Perbandingan Metode Klasifikasi dengan Teknik Resampling pada Kejadian Gagal Ginjal Kronis di Kalimantan Utara

Putri Ramatillah Ramadhana[1)], Nofita Istiana[2)*]
[1)] Badan Pusat Statistik
[2)] Politeknik Statistika STIS

Received: 2 June 2025 Accepted: 23 June 2025 Published: 1 July 2025

* nofita@stis.ac.id

Kata Kunci: Gagal Ginjal Kronis, Teknik Resampling, Metode Klasifikasi

DSI: Jurnal Data Science Indonesia is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

Abstrak : Gagal ginjal kronis adalah penyakit tidak menular yang menjadi permasalahan global. Di Indonesia, Kalimantan Utara memiliki prevalensi kejadian gagal ginjal kronis tertinggi yaitu sebesar 0,64%. Penduduk usia produktif di provinsi tersebut merupakan salah satu kelompok penduduk yang rentan terhadap penyakit ini. Gagal ginjal kronis lebih banyak terjadi pada perempuan, namun perkembangan keparahannya lebih cepat pada laki-laki dimana prevalensi penduduk laki-laki yang melakukan hemodialisis adalah sebesar 18,41% sementara perempuan 0%. Fakta ini disayangkan karena penduduk laki-laki lebih banyak yang bekerja dibanding perempuan dan memiliki produktivitas lebih tinggi. Oleh karena itu, penelitian ini bertujuan mengetahui variabel penting yang berperan pada kejadian gagal ginjal kronis penduduk laki-laki usia produktif melalui pemodelan klasifikasi dan penerapan teknik *resampling* terbaik. Penerapan teknik *resampling* perlu dilakukan karena data yang digunakan memiliki kasus ketidakseimbangan kelas. Teknik *resampling* yang dibandingkan adalah SMOTE dan SMOTE + Tomek Links dengan metode klasifikasi *Random Forest*, *Naïve Bayes*, regresi logistik, dan regresi logistik Firth. Hasil penelitian ini, metode klasifikasi terbaik adalah *Naïve Bayes* dan teknik *resampling* terbaiknya adalah SMOTE. Variabel penting yang berperan pada kejadian gagal ginjal kronis laki-laki usia produktif ini adalah usia, konsumsi minuman energi, klasifikasi wilayah, konsumsi makanan instan, dan konsumsi minuman manis.

PENDAHULUAN

Kesehatan merupakan kunci utama yang penting untuk diperhatikan agar sumber daya manusia yang dimiliki Indonesia mempunyai kehidupan yang berkualitas dan dapat berperan dalam membangun kesejahteraan masyarakat. Hal ini sesuai dengan tujuan ke-3 dari *Sustainable Development* Goals (SDGs) yaitu menjamin kehidupan sehat serta mewujudkan kesejahteraan bagi semua orang di segala usia. Salah satu permasalahan kesehatan yang menjadi ancaman serius bagi kehidupan manusia adalah penyakit tidak menular (PTM). Penyakit ini terjadi dalam durasi yang panjang dan merupakan penyebab dari 74% kematian di seluruh dunia [1].

Gagal ginjal kronis adalah penyakit tidak menular yang terjadi ketika ginjal mengalami kerusakan atau penurunan fungsi setidaknya dalam tiga bulan apapun penyebabnya [2]. Kerusakan ginjal umumnya mengacu pada kelainan ginjal yang dapat disimpulkan melalui pencitraan, biopsi atau penanda klinis misalnya peningkatan albuminuria (rasio albumin terhadap kreatinin lebih dari 30 mg/g) atau perubahan sedimen urin. Penurunan fungsi ginjal adalah ketika laju filtrasi glomerulus yang bernilai kurang dari 60 ml/min/1,73 m^2 [3].

Menurut jurnal The Lancet, penyakit gagal ginjal kronis merupakan penyebab kematian 1,2 juta orang secara global. Penyakit ini menempati posisi ke-12 sebagai penyakit penyebab kematian di dunia tertinggi dengan prevalensi sebesar 9,1%. Secara kasar, terdapat sekitar 700 juta kasus penyakit ini pada tahun 2017

[4]. Sementara menurut World Health Organization (WHO), penyakit ginjal berada pada peringkat 10 besar sebagai penyakit penyebab kematian tertinggi dengan kematian mencapai 1,3 juta pada tahun 2019 [5]. Di Indonesia, prevalensi gagal ginjal kronis meningkat dari 0,2% pada tahun 2013 menjadi 0,38% pada tahun 2018 [6]. Angka tersebut meningkat hampir 2 kali lipat dalam rentang waktu 5 tahun saja. Penyakit gagal ginjal kronis juga menempati posisi ke-4 sebagai penyakit yang paling banyak menyerap dana jaminan sosial dengan 1,93 juta kasus dan menghabiskan dana sebesar 2,79 triliun rupiah pada tahun 2019 [7].

Prevalensi gagal ginjal kronis terbanyak pada tahun 2018 adalah Kalimantan Utara. Prevalensi gagal ginjal kronis di provinsi tersebut mencapai 0,64%. Penduduk usia produktif merupakan salah satu kelompok penduduk yang memiliki prevalensi kejadian gagal ginjal kronis tertinggi di provinsi ini. Prevalensi kejadian gagal ginjal kronis tertinggi pada penduduk usia produktif berada pada usia 55-64 tahun dan 45-54 tahun. Kejadian gagal ginjal kronis di provinsi ini lebih banyak terjadi pada perempuan dibandingkan laki-laki, yaitu berturut-turut 0,71% dan 0,58%. Meskipun demikian, penduduk laki-laki yang mengidap gagal ginjal kronis memiliki proporsi yang lebih besar untuk melakukan hemodialisis. Proporsi laki-laki yang melakukan hemodialisis adalah sebesar 18,41% sementara perempuan 0%. Ini menunjukkan bahwa penduduk laki-laki yang mengalami gagal ginjal kronis memiliki perkembangan keparahan yang lebih cepat dibandingkan perempuan. Hal ini sejalan dengan penelitian yang menyatakan bahwa laki-laki mengalami perkembangan gagal ginjal kronis yang cepat [8]. Kejadian gagal ginjal kronis dan tingkat keparahan yang cepat berkembang pada laki-laki di Kalimantan Utara seharusnya dihindari. Penduduk laki-laki di Kalimantan Utara berjumlah lebih banyak dibandingkan penduduk perempuan. Penduduk usia produktif di provinsi tersebut pada tahun 2022 terdiri dari 291.712 laki-laki dan 255.458 perempuan. Sebanyak 232.375 penduduk laki-laki pada usia produktif berstatus bekerja, sementara perempuan sebanyak 121.609 orang [9]. Angka ini menunjukkan bahwa penduduk laki-laki secara ekonomi lebih produktif dibandingkan perempuan di provinsi tersebut. Keberadaan penyakit gagal ginjal kronis pada laki-laki usia produktif ini akan memiliki efek pada produktivitas penderita dan kesejahteraan dirinya sendiri ataupun orang yang menjadi tanggungannya.

