Klasifikasi Indeks Pembangunan Gender Di Indonesia Tahun 2020 Menggunakan Supervised Machine Learning Algorithms

Artanti Indrasetianingsih[1*], Fenny Fitriani[2],Prasdianitaningtiyas Junita Kusuma[3]
[1,2,3]Program Studi Statistika, Fakultas Sains dan Teknologi, Universitas PGRI Adi Buana Surabaya
*Corresponding author: artanti.indra@unipasby.ac.id

Received: 27 September 2021
Accepted: 29 September 2021

ABSTRACT – **Indeks Pembangunan Gender (IPG) merupakan indikator yang digunakan untuk menggambarkan kesenjangan pencapaian pembangunan manusia antara laki-laki dan perempuan. Capaian IPG Indonesia pada tahun 2020 sebesar 91,06. IPG dapat diklasifikasikan menjadi 2 kategori, yaitu kategori rendah jika nilai IPG kurang dari 90 dan kategori tinggi jika nilai IPG lebih besar sama dengan 90. Berdasarkan sebaran kabupaten/kota, pada tahun 2020 terdapat 280 dari 514 kabupaten/kota yang mencapai angka IPG di atas 90. Hal ini menunjukkan bahwa capaian IPG di Indonesia belum merata. Tujuan dalam penelitian ini adalah untuk mengklasifikasikan dan membandingkan hasil ketepatan klasifikasi tentang IPG di Indonesia tahun 2020 dengan menggunakan algoritma supervised machine learning yaitu Regresi Logistik Biner dan K-Nearest Neighbor (K-NN). Hasil penelitian diperoleh bahwa variabel yang berpengaruh signifikan terhadap IPG yaitu Angka Partisipasi Sekolah SMA, persentase penduduk yang mempunyai keluhan kesehatan, persentase Pegawai Negeri Sipil perempuan, sumbangan pendapatan perempuan, dan rasio jenis kelamin. Hasil perbandingan kedua metode yang digunakan menunjukkan bahwa metode terbaik untuk mengklasifikasikan IPG kabupaten/kota di Indonesia tahun 2020 yaitu menggunakan K-NN, dengan nilai akurasi, sensitivitas, spesifisitas, dan AUC yang diperoleh masing-masing sebesar 71,88%, 65,52%, 77,14%, dan 71,33%. Nilai AUC sebesar 0,7133 atau 71,33% menunjukkan bahwa hasil klasifikasi termasuk dalam tingkat klasifikasi yang baik.**

Keywords– **IPG, Klasifikasi, Supervised Machine Learning Algorithms, Regresi Logistik Biner, K-NN.**

I. PENDAHULUAN

Salah satu faktor yang dapat dijadikan sebagai daya saing dari suatu negara yaitu kualitas sumber daya manusia, baik ditinjau dari segi kemampuan, keterampilan, ataupun produktivitasnya. Oleh karena itu, dalam mewujudkan bangsa yang mempunyai daya saing perlu dilakukan suatu upaya untuk membangun kualitas dari sumber daya manusia tersebut. Upaya pembangunan manusia diperuntukkan bagi seluruh penduduk di suatu negara, tanpa harus memandang adanya perbedaan jenis kelamin. Pada konteks pembangunan manusia terdapat istilah mengenai pembangunan berbasis gender yang mengukur capaian pembangunan antara perempuan dan laki-laki.

Pembangunan berbasis gender secara tegas tercantum dalam tujuan ke lima dari Tujuan Pembangunan Berkelanjutan (TPB) atau *Sustainable Development Goals* (SDGs) yaitu "Mencapai Kesetaraan Gender Dan Memberdayakan Kaum Perempuan" (Badan Perencanaan Pebangunan Nasional, 2020). Selain itu, isu kesetaraan gender secara tegas sebagai tujuan dari Rencana Pembangunan Jangka Panjang Nasional (RPJPN) tahun 2005-2025, yang dijabarkan dalam Rencana Pembangunan Jangka Menengah Nasional (RPJMN) tahun 2020-2024 serta menjadi rencana strategis dari Kementerian Pemberdayaan Perempuan dan Perlindungan Anak. Kesetaraan gender merupakan salah satu indikator yang tidak boleh diabaikan, hal ini inti dari pembangunan manusia itu sendiri adalah perempuan dan laki-laki [1].

