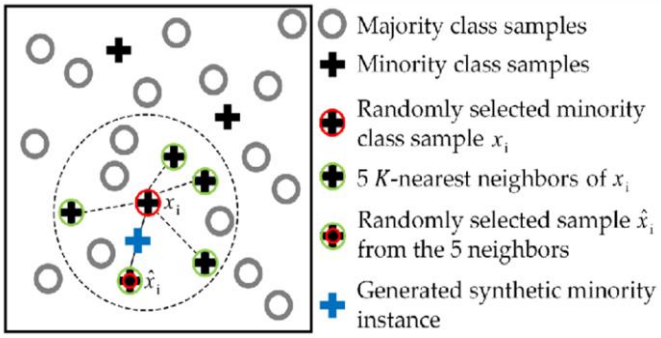
Langkah-langkah tersebut dapat diformulasikan ke dalam bentuk persamaan berikut

$$x_{syn} = x_i + (k_{knn} - x_i) \times \tau \quad (3)$$

Keterangan:

x_{syn} : data sintetis hasil replikasi
x_i : data variabel bebas ke-*i* yang direplikasi
x_{knn} : data yang memiliki jarak terdekat dengan data yang direplikasi
τ : bilangan acak antara nol sampai satu

Untuk data kategorik, maka akan dipilih atribut mayoritas atau modus antara vektor utama yang dipertimbangkan dengan k-tetangga terdekatnya. Jika terjadi nilai sama, maka dipilih secara acak. Selanjutnya, nilai tersebut akan dijadikan atribut data pada kelas buatan baru. Visualisasi cara kerja SMOTE dapat dilihat melalui gambar berikut [22]:



Gambar 1. Ilustrasi SMOTE

D. Evaluasi Ketepatan Klasifikasi

Mengevaluasi ketepatan klasifikasi data merupakan hal yang sangat penting dilakukan untuk mengetahui performa suatu sistem dalam mengklasifikasikan data. Pada hasil klasifikasi yang menyatakan sukses (Y = 1) dan gagal (Y = 0) dapat menggunakan *confusion matrix* atau tabel kesesuaian klasifikasi [23]. *Confusion matrix* akan menghasilkan ukuran ketepatan yaitu sensitivitas (akurasi pada kelas positif) dan spesifisitas (akurasi pada kelas negatif) seperti pada tabel berikut [21]:

Tabel 1. *Confusion matrix*

Nilai Aktual	Nilai Prediksi		Ketepatan
	Positif	Negatif	
Positif	*True Positive* (TP)	*False Negative* (FN)	Sensitivitas = TP/(TP+FN)
Negatif	*False Positive* (FP)	*True Negative* (TN)	Spesifisitas = TN/(TN+FP)
	Akurasi Total		(TP+TN)/(TP+FN+FP+TN)

Selain *confusion matrix,* metode lain yang dapat digunakan dalam mengukur akurasi suatu klasifikasi adalah kurva ROC (*Receiver Operating Characteristic*). Kurva ROC yaitu plot antara *sensitivity* dan 1-*specificity* dalam semua kemungkinan *cut points.* Kurva ini merangkum *predictive power* untuk keseluruhan kemungkinan *cut points,* sehingga lebih informatif bila dibandingkan dengan tabel klasifikasi [24]. Area dibawah kurva ROC disebut dengan AUC (*Area Under the ROC Curve*). AUC menunjukkan kemungkinan amatan untuk kejadian "sukses" lebih tinggi daripada kemungkinan untuk kejadian "gagal". Nilai AUC berada di antara nilai nol dan satu. Apabila nilai AUC mendekati satu, maka model klasifikasi yang terbentuk akan semakin akurat [25].