Dari penjabaran sebelumnya, gagal ginjal kronis merupakan penyakit yang dapat mengancam kesejahteraan masyarakat, bukan hanya dari segi kesehatan tetapi juga dari segi ekonomi. Dengan demikian, variabel penting yang berperan pada kejadian gagal ginjal kronis perlu untuk diketahui dimana pengetahuan tersebut dapat digunakan untuk mencegah kenaikan angka prevalensi penyakit gagal ginjal kronis. Salah satu cara untuk mengetahui variabel penting tersebut adalah melalui suatu model klasifikasi. Namun, kejadian gagal ginjal kronis pada laki-laki usia produktif di Kalimantan Utara pada tahun 2018 memiliki data dengan kasus ketidakseimbangan kelas sehingga metode klasifikasi standar tidak dapat bekerja dengan baik pada data dengan kasus demikian. Variabel dependen yaitu status gagal ginjal kronis memiliki jumlah observasi kelas minoritas yang berbeda jauh dari kelas mayoritasnya. Persentase kelas minoritas berada di bawah 1% sehingga data tersebut disebut dengan *extremely imbalanced data* [10]. Penerapan teknik *resampling* merupakan solusi yang dapat dilakukan untuk meningkatkan kinerja dari metode klasifikasi yang dipakai, yaitu *Random Forest*, *Naïve Bayes*, regresi logistik dan regresi logistik Firth. Teknik *resampling* yang digunakan pada penelitian ini adalah SMOTE dan SMOTE + Tomek Links. Dengan demikian, tujuan penelitian ini untuk mengetahui metode klasifikasi dan teknik *resampling* terbaik untuk mengklasifikasikan kejadian gagal ginjal kronis serta mengetahui variabel penting yang berperan pada pengklasifikasian tersebut.

Sebelumnya telah terdapat penelitian mengenai pengklasifikasian kejadian gagal ginjal kronis dengan metode *Random Forest*, *Naïve Bayes*, dan regresi logistik. Metode ini telah digunakan oleh beberapa peneliti untuk memprediksi kejadian gagal ginjal kronis dan menghasilkan akurasi yang baik [11][12]. Data yang digunakan pada penelitian tersebut berasal dari *UCI Machine Learning Repository*, dimana data tersebut seimbang. Keadaan ini berbeda dengan kasus kejadian gagal ginjal kronis yang terjadi di Kalimantan Utara pada tahun 2018 yang merupakan data yang jarang dan tidak seimbang.

Penelitian lainnya yang meneliti kinerja klasifikasi pada data yang tidak seimbang adalah penelitian yang mengklasifikasikan masyarakat penerima bantuan langsung tunai (BLT) desa dengan *Naïve Bayes* dan SMOTE. Hasil *Naïve Bayes* dengan SMOTE menghasilkan akurasi sebesar 97,8% dan AUC sebesar 0,99. Model ini lebih baik dibandingkan dengan pemodelan *Naïve Bayes* tanpa SMOTE [13]. Sementara itu,

terdapat penelitian yang menggunakan teknik *resampling* SMOTE dan klasterisasi k-*means* sebagai bagian dari *preprocessing data*. Metode klasifikasi yang digunakan adalah regresi logistik untuk mengklasifikasikan kejadian diabetes pada pasien perempuan pada data Pima Indian Diabetes, data ini memiliki kasus ketidakseimbangan kelas. Akurasi yang didapatkan bernilai 99,6% untuk pemodelan dengan SMOTE, sementara pemodelan tanpa SMOTE memiliki akurasi bernilai 99,3%. Ini menunjukkan bahwa metode *preprocessing* dengan SMOTE dan k-*means* berhasil membuat kinerja metode klasifikasi menjadi lebih baik [14].

Kemudian, terdapat penelitian yang mengungkapkan faktor yang berpengaruh pada kejadian gagal ginjal kronis pada pasien dengan hemodialisis di RSUD Arjawinangun dan Waled Kabupaten Cirebon. Hasil dari penelitian ini adalah faktor risiko yang memiliki hubungan bermakna dengan penyakit gagal ginjal kronis diantaranya pendidikan, pekerjaan intelektual, riwayat penyakit hipertensi, riwayat penyakit diabetes melitus, konsumsi air putih, konsumsi minuman yang mengandung gula, dan konsumsi makanan yang mengandung garam tinggi [15]. Penelitian lainnya juga menyatakan bahwa jenis kelamin, usia, diabetes melitus, obesitas, dan status ekonomi berpengaruh signifikan pada kejadian gagal ginjal kronis [16].

TINJAUAN LITERATUR

Gagal ginjal kronis yaitu kerusakan ginjal atau penurunan fungsi ginjal yang terjadi minimal tiga bulan terlepas dari penyebabnya[2]. Kriteria gagal ginjal kronis berdasarkan *Kidney Disease Improving Global Outcomes* (KDIGO) adalah [17]:

1. Apabila ginjal mengalami kerusakan berupa kelainan baik secara struktural maupun fungsional selama lebih dari 3 bulan dengan atau tanpa penurunan laju filtrasi glomerulus, misalnya albuminuria, kelainan sedimen urin, dan gangguan tubulus ginjal.
2. Laju filtrasi glomerulus (LFG) < 60 ml/min/1,73m^2 dalam waktu lebih dari 3 bulan, dengan atau tanpa adanya kerusakan pada ginjal.

Faktor risiko gagal ginjal kronis telah disistematisasi oleh *Australian Institute of Health and Welfare* yang mengelompokkan faktor risiko tersebut menjadi empat kelompok yaitu faktor lingkungan-sosial, biomedik, perilaku dan predisposisi [18]. Dalam penelitian ini, faktor lingkungan sosial terdiri dari variabel klasifikasi wilayah, faktor risiko biomedik terdiri dari penyakit jantung dan diabetes melitus, faktor risiko perilaku terdiri dari konsumsi minuman manis, konsumsi minuman berenergi, dan konsumsi makanan instan, dan faktor predisposisi terdiri dari variabel usia.

Data tidak seimbang atau *imbalanced data* merupakan salah satu permasalahan klasifikasi dimana jumlah observasi dari suatu kelas lebih rendah dibandingkan kelas yang lain [19]. Dataset disebut *extremely imbalanced* ketika rasio ketidakseimbangannya bernilai 100:1, 1000:1, dan 10.000:1 [10]. Ini berarti ketika kelas minoritas bernilai 1% atau kurang dari 1% dari seluruh observasi. Ketidakseimbangan kelas ekstrim ini dapat memengaruhi kinerja dari metode klasifikasi untuk mengabaikan kelas minoritas dan cenderung mengklasifikasikan data ke kelas mayoritas. Hal ini terjadi karena metode klasifikasi tidak dapat mempelajari karakteristik data pada kelas minoritas dengan baik karena jumlahnya yang berbeda jauh dengan kelas mayoritas. Akibatnya, data yang seharusnya berada pada kelas minoritas akan diklasifikasikan oleh metode klasifikasi menjadi berada pada kelas mayoritas. Solusi yang dapat diterapkan salah satunya adalah menyeimbangkan kembali dataset aslinya melalui *oversampling* kelas minoritas atau *undersampling* kelas mayoritas [19].

Teknik *Resampling*

a. SMOTE (*Synthetic Minority Oversampling Technique*)

Metode SMOTE dapat mengatasi kekurangan dari metode *Random Oversampling* yaitu permasalahan *overfitting* yang terjadi karena metode tersebut mereplikasi data dari kelas minoritas [20]. Metode SMOTE menyeimbangkan kelas data dengan membentuk sampel sintetis pada kelas minoritas berdasarkan sampel dari kelas tersebut dan k tetangga terdekatnya (Chawla et al., 2002).

Sampel sintetis pada data dengan tipe kategorik dibentuk berdasarkan modus dari nilai variabel sampel minoritas dan k tetangga terdekatnya. Sementara penentuan tetangga terdekat dilakukan dengan VDM (*Value Difference Metric*) [21].

$$\Delta(X,Y) = \sum_{i=1}^{p} \delta(V_{1i}, V_{2i}) \quad (2\text{-}1)$$

Dimana:

$\Delta(X,Y)$: Jarak antara amatan X dan amatan Y
$\delta(V_{1i}, V_{2i})$: Jarak antara amatan X dan Y untuk tiap variabel prediktor
p : Banyaknya variabel prediktor.