Adanya tujuan kesetaraan gender sebagai salah satu tujuan dari SDGs menjadikan kesetaraan gender sebagai suatu urgensi dalam pembangunan manusia. Hal ini karena kemajuan dari suatu negara tidak dapat sepenuhnya tercapai tanpa adanya kesetaraan gender. Terdapat beberapa ukuran atau indikator yang dapat digunakan untuk mengevaluasi sejauh mana kesetaraan atau kesenjangan pembangunan manusia berbasis gender di suatu negara atau daerah. Salah satu indikator yang dapat digunakan yaitu Indeks Pembangunan Gender (IPG). Menurut Badan Pusat Statistik [2] Indeks Pembangunan Gender adalah perbandingan (rasio) capaian antara Indeks Pembangunan Manusia (IPM) laki-laki dan IPM perempuan.

Berdasarkan data Badan Pusat Statistik [2], capaian IPG Indonesia pada tahun 2020 sebesar 91,06. Capaian IPG ini mengalami penurunan sebesar 0,01 poin dari tahun 2019. Di Indonesia pada tahun 2020, terdapat 18 provinsi yang memiliki angka IPG dibawah rata-rata angka nasional yaitu 91,06. Menurut Kemen PPPA (2020), IPG dapat diklasifikasikan menjadi 2 kategori, yaitu kategori rendah jika nilai IPG kurang dari 90,00 dan kategori tinggi jika nilai IPG lebih besar sama dengan 90,00. Berdasarkan pengklasifikasian tersebut, terdapat 54,47% atau 280 dari 514 kabupaten/kota di Indonesia atau yang memiliki angka IPG di atas 90 dalam sebaran provinsi yang bervariasi. Hal ini menunjukkan bahwa capaian IPG di Indonesia baik di tingkat provinsi ataupun kabupaten/kota belum merata. Oleh karena itu, perlu dilakukan suatu penelitian untuk mengklasifikasikan atau mengelompokkan wilayah berdasarkan indikator IPG yang telah ditentukan.

Pengklasifikasian merupakan suatu metode yang digunakan untuk mengelompokkan atau mengklasifikasikan suatu observasi ke dalam kelas dengan memperhatikan atribut atau indikator yang ada. Pengklasifikasian dapat

menggunakan algoritma dari *machine learning* yang memiliki beberapa jenis yaitu *supervised learning, unsupervised learning, semi-supervised learning* dan *reinforcement learning*. Pada penelitian ini berfokus pada algoritma *supervised learning*. Menurut Yahya [3] *Supervised machine learning* merupakan sebuah pendekatan yang bertujuan untuk mengelompokkan suatu data ke data yang sudah ada dengan cara melatih data yang sudah ada tersebut dan terdapat variabel yang ditargetkan. Beberapa metode *supervised learning* yang sering digunakan untuk kasus klasifikasi antara lain yaitu Regresi Logistik, *K-Nearest Neighbor* (K-NN), *Support Vector Machine* (SVM), *Naïve Bayes*, dan *Decission Trees*. Metode klasifikasi yang menjadi fokus dalam penelitian ini yaitu Regresi Logistik Biner dan K-NN.