III. METODOLOGI

Penelitian ini merupakan penelitian kuantitatif dengan menggunakan data sekunder yakni *raw data* dari Survei Sosial Ekonomi Nasional (Susenas) tahun 2021 yang dilaksanakan oleh Badan Pusat Statistik dengan unit analisis anak usia 5-17 tahun di Nusa Tenggara Barat yaitu sebanyak 5891 anak. Dua jenis variabel yang digunakan pada penelitian ini yaitu variabel penjelas (*independent*) dan variabel respon (*dependen*) yang ditunjukkan pada Tabel 2.

Tabel 2. Ringkasan kategori variabel

Variabel	Kategori
Perilaku merokok pada anak (Y)	0 = tidak merokok*, 1 = merokok
Jenis kelamin (X_1)	0 = perempuan*, 1 = laki-laki
Status ekonomi (X_2)	0 = tidak miskin*, 1 = miskin
Usia (X_3)	0 = kanak-kanak (usia 5-11 tahun)*, 1 = remaja (usia 12-17 tahun)
Perilaku merokok ART (X_4)	0 = tidak terdapat ART merokok*, 1 = terdapat ART merokok
Akses internet (X_5)	0 = tidak menggunakan internet*, 1 = menggunakan internet
Status wilayah tempat tinggal (X_6)	0 = perdesaan*, 1 = perkotaan
Tingkat pendidikan KRT (X_7)	0 = ≤SMP/sederajat*, 1 = >SMP/sederajat
Status bersekolah (X_8)	0 = masih bersekolah*, 1 = tidak bersekolah

Keterangan: *) = kategori referensi

IV. HASIL DAN PEMBAHASAN

Gambar 2. Persentase status merokok anak di Nusa Tenggara Barat tahun 2021

Dalam penelitian ini, sampel yang diperoleh yaitu sebanyak 5891 penduduk yang berusia 5-17 tahun. Gambar 2 menunjukkan persentase anak di Nusa Tenggara Barat yang dibagi menjadi dua kategori (merokok dan tidak merokok). Persentase anak yang merokok dihitung dengan cara membagi jumlah anak yang merokok dengan jumlah anak seluruhnya. Hanya terdapat 1,97 persen anak yang merokok dan sisanya sebanyak 98,03 persen tidak merokok. Dari perbandingan jumlah data antara kategori merokok dan tidak merokok mengindikasikan bahwa adanya ketidakseimbangan data pada kasus merokok pada anak di Nusa Tenggara Barat tahun 2021.

Tabel 3. Karakteristik status merokok anak di Nusa Tenggara Barat tahun 2021

No	Nama Variabel	Kategori	Status Merokok Anak	
			Merokok	Tidak Merokok
1	Jenis kelamin (X_1)	Perempuan	0,04%	99,96%
		Laki-laki	3,75%	96,25%
2	Status ekonomi (X_2)	Tidak miskin	2,11%	97,89%
		Miskin	1,29%	98,71%
3	Umur (X_3)	Kanak-kanak	0,03%	99,97%
		Remaja	4,59%	95,41%
4	Perilaku merokok anggota rumah tangga (X_4)	Tidak terdapat ART merokok	1,80%	98,20%
		Terdapat ART merokok	2,04%	97,96%
5	Akses internet (X_5)	Tidak menggunakan internet	0,59%	99,41%
		Menggunakan internet	2,79%	97,21%
6	Status wilayah tempat tinggal (X_6)	Perdesaan	2,31%	97,69%
		Perkotaan	1,52%	98,48%
7	Tingkat pendidikan Kepala Rumah Tangga / KRT (X_7)	≤ SMP/sederajat	2,68%	97,32%
		> SMP/sederajat	0,78%	99,22%
8	Status bersekolah (X_8)	Masih bersekolah	1,65%	98,35%
		Tidak bersekolah	3,71%	96,29%

Sumber: Susenas(diolah)

Tabel 3 menunjukkan gambaran umum karakteristik status merokok anak di Nusa Tenggara Barat tahun 2021. Terlihat bahwa persentase perokok anak laki-laki (3,75 persen) lebih banyak daripada perokok anak perempuan (0,04 persen). Hal ini sesuai dengan Profil Statistik Kesehatan yang menyatakan bahwa perokok anak didominasi oleh laki-laki [26].