Jarak antara amatan X dan amatan Y untuk tiap variabel dihitung dengan rumus sebagai berikut [21].

$$\delta(V_1, V_2) = \sum_{i=1}^{k} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right| \quad (\textbf{Error! No text of specified style in document.}\text{-}2)$$

Dimana:

k : Banyaknya kelas pada variabel dependen
C_{1i} : Banyaknya kemunculan nilai variabel prediktor V_1 pada kelas ke$-i$
C_{2i} : Banyaknya kemunculan nilai variabel prediktor V_2 pada kelas ke$-i$
C_1 : Total kemunculan nilai variabel prediktor V_1 dalam dataset
C_2 : Total kemunculan nilai variabel prediktor V_2 dalam dataset

b. SMOTE + Tomek Links

Tomek Links merupakan metode *undersampling* yang bekerja dengan menghilangkan sampel kelas mayoritas yang menginvasi kelas minoritas [22]. Tautan Tomek (Tomek link) merupakan sebutan untuk dua observasi dari dua kelas yang berbeda, yaitu kelas minoritas dan mayoritas, yang memiliki jarak terpendek. Jarak terpendek ini dihitung dan ditetapkan dengan algoritma *K-Nearest Neighbors* [23]. Pada penelitian ini penentuan tetangga terdekat pada Tomek Links sama dengan pada SMOTE.

Metode Klasifikasi

a. *Random Forest*

Random Forest merupakan metode klasifikasi dengan mengkombinasikan kumpulan pohon keputusan dimana tiap pohon dibangun berdasarkan sampel acak yang berasal dari data aslinya [24]. Pembentukan *node* pada setiap pohon *Random Forest* dibentuk berdasarkan nilai *entropy* dan *gain* yang dihasilkan dari tiap variabel dengan rumus sebagai berikut [25].

$$Entropy = -\sum_{i=1}^{n} p_i \log_2 p_i \quad (\textbf{Error! No text of specified style in document.}\text{-}3)$$

$$Gain\ (A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S_i) \quad (\textbf{Error! No text of specified style in document.}\text{-}4)$$

Dimana:

A: Variabel tertentu yang akan dihitung *gain*-nya
S: Total jumlah data
S_i: Total jumlah data pada kategori ke-i dari variabel tertentu
p_i: Peluang data di suatu kelas dibandingkan dengan total data pada variabel yang ingin dihitung *entropy*-nya.

b. *Naïve Bayes*

Naïve Bayes atau disebut juga Bayesian *classifiers* merupakan metode klasifikasi statistik yang memprediksi peluang keanggotaan sebuah kelas berdasarkan variabel prediktor yang ada. *Naïve Bayes* melakukan pengklasifikasian data berdasarkan Bayes' *theorem* [26].

$$P(H|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|H)\ P(H)}{P(\boldsymbol{X})} \quad (2\text{-}5)$$

Dimana:

$\boldsymbol{X}$: Tupel data (sejumlah nilai dari variabel prediktor pada suatu amatan)

H : Hipotesis bahwa $\boldsymbol{X}$ merupakan anggota kelas C

$P(H|\boldsymbol{X})$: Peluang *posterior* (Peluang H bersyarat **X**)

$P(H)$: Peluang *prior*

$P(\boldsymbol{X})$: Peluang tupel **X** yang dihitung dengan membagi jumlah tupel **X** dengan total dataset.

$P(\boldsymbol{X}|H)$: Peluang **X** bersyarat H)

c. Regresi Logistik

Regresi logistik adalah teknik klasifikasi umum yang digunakan dalam permasalahan klasifikasi biner. Metode ini digunakan untuk mengestimasi variabel dependen biner berdasarkan satu atau sekumpulan variabel penjelas. Regresi logistik melakukan pengklasifikasian data berdasarkan *logistic function* [27].

$$p(x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}} = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}} \quad (2\text{Error! No text of specified style in document.}-6)$$

Dimana:

$p(x)$: Peluang kejadian sukses

β_k : Nilai parameter ke-k

x_k : Nilai dari variabel prediktor ke-k

d. Regresi Logistik Firth

Regresi logistik Firth adalah teknik klasifikasi regresi logistik yang diaplikasikan pada data dengan kasus ketidakseimbangan kelas. metode ini digunakan untuk mengurangi bias estimasi *Maximum Likelihood* pada data yang tidak seimbang. Pengurangan bias ini dilakukan dengan menghilangkan bias orde pertama dengan modifikasi fungsi skor. Pemodifikasian dilakukan dengan memasukkan *Jeffreys invariant prior* pada fungsi skor tersebut. Dengan demikian, metode ini dapat mengatasi permasalahan regresi logistik biner konvensional dalam menghadapi data yang tidak seimbang [28].

Relative Risk

Relative risk atau risiko relatif adalah perbandingan kemungkinan terjadinya suatu kejadian (*event*) dalam kelompok yang terpapar dengan kemungkinan kejadian (*event*) tersebut terjadi dalam kelompok yang tidak terpapar.

Tabel 1.Tabel status terpapar dan status kejadian untuk penghitungan risiko relatif

Status Terpapar	**Status Kejadian**	
	Ya	Tidak
Terpapar	a	b
Tidak Terpapar	c	d

Dari tabel 1, penghitungan risiko relatif (RR) adalah sebagai berikut:

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} \quad (\text{Error! No text of specified style in document.}-7)$$

METODE PENELITIAN

Data dan Sumber Data

Tabel 2. Daftar variabel yang digunakan dan keterangannya

Variabel	Kode	Keterangan
Status Gagal Ginjal Kronis (GGK)	0	Tidak mengalami GGK
	1	Mengalami GGK
Klasifikasi Wilayah	0	Perdesaan
	1	Perkotaan
Status Penyakit Jantung	0	Tidak mengalami penyakit jantung
	1	Mengalami penyakit jantung
Status Diabetes Melitus (DM)	0	Tidak mengalami DM
	1	Mengalami DM
Konsumsi Minuman Manis	0	< 1 kali per minggu
	1	≥ 1 kali per minggu
Konsumsi Minuman Berenergi	0	< 1 kali per minggu
	1	≥ 1 kali per minggu
Konsumsi Makanan Instan	0	< 1 kali per minggu
	1	≥ 1 kali per minggu
Usia	0	15-39 tahun
	1	40-64 tahun

Data yang digunakan merupakan data sekunder yang berasal dari *raw data* Riskesdas (Riset Kesehatan Dasar) tahun 2018. Variabel dependen yang dipakai adalah status gagal ginjal kronis yang merupakan variabel bertipe nominal dengan 2 kategori yaitu Ya dan Tidak. Jumlah sampel penduduk laki-laki usia produktif di Kalimantan Utara yang menjadi observasi pada penelitian ini adalah sebesar 2291 observasi, dimana 7 diantaranya mengalami gagal ginjal kronis dan sisanya tidak mengalami gagal ginjal kronis (Tabel 2).

Tahapan Penelitian

Tahapan analisis data yang dilakukan adalah sebagai berikut.