Regresi logistik biner dapat digunakan untuk menyelesaikan kasus klasifikasi. Pemilihan metode regresi logistik biner pada penelitian ini karena metode ini merupakan metode statistik klasik yang sering digunakan dalam berbagai penelitian karena mudah dalam pengaplikasiannya. Menurut Pratama, dkk. [4] *Support Vector Machine* (SVM) merupakan salah satu metode terbaik yang bisa digunakan untuk masalah klasifikasi. Pada beberapa kasus, data yang diperoleh tidak dapat diklasifikasi secara linier, sehingga untuk mengklasifikasikan data yang non linier SVM dapat dimodifikasi dengan cara menambahkan fungsi kernel di dalamnya. Sedangkan pemilihan metode K-NN dalam penelitian ini karena merupakan metode lama yang sederhana dan mudah untuk diimplementasikan. Menurut Hamamoto, dkk. [5] mengatakan bahwa metode KNN memiliki tingkat efisiensi yang tinggi, selain itu dalam beberapa kasus klasifikasi dapat memberikan tingkat akurasi yang tinggi.

Penelitian sebelumnya yang berkaitan dengan Indeks Pembangunan Gender pernah dilakukan oleh Fitarisca [6] dengan judul "Analisis Faktor-Faktor Yang Mempengaruhi Indeks Pembangunan Gender Dengan Menggunakan Regresi Probit". Hasil penelitian yang diperoleh menunjukkan bahwa faktor-faktor yang dapat mempengaruhi IPG pada penduduk laki-laki antara lain Angka Partisipasi Sekolah (APS) SD/sederajat dan rasio jenis kelamin saat lahir. Sedangkan faktor yang mempengaruhi IPG pada penduduk perempuan yaitu APS SMA/sederajat, Tingkat Partisipasi Angkatan Kerja (TPAK), Pengeluaran Perkapita yang disesuaikan (PPP), dan rasio jenis kelamin saat lahir.

Penelitian sebelumnya yang berkaitan dengan klasifikasi pernah dilakukan oleh Pusporani, dkk. [7] dengan judul "Klasifikasi Pasien Penderita Penyakit Liver dengan Pendekatan *Machine Learning*". Penelitian ini membandingkan antara metode regresi logistik, *Decission Trees, Naïve Bayes*, K-NN, dan SVM. Hasil penelitian menunjukkan bahwa metode SVM memberikan hasil yang terbaik berdasarkan nilai akurasi dan presisi, akan tetapi berdasarkan nilai *recall* maka metode K-NN memberikan hasil terbaik. Penelitian serupa pernah dilakukan oleh Utami, dkk. [8] dengan judul "Perbandingan Klasifikasi Status Pendonor Darah Dengan Menggunakan Regresi Logistik Dan *K-Nearest Neighbor*". Hasil penelitian menunjukkan bahwa metode yang memiliki ketepatan klasifikasi lebih tinggi yaitu regresi logistik biner dengan tingkat akurasi yang diperoleh sebesar 93%.

Berdasarkan uraian latar belakang yang telah dipaparkan di atas, maka penelitian ini bertujuan untuk mengeahui faktor-faktor yang berpengaruh terhadap nilai IPG di Indonesia tahun 2020 beserta mengklasifikasikan dan membandingkan hasil ketepatan klasifikasi Indeks Pembangunan Gender di Indonesia dengan menggunakan metode klasifikasi yaitu regresi logistik dan K-NN.

II. METODE PENELITIAN

A. Sumber Data

Sumber data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari publikasi Badan Pusat Statistik (BPS). Data yang digunakan merupakan data Indeks Pembangunan Gender di setiap kabupaten/kota di Indonesia tahun 2020 beserta faktor-faktor yang diduga mempengaruhinya. Unit observasi yang digunakan pada penelitian ini adalah 514 kabupaten dan kota yang ada di Indonesia dengan rincian 416 kabupaten dan 98 kota.