Persentase anak dengan status ekonomi tidak miskin yang merokok sebesar 2,11 persen. Jumlah ini lebih besar jika dibandingkan dengan persentase anak dengan status ekonomi miskin yang merokok yaitu sebesar 1,29 persen. Hal ini memperlihatkan bahwa banyak perokok anak berasal dari keluarga dengan status ekonomi tidak miskin. Pada tabel 3 juga dapat dilihat hubungan status merokok anak dengan umur. Persentase anak yang merokok dengan rentang usia 12-17 tahun (4,59 persen) lebih banyak jika dibandingkan dengan anak yang merokok dengan rentang usia 5-11 tahun (0,03 persen). Hal ini sesuai dengan penelitian Pandelaki yang mengemukakan bahwa semakin bertambahnya usia seorang anak, maka semakin tinggi juga persentase perilaku merokok pada anak [27].

Selanjutnya, persentase perokok anak dengan anggota rumah tangga yang juga perokok (2,04 persen) lebih

tinggi daripada perokok anak dengan anggota rumah tangga bukan perokok (1,80 persen). Hal ini sesuai dengan penelitian Maharani & Harsanti yang menjelaskan bahwa remaja yang tinggal pada rumah yang terpapar asap rokok cenderung berperilaku untuk menjadi perokok berat dibandingkan dengan remaja yang tinggal pada rumah yang tidak terpapar asap rokok [28].

Persentase anak yang merokok dan menggunakan internet (2,79 persen) lebih banyak jika dibandingkan dengan anak yang merokok dan tidak menggunakan internet (0,59 persen). Hal ini menunjukkan bahwa anak yang merokok lebih banyak menggunakan internet. Selain itu, berdasarkan hasil pengolahan juga terlihat bahwa persentase perokok anak yang tinggal di perdesaan (2,31 persen) lebih tinggi dibandingkan di perkotaan (1,52 persen). Hal ini sesuai dengan publikasi BPS yang menyatakan bahwa mayoritas anak umur 5-17 tahun yang merokok mayoritas tinggal di daerah perdesaan [8].

Tabel 3 juga menunjukkan bahwa persentase anak yang merokok dengan KRT yang pendidikan terakhirnya adalah pendidikan dasar atau ≤ SMP/sederajat (2,68 persen) lebih banyak jumlahnya dibandingkan KRT berpendidikan menengah keatas atau > SMP/sederajat (0,78 persen). Dari pola antara status merokok anak dengan jenjang pendidikan KRT memperlihatkan bahwa semakin tinggi tingkat pendidikan KRT maka semakin turun persentase perilaku merokok pada anaknya. Terakhir, Persentase anak merokok yang tidak bersekolah (3,71 persen) dua kali lipat lebih banyak dibandingkan dengan anak merokok yang bersekolah (1,65 persen). Hal ini sesuai dengan penelitian Maharani & Harsanti yang mengemukakan bahwa proporsi perokok remaja yang masih bersekolah lebih rendah dibandingkan dengan perokok remaja yang tidak bersekolah [28].

Berdasarkan gambar 3, terlihat bahwa perbandingan jumlah data antara kategori merokok dan tidak merokok pada kasus merokok anak di Nusa Tenggara Barat tahun 2021 mengindikasikan adanya ketidakseimbangan data. Apabila dilakukan pemodelan, data seperti ini akan mengakibatkan hasil estimasinya menjadi tidak representatif. Oleh sebab itu, dilakukan pendekatan SMOTE agar data lebih seimbang.