1. *Rawdata* dilakukan *preprocessing* terlebih dahulu untuk memilih variabel, melakukan *recode data* pada SPSS, memeriksa apabila adanya *missing data* dan duplikasi data, menghilangkan *missing data* dan duplikasi data, dan mengubah tipe data pada variabel yang diperlukan misal dari numerik menjadi faktor.
2. Analisis deskriptif dan penghitungan *relative risk* dilakukan setetelah data selesai dilakukan *preprocessing*. Ini dilakukan bertujuan mengetahui gambaran umum mengenai data.
3. *Split data* dilakukan untuk membagi data menjadi dua bagian yang selanjutnya disebut sebagai data latih dan data uji dengan perbandingan 70:30. Proses *split data* ini dilakukan dengan acak pada *software* R Studio.
4. Proses *resampling* diterapkan pada data latih yang telah terbentuk untuk menyeimbangkan kelas minoritas dan mayoritas. Teknik *resampling* yang diterapkan adalah SMOTE dan SMOTE + Tomek Links.
5. Pemodelan data dengan metode klasifikasi *Random Forest, Naive Bayes,* dan regresi logistik Firth dilakukan pada data latih original (tidak dilakukan *resampling*). Metode klasifikasi *Random Forest* dilakukan dengan mtry 2 dan ntree 500. Pemilihan mtry dan ntree ini dilakukan setelah beberapa nilai mtry dan ntree dicoba dan memberikan hasil terbaik.
6. Pada data latih dengan teknik *resampling* SMOTE dan SMOTE + Tomek Links, metode klasifikasi yang diterapkan adalah *Random Forest, Naive Bayes,* dan regresi logistik.
7. Setelah model di tiap klasifikasi dibentuk, akan dilakukan pengujian model dengan melakukan prediksi pada data yang baru yaitu data uji yang telah dibangun pada tahap *split data*.
8. Model yang telah diuji lalu dievaluasi dengan menghitung sensitivitas, spesifisitas, akurasi, dan nilai AUC. Akurasi adalah persentase data *testing* yang diklasifikasikan secara benar oleh model. Sensitivitas adalah proposi kelas positif yang diklasifikasikan dengan tepat, sementara

spesifisitas adalah proporsi kelas negatif yang diklasifikasikan dengan tepat. AUC adalah luas area di bawah kurva ROC.

9. Model dengan nilai sensitivitas, spesifisitas, akurasi, dan AUC tertinggi dipilih sebagai model yang terbaik.
10. Variabel penting yang berperan pada pengklasifikasian kelas pada model terbaik diidentifikasi dengan menggunakan *package* 'caret' pada R Studio.

HASIL PENELITIAN

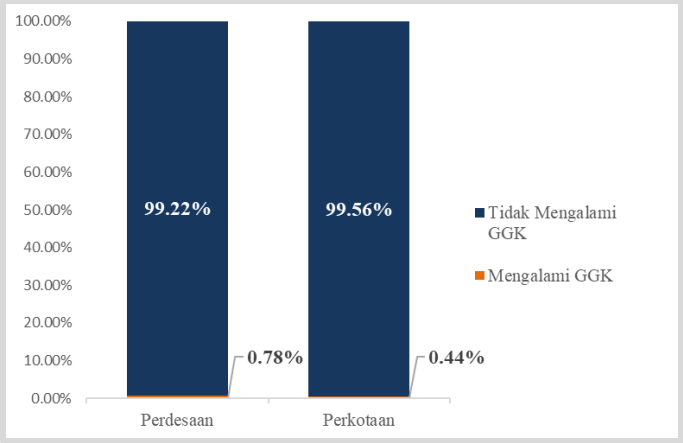
Gambaran Kejadian Gagal Ginjal Kronis pada Laki-Laki Usia Produktif di Kalimantan Utara Tahun 2018

Kejadian gagal ginjal kronis merupakan salah satu kejadian yang jarang terjadi sehingga terdapat ketidakseimbangan antara jumlah penduduk yang menderita gagal ginjal kronis dan yang tidak menderita penyakit tersebut. Meskipun demikian, kejadian gagal ginjal kronis di Kalimantan Utara memiliki prevalensi yang tertinggi di Indonesia.



Gambar 1 Kejadian Gagal Ginjal Kronis pada Laki-Laki Usia Produktif di Indonesia

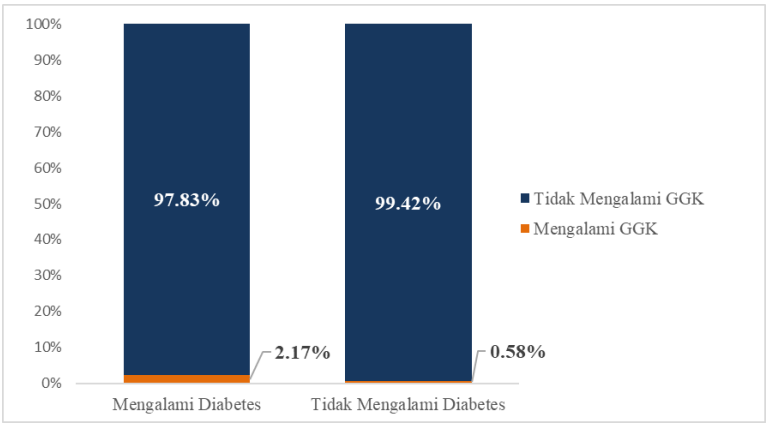
Berdasarkan hasil pengolahan data Riskesdas tahun 2018 didapatkan persentase penduduk laki-laki usia produktif yang mengalami gagal ginjal kronis adalah sebesar 0,61%. Angka tersebut menunjukkan bahwa sebanyak 6 sampai 7 dari 1000 penduduk laki-laki usia produktif di Kalimantan Utara mengidap penyakit gagal ginjal kronis. Sementara sisanya, yaitu 99,39% penduduk laki-laki usia produktif tidak mengalami gagal ginjal kronis (Gambar 1).



Gambar 2. Persentase Kejadian Gagal Ginjal Kronis Penduduk Laki-Laki Usia Produktif Berdasarkan Klasifikasi Wilayah

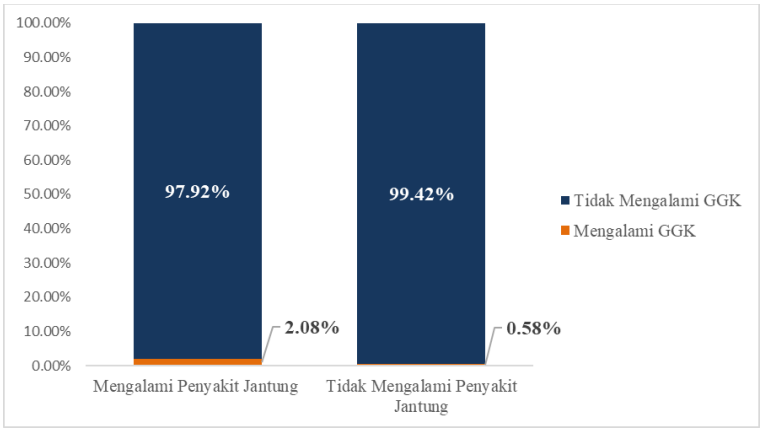
Persentase laki-laki usia produktif yang mengalami gagal ginjal kronis di perdesaan lebih besar dibandingkan perkotaan (Gambar 2). Persentase penduduk laki-laki usia produktif yang mengalami gagal ginjal kronis di perdesaan adalah sebesar 0,78%, ini berarti bahwa sebanyak 7 hingga 8 dari 1000 penduduk

laki-laki usia produktif di perdesaan mengalami gagal ginjal kronis. *Relative risk* yang didapatkan adalah sebesar 0,56. Ini menunjukkan bahwa penduduk laki-laki usia produktif yang tinggal di perkotaan memiliki peluang 0,56 kali mengalami gagal ginjal kronis dibandingkan yang tinggal di perdesaan.