B. Variabel Penelitian

Variabel penelitian yang digunakan dalam penelitian ini terdiri dari variabel respon (Y) yaitu Indeks Pembangunan Gender (IPG). IPG diklasifikasikan menjadi dua kategori yaitu, IPG rendah (0) jika nilai IPG < 90 dan IPG tinggi (1) jika nilai $\geq$ 90. Variabel prediktor yang digunakan sebanyak 9 variabel, yaitu Angka Partisipasi Sekolah (APS) jenjang SD/Sederajat (X_1), Angka Partisipasi Sekolah jenjang SMP/Sederajat (X_2), Angka Partisipasi Sekolah jenjang SMA/Sederajat (X_3), persentase penduduk mempunyai keluhan kesehatan (X_4), persentase Pegawai Negeri Sipil (PNS) perempuan (X_5), Tingkat Partisipasi Angkatan Kerja (X_6), sumbangan pendapatan perempuan (X_7), rasio jenis kelamin (X_8), dan persentase penduduk miskin (X_9).

C. Langkah-langkah Penelitian

Langkah-langkah analisis yang digunakan dalam mencapai tujuan penelitian adalah sebagai berikut.

1. Melakukan eksplorasi data dengan menggunakan statistika deskriptif dari masing-masing variabel yang digunakan dalam penelitian.
2. Membagi data menjadi 2 bagian yaitu data *training* (75%) dan data *testing* (25%).
3. Melakukan klasifikasi menggunakan regresi logistik biner dengan langkah-langkah sebagai berikut.
   a. Melakukan pemeriksaan multikolinearitas antar peubah prediktor dengan menghitung nilai VIF.
   b. Membentuk model regresi logistik biner dengan menggunakan seluruh variabel prediktor.
   c. Melakukan uji signifikansi parameter model regresi logistik biner secara serentak maupun individu.

d. Melakukan uji kesesuaian model menggunakan uji *Hosmer and Lemeshow*.
e. Menghitung ketepatan klasifikasi model.

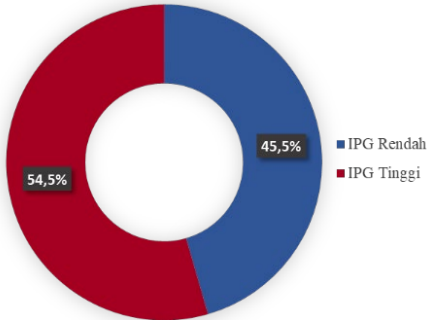
4. Melakukan klasifikasi menggunakan *K-Nearest Neighbor* dengan langkah-langkah sebagai berikut.
   a. Menentukan nilai parameter k.
   b. Menghitung kuadrat jarak *euclid* (*query distance*) masing-masing objek terhadap training data yang diberikan.
   c. Mencari jarak yang terkecil dengan mengurutkan hasil perhitungan berdasarkan *euclidean distance* secara *ascending* (berurutan dari jarak terkecil ke jarak terbesar terbesar).
   d. Pemeringkatan hasil pengurutan sesuai dengan nilai k.
   e. Memberi label dari data yang telah diperingkatkan tersebut berdasarkan kategori tetangga terdekat yang paling banyak.
   f. Menghitung nilai ketepatan klasifikasi.
5. Membandingkan tingkat ketepatan klasifikasi antara metode regresi logistik biner dan K-NN.
6. Menarik kesimpulan.

III. HASIL DAN PEMBAHASAN

Analisis dan pembahasan dalam penelitian ini meliputi statistika deskriptif terhadap variabel-variabel penelitian yang digunakan dan klasifikasi Indeks Pembangunan Gender kabupaten/kota di Indonesia menggunakan metode Regresi Logistik Biner dan *K-Nearest Neighbor*.

A. Statistika Deskriptif

Gambaran umum tentang variabel respon pada penelitian ini yaitu tentang Indeks Pembangunan Gender (IPG). Berikut statistik deskriptif variabel IPG yang disajikan pada Gambar 1.