Pembentukan model pada penelitian ini menggunakan regresi logistik biner dengan data sebelum dan sesudah SMOTE. Berdasarkan pengolahan regresi logistik biner pada data dengan variabel tidak bebas berupa status merokok pada anak, dilakukan pengujian parameter secara simultan menggunakan statistik uji G dengan hasil sebagai berikut

Tabel 4. Omnibus Test of Model Coefficients

Model	***Chi-square***	***Df***	***p-value***
Tanpa SMOTE	298,25	8	0,000*
Dengan SMOTE	318,32	8	0,000*

Sumber: Susenas 2021 (diolah)

Keterangan: *) = signifikan pada taraf uji 5 persen

Berdasarkan hasil simultan pada tabel 4 diperoleh *p-value* untuk model tanpa SMOTE dan dengan SMOTE adalah 0,000. Dikarenakan nilai *p-value* kurang dari 0,05 maka disimpulkan bahwa dengan taraf uji 5 persen, cukup bukti untuk menyatakan minimal terdapat satu variabel bebas yang berpengaruh terhadap perilaku merokok pada anak di Nusa Tenggara Barat tahun 2021. Setelah itu, dilakukan pengujian parameter secara parsial untuk mengetahui apakah masing-masing variabel penjelas berpengaruh terhadap perilaku merokok anak. Diperoleh hasil pendugaan parameter yang ditunjukkan pada tabel 5.

Tabel 5. Hasil pendugaan parameter regresi logistik biner

Variabel	Tanpa SMOTE			Dengan SMOTE		
	$\hat{\beta}$	*Standard error*	*p-value*	$\hat{\beta}$	*Standard error*	*p-value*
Konstanta	-12,539	1,485	0,000	-6,685	1,071	0,000
Jenis Kelamin (X_1)	4,641	1,011	0,000*	3,693	0,447	0,000*
Status Ekonomi (X_2)	-0,467	0,383	0,223	0,969	0,441	0,028*
Umur (X_3)	5,045	1,026	0,000*	5,423	0,849	0,000*
Perilaku Merokok ART (X_4)	-0,086	0,275	0,754	-0,373	0,366	0,309
Akses Internet (X_5)	0,993	0,381	0,009*	0,104	0,424	0,807
Status Wilayah Tempat Tinggal (X_6)	-0,191	0,256	0,442	-0,737	0,351	0,036*
Tingkat Pendidikan KRT (X_7)	-1,327	0,357	0,000*	-1,581	0,396	0,000*
Status Bersekolah (X_8)	1,973	0,337	0,000*	1,700	0,571	0,003*

Sumber: Susenas 2021 (diolah)

Keterangan: *) = signifikan pada taraf uji 5 persen

Berdasarkan hasil pengujian parsial pada tabel 5, hasil dari model tanpa SMOTE jika dilihat dari *p-value* yang nilainya kurang dari 0,05 diperoleh 5 variabel yang berpengaruh terhadap perilaku merokok pada anak di Nusa

Tenggara Barat tahun 2021. Sedangkan untuk hasil dari model dengan SMOTE diperoleh 6 variabel yang berpengaruh terhadap perilaku merokok pada anak di Nusa Tenggara Barat tahun 2021.

Setelah dilakukan analisis, selanjutnya dilakukan perbandingan hasil evaluasi untuk model regresi logistik biner tanpa SMOTE dan dengan SMOTE. Ukuran evaluasi yang digunakan untuk membandingkan metode meliputi nilai sensitivitas, spesifisitas, akurasi, dan AUC.