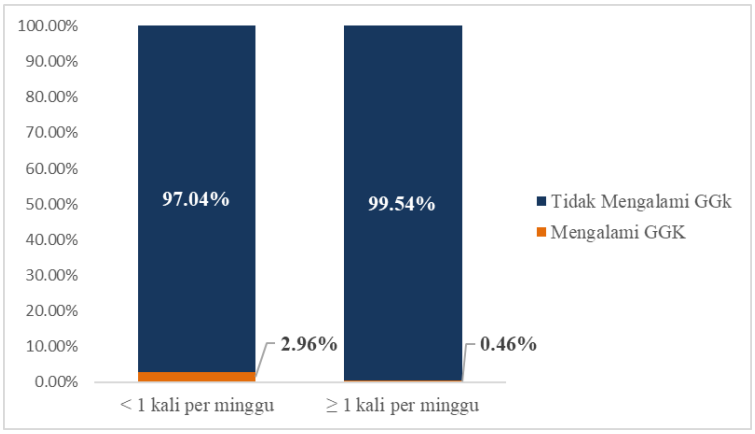
Gambar 3. Persentase Kejadian Gagal Ginjal Kronis Penduduk Laki-Laki Usia Produktif Berdasarkan Kejadian Diabetes Melitus

Persentase penduduk laki-laki usia produktif yang mengalami gagal ginjal kronis lebih besar pada penduduk laki-laki usia produktif dengan diabetes melitus dibandingkan dengan yang tidak mengalami diabetes melitus (Gambar 3). *Relative risk* yang didapatkan adalah sebesar 3,75. Ini menunjukkan bahwa penduduk laki-laki usia produktif yang mengalami diabetes melitus memiliki peluang 3,75 kali lebih besar untuk gagal ginjal kronis dibandingkan yang tidak mengalami diabetes melitus.



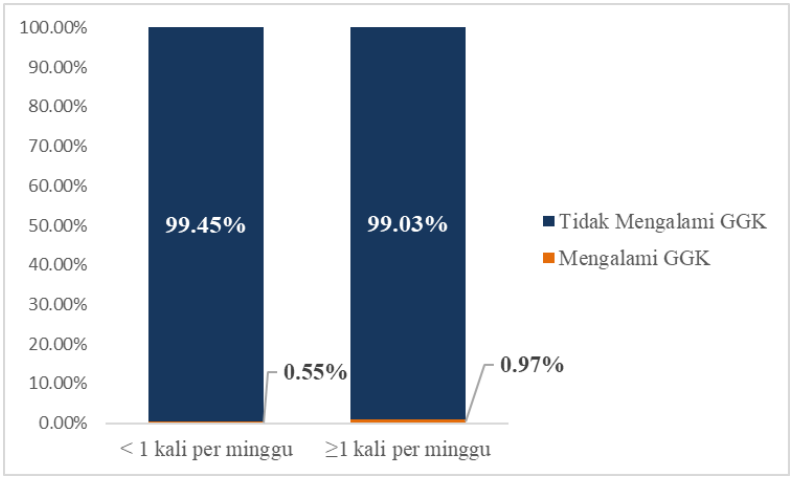
Gambar 4. Persentase Kejadian Gagal Ginjal Kronis Penduduk Laki-Laki Usia Produktif Berdasarkan Kejadian penyakit Jantung

Penduduk laki-laki usia produktif di Kalimantan Utara yang mengalami gagal ginjal kronis memiliki proporsi yang lebih besar pada penduduk dengan penyakit jantung dibandingkan dengan yang tidak mengalami penyakit jantung (Gambar 4), *Relative risk* yang didapatkan adalah sebesar 3,59. Ini menunjukkan bahwa penduduk laki-laki usia produktif yang mengalami penyakit jantung memiliki peluang 3,59 kali lebih besar untuk mengalami gagal ginjal kronis dibandingkan yang tidak mengalami penyakit jantung.



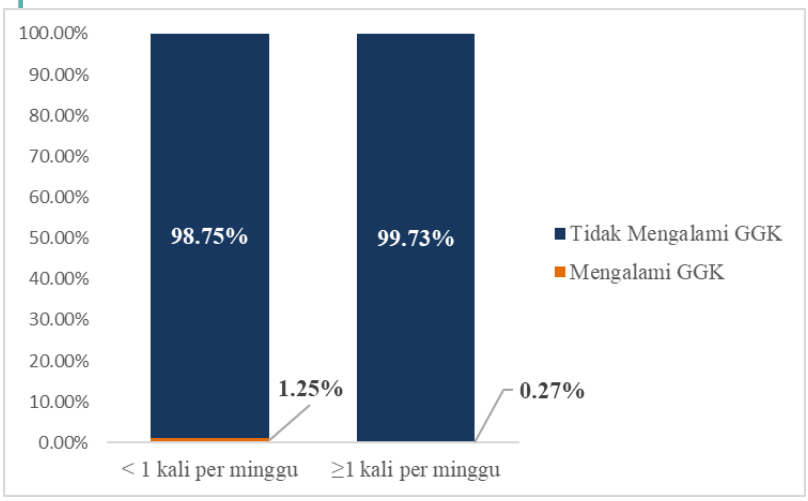
Gambar 5. Persentase Kejadian Gagal Ginjal Kronis Penduduk Laki-Laki Usia Produktif Berdasarkan Frekuensi Konsumsi Minuman Manis

Dapat dilihat pada gambar di atas, persentase penduduk laki-laki usia produktif yang mengalami kejadian gagal ginjal kronis lebih besar pada kategori konsumsi minuman manis kurang dari satu kali per minggu (Gambar 5). *Relative risk* yang didapatkan adalah sebesar 0,16. Ini menunjukkan bahwa penduduk laki-laki usia produktif yang mengonsumsi minuman manis lebih dari sama dengan satu kali per minggu memiliki peluang 0,16 kali untuk mengalami gagal ginjal kronis dibandingkan yang mengonsumsi minuman manis kurang dari satu kali per minggu.



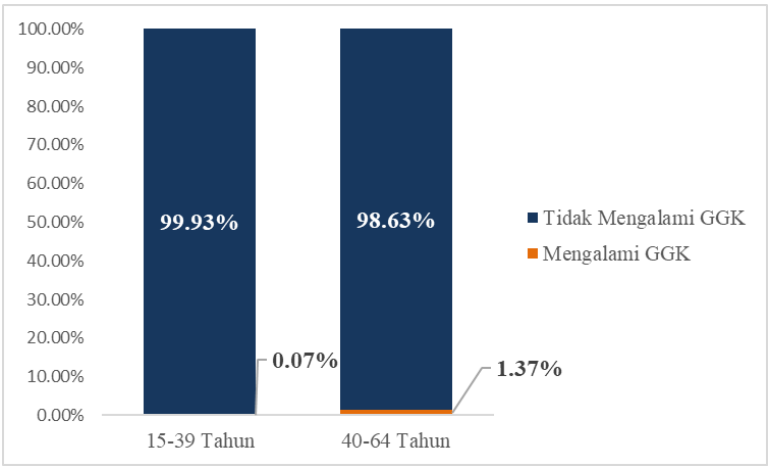
Gambar 6. Persentase Kejadian Gagal Ginjal Kronis Penduduk Laki-Laki Usia Produktif Berdasarkan Frekuensi Konsumsi Minuman Berenergi

Pada gambar di atas dapat dilihat bahwa persentase penduduk laki-laki usia produktif yang mengalami gagal ginjal kronis lebih besar pada penduduk laki-laki usia produktif yang mengonsumsi mimuman berenergi lebih dari sama dengan satu kali per minggu (Gambar 6). *Relative risk* yang didapatkan adalah sebesar 1,76. Ini menunjukkan bahwa penduduk laki-laki usia produktif yang mengonsumsi minuman berenergi lebih dari sama dengan satu kali per minggu memiliki peluang 1,76 kali lebih tinggi untuk mengalami kejadian gagal ginjal kronis dibandingkan penduduk laki-laki usia produktif yang mengonsumsi minuman berenergi kurang dari satu kali per minggu.



Gambar 7. Persentase Kejadian Gagal Ginjal Kronis Penduduk Laki-Laki Usia Produktif Berdasarkan Frekuensi Konsumsi Makanan Instan

Persentase kejadian gagal ginjal kronis lebih tinggi pada penduduk laki-laki dengan frekuensi konsumsi makanan instan kurang dari satu kali per minggu yaitu sebesar 1,2% (Gambar 7). *Relative risk* yang diperoleh adalah sebesar 0,22. Ini menunjukkan bahwa penduduk laki-laki usia produktif yang mengonsumsi makanan instan lebih dari sama dengan satu kali per minggu memiliki peluang 0,22 kali untuk terkena gagal ginjal kronis dibandingkan penduduk yang mengonsumsi makanan tersebut kurang dari satu kali per minggu.



Gambar 8. Persentase Kejadian Gagal Ginjal Kronis Penduduk Laki-Laki Usia Produktif Berdasarkan Kategori Usia

Penduduk laki-laki usia produktif di Kalimantan Utara yang berusia 40-64 tahun memiliki persentase kejadian gagal ginjal kronis yang lebih besar dibandingkan penduduk usia produktif dengan usia 15-39 tahun (Gambar 8). *Relative risk* yang didapatkan adalah sebesar 18,48. Ini berarti laki-laki usia produktif yang berusia 40-64 tahun memiliki peluang 18,48 kali lebih tinggi untuk terkena gagal ginjal kronis dibandingkan dengan laki-laki usia produktif yang berusia 15-39 tahun.

Analisis Perbandingan Penerapan Teknik *Resampling* dan Metode Klasifikasi

Teknik *resampling* diterapkan pada data latih sebelum pemodelan dengan metode klasifikasi dilakukan (Tabel 3).

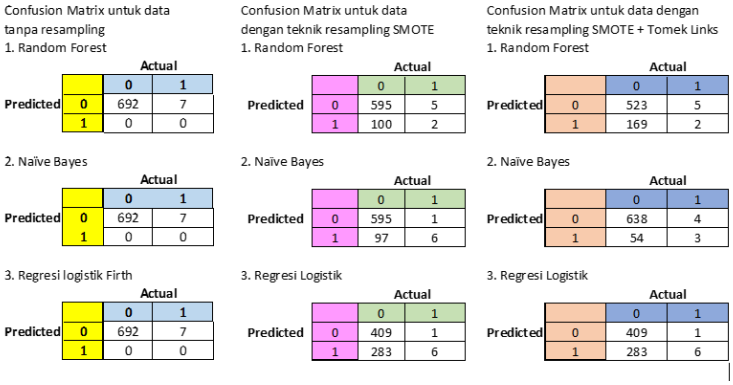
Tabel 3. Jumlah dan Proporsi Data Latih Sebelum dan Setelah Penerapan Teknik *Resampling*

Teknik Resampling	Mengalami GGK	Tidak Mengalami GGK
Tanpa Resampling	7 0.44%	1585 99,56%

Teknik Resampling	Mengalami GGK	Tidak Mengalami GGK
SMOTE	77	105
	42,31%	57,69%
SMOTE + Tomek Links	77	99
	43,75%	56,35%

Proporsi data latih sebelum dan sesudah penerapan teknik *resampling* menunjukkan bahwa SMOTE dan SMOTE + Tomek Links berhasil mengatasi kasus ketidakseimbangan ekstrim pada data status gagal ginjal kronis. Setelah tahap ini berhasil dilakukan, akan dilanjutkan proses pemodelan data dengan metode klasifikasi.

Analisis perbandingan teknik *resampling* dan metode klasifikasi dilakukan dengan membandingkan seluruh ukuran evaluasi yang dihasilkan tiap kombinasi teknik *resampling* dan metode klasifikasi yang digunakan. Ukuran evaluasi tersebut dihitung dari *confusion matrix* yang ditunjukan dibawah ini.



Gambar 9. *Confusion matrix* dari tiap metode dan teknik *resampling* yang digunakan

Setiap pengujian dengan data uji menghasilkan hasil yang berbeda-beda untuk setiap teknik *resampling* dan metode klasifikasi yang digunakan seperti yang terlihat pada gambar 9. Kemudian, seluruh ukuran evaluasi dihitung dari *confusion matrix* tersebut seperti yang tertera pada table di bawah ini.

Tabel 4. Perbandingan Teknik *Resampling* dan Metode Klasifikasi yang Digunakan

Teknik *Resampling*	Metode Klasifikasi	Akurasi	Sensitivitas	Spesifisitas	AUC
Tanpa *Resampling*	*Random Forest*	0,9899	0	1	0,50
	Naïve Bayes	0,9899	0	1	0,50
	Regresi Logistik Firth	0,9899	0	1	0,50
SMOTE	*Random Forest*	0,8497	0,2857	0,8555	0,5706
	Naïve Bayes	0,8598	0,8571	0,8598	0,8585
	Regresi Logistik	0,9185	0,4286	0,9234	0,6759
SMOTE + Tomek Links	*Random Forest*	0,7396	0,4285	0,7428	0,5857
	Naïve Bayes	0,5879	0,8571	0,5853	0,7212
	Regresi Logistik	0,9185	0,4286	0,9234	0,6759

Tabel di atas menunjukkan ringkasan ukuran nilai evaluasi model untuk penerapan metode klasifikasi tanpa *resampling*, dengan SMOTE, dan dengan SMOTE + Tomek Links. Dapat dilihat pada tabel 4, model tanpa *resampling* menghasilkan akurasi yang begitu tinggi yaitu 0,9899, nilai ini mendekati sempurna. Namun sensitivitasnya bernilai 0 dan spesifisitasnya bernilai 1. Ini menandakan bahwa model tanpa *resampling* tidak bisa mengidentifikasi kelas minoritas dan gagal mengklasifikasikan kelas positif atau kejadian laki-laki usia produktif yang mengalami gagal ginjal kronis dengan tepat. Nilai AUC untuk ketiga model tanpa *resampling* bernilai 0,5 yang berarti bahwa model tersebut tidak menunjukkan kinerja yang lebih baik dari pengklasifikasian acak.

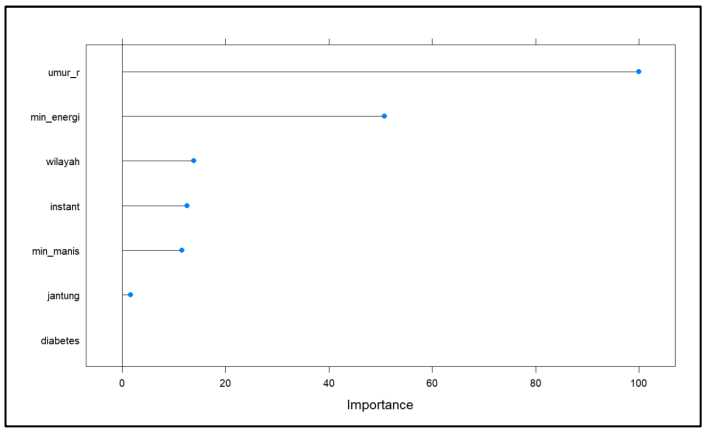
Model dengan SMOTE menunjukkan adanya penurunan pada nilai akurasi, namun terjadi peningkatan pada nilai sensitivitas dan AUC. Penurunan pada nilai spesifisitas dan peningkatan pada nilai