Tabel 6. Perbandingan hasil evaluasi model tanpa SMOTE dan model dengan SMOTE

Pendekatan	Sensitivitas	Spesifisitas	Akurasi	AUC
Tanpa SMOTE	23,53%	99,65%	98,19%	0,616
Dengan SMOTE	97,06%	80,14%	80,46%	0,886

Berdasarkan hasil perbandingan pada tabel 6, analisis regresi logistik biner tanpa pendekatan SMOTE menghasilkan nilai akurasi yang lebih tinggi dibandingkan dengan regresi logistik biner dengan pendekatan SMOTE. Namun, ternyata akurasi saja belum cukup dalam mengukur ketepatan model. Jika dilihat dari nilai sensitivitas dan spesifisitas pada model tanpa SMOTE, terindikasi adanya ketimpangan yaitu 23,53 persen dan 99,65 persen. Hal tersebut menunjukkan bahwa model analisis regresi logistik biner tanpa pendekatan SMOTE tidak dapat mengklasifikasikan kelas minoritas dengan baik. Sedangkan, pada analisis logistik biner dengan pendekatan SMOTE, nilai sensitivitasnya meningkat dan cenderung seimbang dengan nilai spesifisitasnya serta memiliki nilai AUC yang lebih tinggi jika dibandingkan dengan model tanpa SMOTE. Sehingga, dapat disimpulkan bahwa analisis regresi logistik biner dengan pendekatan SMOTE lebih baik digunakan dan mencirikan karakteristik anak yang merokok di Nusa Tenggara Barat tahun 2021. Berdasarkan nilai $\hat{\beta}$ pada tabel 5, dibentuk persamaan regresi logistik biner untuk status merokok pada anak sebagai berikut:

$$\hat{g}(x) = -6{,}685 + 3{,}693X_1 + 0{,}969X_2 + 5{,}423X_3 - 0{,}373X_4 + 0{,}104X_5 - 0{,}737X_6 - 1{,}581X_7 + 1{,}700X_8$$

Menginterpretasikan hasil estimasi model regresi logistik biner dapat melihat rasio kecenderungan (*odds ratio*) dari setiap variabel bebas. *Odds Ratio* untuk masing-masing variabel bebas dihitung dengan exp ($\hat{\beta}$) berdasarkan nilai dari masing-masing koefisien $\hat{\beta}$. Sehingga interpretasi *odds ratio* untuk tiap variabel bebas adalah sebagai berikut:

a. *Odds ratio* untuk variabel status ekonomi ialah sebesar exp (0,969) = 2,635 yang artinya bahwa anak yang berstatus miskin memiliki kecenderungan untuk merokok 2,635 kali lebih besar dibandingkan dengan anak yang berstatus tidak miskin. Hal ini sejalan dengan temuan Amponsah et al. yang mengemukakan bahwa laki-laki berstatus miskin cenderung untuk merokok [29]. Adanya perubahan arah koefisien regresi setelah dilakukan pendekatan SMOTE pada variabel status ekonomi dan arah hubungan variabel yang tidak bersesuaian dengan hasil analisis deskriptif juga terjadi pada penelitian Albaihaqi [30].
b. *Odds ratio* untuk variabel status wilayah tempat tinggal ialah sebesar exp (-0,737) = 0,479 artinya anak yang tinggal di perkotaan 0,479 kali lebih kecil kemungkinannya untuk merokok dibandingkan di perdesaan. Atau dengan kata lain, anak yang tinggal di perdesaan memiliki kecenderungan untuk merokok sebesar (1/0,479) yaitu sebesar 2,089 kali lebih besar dibandingkan dengan anak yang tinggal di perkotaan.
c. *Odds ratio* untuk variabel tingkat pendidikan KRT ialah sebesar exp (-1,581) = 0,206 artinya bahwa anak dengan KRT berpendidikan menengah keatas 0,206 kali lebih kecil kemungkinannya untuk merokok dibandingkan anak dengan KRT yang pendidikan terakhirnya adalah pendidikan dasar. Atau dengan kata lain, anak dengan KRT yang pendidikan terakhirnya adalah pendidikan dasar memiliki kecenderungan untuk merokok sebesar (1/0,206) yaitu sebesar 4,859 kali lebih besar dibandingkan dengan anak dengan KRT berpendidikan menengah keatas. Sejalan dengan temuan Purnaningrum et al. dan Tyas & Pederson yaitu semakin rendah jenjang pendidikan orang tua akan meningkatkan perilaku merokok pada remaja [31][32].
d. *Odds ratio* untuk variabel status bersekolah ialah sebesar exp (1,700) = 5,474 artinya bahwa anak tidak bersekolah memiliki kecenderungan untuk merokok 5,474 kali lebih besar dibandingkan dengan anak yang masih bersekolah.

V. KESIMPULAN DAN SARAN

Hasil evaluasi model menunjukkan bahwa analisis regresi biner pendekatan SMOTE dapat digunakan untuk mengatasi ketidakseimbangan pada data. Terdapat 6 variabel bebas yang memengaruhi perilaku merokok pada anak, diantaranya jenis kelamin, status ekonomi, usia, status wilayah tempat tinggal, tingkat pendidikan KRT, dan status bersekolah. Kecenderungan anak berjenis kelamin laki-laki, status ekonomi miskin, usia 12-17 tahun, tinggal di wilayah perdesaan, memiliki KRT yang pendidikan terakhirnya adalah pendidikan dasar, dan tidak bersekolah lebih besar untuk berperilaku merokok dengan anak yang tidak bersekolah memiliki kecenderungan terbesar untuk merokok. Sosialiasi mengenai pentingnya bersekolah perlu ditingkatkan dengan harapan pikiran dari orang tua bisa terbuka dan menyadari akan pentingnya bersekolah. Pemberian dana kepada orang yang tidak mampu agar dapat menikmati pendidikan yang layak juga penting dilakukan dengan harapan anak yang putus sekolah atau tidak bersekolah lagi bisa melanjutkan pendidikannya.

REFERENCES

[1] BPS, "Indikator Tujuan Pembangunan Berkelanjutan 2021," pp. 1–253, 2021.
[2] WHO, "WHO Report on The Global Tobacco Epidemic," *Heal. Promot.*, 2021, [Online]. Available: https://www.who.int/teams/health-promotion/tobacco-control/global-tobacco-report-2021.
[3] BPS, *Profil Anak Usia Dini 2020*. 2020.
[4] SEATCA, "The tobacco control atlas: ASEAN region," *Southeast Asia Tob. Control Alliance*, no. December, 2021, [Online]. Available: https://seatca.org/dmdocuments/SEATCA ASEAN Tobacco Control Atlas_5th Ed.pdf.
[5] D. Komasari and A. F. Helmi, "Faktor Faktor Penyebab Merokok Pada Remaja," *J. Psikol.*, vol. 27, no. 1, pp. 37–47, 2011.
[6] S. Rezeki and D. M. Utari, "Faktor-Faktor yang Mempengaruhi Perilaku Merokok Pada Anak Sekolah Dasar di SD Pinggiran Banda AcehTahun 2021," *J. Healthc. Technol. Med.*, vol. 47, no. 4, pp. 124–134, 2021, doi: 10.31857/s013116462104007x.
[7] I. K. Nasution, "PERILAKU MEROKOK PADA REMAJA," *Rev. Esp. Enfermedades Dig.*, vol. 94, no. 2, pp. 101–103, 2007.
[8] BPS, "Profil Statistik Kesehatan 2021," *Badan Pus. Stat.*, p. 404, 2021, [Online]. Available: bps.go.id.
[9] A. Agresti, "Categorical data analysis (Vol. 792).," *John Wiley Sons*, 2012.
[10] P. Harrington, "Machine Learning in Action," in *New York: Manning Publications Co*, 2012.
[11] J. Brownlee, "Data Preparation for Machine Learning," *San Fr. Mach. Learn. Mastery.*, 2020.
[12] G. King and L. Zeng, "Logistic Regression in Rare Events Data," *Polit. Anal.*, vol. 9, no. 2, pp. 137–163, 2001, doi: 10.1093/oxfordjournals.pan.a004868.
[13] G. E. Batista, R. C. Prati, and M. C. Monard, "Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *ACM SIGKDD Explor. newsletter, 6(1), pp.20-29*, 2004.
[14] O. Komori and S. Eguchi, "Statistical methods for imbalanced data in ecological and biological studies," *Springer Japan*, 2019.