sensitivitas menunjukkan bahwa model telah mampu mengidentifikasi kelas minoritas dan tidak mengabaikan kelas tersebut lagi. Namun, selisih yang besar antara sensitivitas dan spesifisitas pada model *Random Forest* dan regresi logistik mengindikasikan bahwa meskipun model telah bisa mengidentifikasi kelas minoritas, kemampuan itu masih jauh lebih buruk dibandingkan dengan pengidentifikasian kelas mayoritas sehingga spesifisitas yang dihasilkan tinggi. Model *Naïve Bayes* pada penerapan teknik *resampling* dengan SMOTE menunjukkan nilai akurasi 0,8598, spesifisitas 0,8598, sensitivitas 0,8571, dan AUC yang bernilai 0,8585. Nilai sensitivitas dan spesifisitas yang cenderung seimbang menunjukkan model sama baiknya dalam mengidentifikasi dan mengklasifikasikan kelas minoritas dan mayoritas dengan tepat. Nilai AUC yang bernilai 0,8585 dan merupakan nilai AUC terbesar dari seluruh model menunjukkan bahwa model *Naïve Bayes* merupakan model klasifikasi yang baik. Selanjutnya, penerapan model klasifikasi dengan teknik *resampling* SMOTE + Tomek Links tidak menunjukkan hasil yang lebih baik dibandingkan dengan model dengan penerapan teknik *resampling* SMOTE.

Identifikasi Variabel Penting yang Berperan dalam Pengklasifikasian Status Kejadian Gagal Ginjal Kronis Laki-Laki Usia Produktif

Model dengan klasifikasi terbaik yang didapatkan adalah model *Naïve Bayes* dengan teknik *resampling* SMOTE. Penentuan variabel penting pada metode *Naïve Bayes* didasarkan pada nilai AUC tiap prediktor pada kurva ROC. Model akan dibentuk untuk tiap prediktor lalu nilai AUC dari model dengan prediktor tersebut akan dihitung. Prediktor yang memiliki nilai AUC tertinggi dianggap sebagai prediktor yang berperan penting pada pengklasifikasian kejadian gagal ginjal kronis. Pengidentifikasian variabel penting ini dilakukan dengan menggunakan R Studio dengan *package* 'caret'. Dengan menggunakan *package* tersebut, model akan dibangun terlebih dahulu lalu diidentifikasi variabel pentingnya.

Grafik pada gambar 9 dibentuk dengan mengskalakan nilai AUC yang telah dihasilkan untuk tiap prediktor sehingga nilainya berada antara 0 dan 100. Variabel yang memiliki nilai kepentingan dengan skala 100 merupakan variabel dengan tingkat kepentingan tertinggi dibandingkan variabel prediktor lainnya. Plot di atas menunjukkan bahwa 5 variabel prediktor yang memiliki peran penting terhadap pengklasifikasian kejadian gagal ginjal kronis adalah usia, konsumsi minuman energi, klasifikasi wilayah, konsumsi makanan instan, dan konsumsi minuman manis. Variabel prediktor dengan tingkat kepentingan tertinggi adalah variabel usia.



Gambar 10. Plot Variabel Penting pada Pengklasifikasian Kejadian Gagal Ginjal Kronis dengan Model Naïve Bayes dan Teknik Resampling SMOTE

PEMBAHASAN

Prevalensi kejadian gagal ginjal kronis yang tinggi di Kalimantan Utara menjadi isu krusial, terutama karena penduduk usia produktif tergolong rentan dan tingkat keparahan penyakit ini cenderung berkembang lebih cepat pada laki-laki dalam kelompok usia tersebut. Padahal, laki-laki usia produktif memiliki peran penting secara ekonomi. Di sisi lain, analisis terhadap permasalahan ini menghadapi tantangan berupa ketidakseimbangan kelas yang ekstrim, dimana jumlah kejadian gagal ginjal kronis jauh lebih sedikit dibandingkan dengan yang tidak mengalami gagal ginjal kronis. Oleh karena itu, pendekatan

yang dapat dilakukan adalah menerapkan teknik *resmapling* terlebih dahulu sebelum menggunakan metode klasiffikasi untuk analisis data.

Hasil evaluasi model menunjukkan bahwa kinerja model mengalami peningkatan dengan menerapkan teknik *resampling* sebelum penerapan metode klasifikasi dilakukan. Teknik *resampling* yang digunakan juga dapat mengatasi masalah ketidakseimbangan kelas data dengan baik. Model dengan SMOTE dan SMOTE + Tomek Links menunjukkan kinerja yang lebih baik dibandingkan dengan model yang tidak menerapkan teknik *resampling*. Dari seluruh metode klasifikasi dan teknik *resampling* yang digunakan, model terbaik yang didapatkan adalah model *Naïve Bayes* dengan SMOTE. Model ini telah berhasil mengidentifikasi kelas minoritas dan kelas mayoritas dengan baik dan mengklasifikasikan kelas-kelas tersebut dengan tepat. Model ini juga memiliki nilai AUC tinggi yang menempatkan model ini pada kategori model klasifikasi dengan kemampuan klasifikasi baik.

Temuan ini memperkuat bukti sebelumnya bahwa faktor-faktor seperti usia, konsumsi minuman berenergi, wilayah tempat tinggal, serta pola konsumsi makanan instan dan minuman manis memiliki peran penting dalam pengklasifikasian kejadian gagal ginjal kronis. Meski sebelumnya faktor-faktor ini telah diidentifikasi secara umum, hasil ini memberikan konteks yang lebih spesifik yaitu pada laki-laki usia produktif di Kalimantan Utara, kelompok yang terbukti lebih rentan terhadap perkembangan keparahan penyakit gagal ginjal kronis. Hasil ini dapat dimanfaatkan oleh tenaga kesehatan dalam mengidentifikasi individu yang berisiko tinggi mengalami gagal ginjal kronis, khususnya pada kelompok laki-laki usia produktif. Informasi mengenai faktor-faktor penting seperti usia, konsumsi minuman berenergi, dan wilayah tempat tinggal dapat digunakan sebagai dasar skrining dini atau penyuluhan yang lebih tepat sasaran. Selain itu, model klasifikasi yang telah terbukti efektif dapat diadaptasi sebagai alat bantu pengambilan keputusan dalam pelayanan kesehatan berbasis data. Selain itu, hasil analisis gagal ginjal kronis ini juga dapat memperkaya pengetahuan mengenai penerapan ilmu statistik dan sains data dalam menangani permasalahan data yang tidak seimbang terutama terkait pentingnya pemilihan metode yang sesuai dengan karakteristik data. Temuan ini juga dapat dijadikan dasar dalam membentuk kebijakan intervensi kejadian gagal ginjal kronis seperti kebijakan penayangan iklan makanan berisiko yang seharusnya dibatasi atau diberikan kriteria tertentu. Kebijakan program promosi kesehatan dan layanan deteksi dini juga dapat dicanangkan sehingga penduduk yang sudah terkena gagal 'ginjal kronis bisa memperlambat perkembangan keparahan penyakitnya.

Meskipun penelitian ini telah memberikan implikasi yang baik untuk pendidikan, kesehatan, dan kebijakan secara umum, penelitian ini tetap memiliki keterbatasan di antaranya adalah cakupan variabel prediktor yang terbatas dan datanya bersifat *cross-sectional* sehingga pengaruh variabel terhadap perkembangan keparahan penyakit ini tidak dapat diobservasi. Oleh karena itu, celah ini dapat dimanfaatkan untuk penelitian selanjutnya agar dapat memanfaatkan data longitudinal dan menganalisis suatu kelompok yang lebih spesifik agar pengaruh dari variabel bisa dianalisis lebih jelas. Selain itu, penelitian selanjutnya juga dapat menggunakan lebih banyak variabel prediktor serta mengekplorasi algoritma *machine learning* lain untuk melihat potensi peningkatan akurasi prediksi terutama dalam skenario data yang tidak seimbang.