[15] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 30, no. 2, pp. 321–357, 2002, doi: 10.1002/eap.2043.
[16] C. I. Zahrani and I. M. Arcana, "Determinan Perilaku Remaja Merokok Setiap Hari Di Indonesia," *Semin. Nas. Off. Stat.*, vol. 2020, no. 1, pp. 519–528, 2021, doi: 10.34123/semnasoffstat.v2020i1.412.
[17] N. Kusumawardhani, I. Tarigan, Suparmi, and A. Schlotheuber, "Socio-Economic, demographic and geographic coreelates of cigarette smoking among Indonesian adolescent: result from the 2013 Indonesian Basic Health Research (RISKESDAS) survey.," 2018.
[18] Y. Wang, H. Y. Sung, T. Yao, & Lightwood, J., and W. Max, "Infrequent and frequent nondaily smokers and daily smokers: their characteristics and other tobacco use patterns," pp. 741–748, 2018.
[19] F. . Aula, "Stop Merokok," 2010.
[20] R. Jessor, *Problem-Behavior Theory*, vol. 2. 2001.
[21] J. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, "Applied logistic regression (Vol. 398)," *John Wiley Sons*, 2013.
[22] R. P. Saputri, W. S. Winahju, and K. Fithriasari, "Klasifikasi Sentimen Wisatawan Candi Borobudur pada Situs TripAdvisor Menggunakan Support Vector Machine dan K-Nearest Neighbor," *J. Sains dan Seni ITS*, vol. 8, no. 2, pp. 349–356, 2020.
[23] L. A. Andika, P. A. N. Azizah, and R. Respatiwulan, "Analisis Sentimen Masyarakat terhadap Hasil Quick Count Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier," *Indones. J. Appl. Stat.*, pp. 34–41, 2019.
[24] C. S. Imanwardhani, "Pendekatan Synthetic Minority Oversampling Technique Dalam Menangani Klasifikasi Imbalanced Data Biner (Studi Kasus: Status Ketertinggalan Desa di Jawa Timur)," 2018.
[25] F. Gorunescu, "Data Mining: Concepts, models and techniques," *Springer Sci. Bus. Media.*, 2011.
[26] BPS, "Konsep dan Definisi Survei Sosial Ekonomi Nasional Maret 2021," 2021.
[27] S. C. Pandelaki, "Determinan Perilaku Merokok Pada Anak di Indonesia Tahun 2020," 2022.
[28] V. Maharani and T. Harsanti, "Variabel-Variabel yang Mempengaruhi Intensitas Merokok Remaja Pria di Indonesia Tahun 2017 (Variables that affect the smoking intensity of male adolescents in Indonesia in 2017)," vol. 2017, pp. 821–830, 2021.
[29] N. E. Amponsah, G. Afful-Mensah, and S. Ampaw, "Deteminants of cigarette smoking and smoking intensity among adult males in Ghana," *BMC Public Health*, vol. 18, no. 941, 2018.
[30] R. Albaihaqi, *Determinan Perilaku Merokok Anak di Jawa Barat Tahun 2019*. 2021.
[31] W. D. Purnaningrum, H. Joebagio, and B. Murti, "Association between cigarette advertisement, peer group, parental education, family income, and pocket money with smoking behavior among adolescents in Karanganyar District, Central Java," *J. Heal. Promot. Behav.*, vol. 2, no. 2, pp. 148–158, 2017.
[32] S. L. Tyas and L. L. Pederson, "Psychosocial factors related to adolescent smoking: a critical review of the literature," *Tob. Control*, vol. 7, no. 4, pp. 409–420, 1998.