KESIMPULAN

Persentase kejadian gagal ginjal kronis penduduk laki-laki usia produktif di Kalimantan Utara pada tahun 2018 adalah 0,61%. Kejadian gagal ginjal kronis ini memiliki persentase yang lebih besar pada penduduk laki-laki usia produktif dengan usia berada pada rentang 40-64 tahun, tinggal di perdesaan. mengalami penyakit jantung, mengidap penyakit diabetes melitus, mengonsumsi minuman berenergi lebih dari sama dengan satu kali per minggu, mengonsumsi minuman manis kurang dari satu kali per minggu, dan mengonsumsi makanan instan kurang dari satu kali per minggu.

Penerapan teknik *resampling* SMOTE dan SMOTE + Tomek Links terbukti dapat menyeimbangkan kelas data yang awalnya tidak seimbang dan meningkatkan kinerja metode klasifikasi dibandingkan dengan penerapan metode tanpa *resampling*. Teknik *resampling* terbaik yang dapat meningkatkan kinerja model pada kejadian gagal ginjal kronis penduduk laki-laki usia produktif di Kalimantan Utara pada tahun 2018 adalah SMOTE.

Metode klasifikasi terbaik dari hasil evaluasi pada pemodelan kejadian gagal ginjal kronis penduduk laki-laki usia produktif di Kalimantan Utara pada tahun 2018 adalah metode klasifikasi *Naïve Bayes*. Pemilihan variabel penting yang dilakukan pada model terbaik *Naïve Bayes* memberikan hasil bahwa variabel usia, konsumsi minuman berenergi, klasifikasi wilayah, konsumsi makanan instan, dan konsumsi minuman manis merupakan variabel penting yang berperan pada pengklasifikasian kejadian gagal ginjal kronis penduduk laki-laki usia produktif di Kalimantan Utara pada tahun 2018.

REFERENCES

[1] World Health Organization (WHO), "Noncommunicable diseases." Accessed: Aug. 29, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases#:~:text=Key%20facts%201%20Noncommunicable%20diseases%20%28NCDs%29%20kill%2041,are%20in%20low-%20and%20middle-income%20countries.%20More%20items

[2] KDOQI, *Clinical practice guidelines for chronic kidney disease: evaluation, classification and stratification*. 2002.

[3] S. Wilson, P. Mone, S. S. Jankauskas, J. Gambardella, and G. Santulli, "Chronic kidney disease: Definition, updated epidemiology, staging, and mechanisms of increased cardiovascular risk," *Journal of Clinical Hypertension*, vol. 23, no. 4. Blackwell Publishing Inc., pp. 831–834, Apr. 01, 2021. doi: 10.1111/jch.14186.

[4] B. Bikbov *et al.*, "Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017," *The Lancet*, vol. 395, no. 10225, pp. 709–733, Feb. 2020, doi: 10.1016/S0140-6736(20)30045-3.

[5] World Health Organization (WHO), "The Top Ten Causes of Death." Accessed: Feb. 22, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[6] Kemenkes RI, *Laporan Riskesdas 2018 Nasional*. Jakarta: Lembaga Penerbit Badan Penelitian dan Pengembangan Kesehatan, 2019.

[7] Menteri Kesehatan RI, "Keputusan Menteri Kesehatan Republik Indonesia; Pedoman Nasional Pelayanan Kedokteran Tata Laksana Ginjal Kronik," 2023.

[8] O. Swartling, H. Rydell, M. Stendahl, M. Segelmark, Y. Trolle Lagerros, and M. Evans, "CKD Progression and Mortality Among Men and Women: A Nationwide Study in Sweden," *American Journal of Kidney Diseases*, vol. 78, no. 2, pp. 190-199.e1, Aug. 2021, doi: 10.1053/j.ajkd.2020.11.026.

[9] BPS Provinsi Kalimantan Utara, "Keadaan Angkatan Kerja Kalimantan Utara Tahun 2022," 2023.

[10] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans Knowl Data Eng*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[11] Q. A'yuniyah *et al.*, "Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 4, no. 1, p. 72, Sep. 2022, doi: 10.30865/json.v4i1.4781.

[12] M. Gokiladevi, S. Santhoshkumar, and V. Varadarajan, "MACHINE LEARNING ALGORITHM SELECTION FOR CHRONIC KIDNEY DISEASE DIAGNOSIS AND CLASSIFICATION," *Malaysian Journal of Computer Science*, vol. 2022, no. Special Issue 1, pp. 102–115, 2022, doi: 10.22452/mjcs.sp2022no1.8.

[13] D. Kurniadi, F. Nuraeni, and M. Firmansyah, "Klasifikasi Masyarakat Penerima Bantuang Langsung Tunai Dana Desa Menggunakan Naive Bayes dan SMOTE," *Jurnal eknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. Vol.10 No. 2, pp. 309–320, 2023.

[14] A. Taufiq Akbar, H. Prapcoyo, and R. Husaini, "SMOTE and K-Means Preprocessing for Classification by Logistic Regression on Pima Indian Diabetes Dataset Prapemrosesan Menggunakan SMOTE dan K-means untuk Klasifikasi Regresi Logistik pada Data Pima Indian Diabetes," *Jurnal Informatika dan Teknologi Informasi*, vol. 20, no. 2, pp. 238–249, 2023, doi: 10.31515/telematika.v20i2.9676.

[15] D. A. Yani, P. Sarnianto, and Y. Anggriani, "FAKTOR-FAKTOR RISIKO PASIEN HEMODIALISIS DI RSUD ARJAWINANGUN DAN RSUD WALED KABUPATEN CIREBON," *Jurnal Ilmiah Indonesia*, vol. 5, no. 1, 2020.

[16] E. Sulistiowati and S. Idaiani, "Faktor Risiko Penyakit Ginjal Kronik Berdasarkan Analisis Cross-sectional Data Awal Studi Kohort Penyakit Tidak Menular Penduduk Usia 25-65 Tahun di Kelurahan Kebon

Kalapa, Kota Bogor Tahun 2011," *Buletin Penelitian Kesehatan*, vol. Vol. 43, No. 3, pp. 163–172, 2015.
[17] KDIGO, "KDIGO 2023 Clinical Practice Guideline for The Evaluation and Management of Chronic Kidney Disease Public Review Draft," 2023.
[18] AIHW, *Chronic kidney disease in Australia*. Canberra, 2005. [Online]. Available: www.aihw.gov.au
[19] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*. New Jersey: John Wiley & Sons, Inc., 2020.
[20] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*. Canada: john Wiley & Sons, Inc., 2013.
[21] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," 2002.
[22] G. E. A. P. A. Batista, A. L. C. Bazzan, and M. C. Monard, "Balancing Training Data for Automated Annotation of Keywords: a Case Study," *Proceedings of the Second Brazilian Workshop on Bioinformatics*, 2003, [Online]. Available: http://www.cs.waikato.ac.nz/
[23] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.
[24] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, pp. 5–32, 2001.
[25] S. Chakraborty, H. S. Islam, and D. Samanta, *Data Classification and Incremental Clustering in Data Mining and Machine Learning*. Switzerland: Springer, 2022. [Online]. Available: https://link.springer.com/bookseries/15427
[26] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd Edition. Waltham: Elsevier Inc., 2012.
[27] P. Wlodarczak, *Machine Learning and its Applications*, 1st Edition. Australia: Taylor & Francis Group, 2020. doi: 10.4324/9780429448782.
[28] D. Firth, "Bias Reduction of Maximum Likelihood Estimates," 1993.