Gambar 1. Statsistik Deskriptif Variabel IPG

Berdasarkan Gambar 1 dapat diketahui bahwa terdapat kabupaten/kota di Indonesia yang memiliki nilai IPG rendah sebanyak 234 kabupaten/kota atau sebesar 45,5%. Sedangkan jumlah kabupaten/kota yang memiliki IPG tinggi sebanyak 280 kabupaten/kota atau sebesar 54,5%. Berikut statistika deskriptif dari variabel prediktor yang dipakai pada penelitian ini.

Tabel 1. Statistik Deskriptif Variabel Prediktor

Variabel	Y	Min	Mean	Std.Dev	Maks
X_1	0	52,22	97,49	6,68	99,99
	1	76,66	99,32	1,51	99,99
X_2	0	32,43	93,47	8,59	99,71
	1	83,93	96,55	2,64	100
X_3	0	22,55	70,41	11,35	93,26
	1	55,93	77,03	8,23	98,89
X_4	0	0,23	26,90	8,63	54,92
	1	11,46	30,68	7,08	57,25
X_5	0	22,04	51,24	8,82	69,8
	1	28,52	57,90	7,26	72,54
X_6	0	51,83	69,66	6,66	96,25
	1	36,65	64,64	6,03	88,95
X_7	0	15,25	32,34	8,19	77,61
	1	20,11	34,19	7,00	52,39
X_8	0	93,3	105,42	5,60	139,02
	1	93,02	101,96	3,08	115
X_9	0	2,57	13,95	8,52	41,76
	1	2,02	9,82	5,65	34,49

Berdasarkan Tabel 1, dapat diketahui nilai minimum, mean (rata-rata), standar deviasi, dan maksimum dari masing-masing variabel prediktor. Selisih antara nilai minimum dan maksimum yang besar menunjukkan bahwa adanya perbedaan kondisi di masing-masing kabupaten/kota di Indonesia. Pada tahun 2020, rata-rata nilai APS

SD/Sederajat kabupaten/kota pada kategori IPG rendah adalah sebesar 97,49 yang berarti bahwa proporsi anak sekolah pada usia 7-12 tahun yang bersekolah pada jenjang SD terhadap jumlah penduduk usia 7-12 tahun adalah sebesar 97,49%. Sedangkan rata-rata APS SD/Sederajat kabupaten/kota pada kategori IPG tinggi adalah sebesar 99.32 yang berarti bahwa proporsi anak sekolah pada usia 7-12 tahun yang bersekolah pada jenjang SD terhadap jumlah penduduk usia 7-12 adalah sebesar 99,32%. Nilai APS SD/Sederajat di kabupaten/kota di Indonesia pada kategori IPG rendah masih memiliki perbedaan yang jauh, dimana APS SD/Sederajat terendah adalah kabupaten Puncak yaitu sebesar 52,22, sedangkan APS SD/Sederajat tertinggi adalah kabupaten Aceh Timur, Aceh Barat Daya, Seluma, Belitung, Banyumas, Rembang, dan Hulu Sungai Selatan yaitu sebesar 99,99. Nilai APS SD/Sederajat terendah pada kategori IPG tinggi yaitu sebesar 76.66 berada di kabupaten Lanny Jaya dan tertinggi sebesar 99,99 pada 22 kabupaten/kota di Indonesia, yaitu kabupaten Sleman, kabupaten Pekalongan, kota Magelang, dll. Standart deviasi dari kategori IPG tinggi (1,51) lebih kecil dari IPG rendah (6,68), artinya keragaman APS SD/Sederajat pada kategori IPG tinggi lebih kecil dari APS SD/Sederajat pada IPG rendah.

B. Klasifikasi Menggunakan Regresi Logistik

Pemeriksaan Multikolinieritas

Pemeriksaan multikolinieritas digunakan untuk melihat kebebasan antar variabel prediktor. Terdapat salah satu ukuran yang dapat digunakan untuk mendeteksi ada tidaknya kasus multikolinieritas yaitu nilai *Variance Inflation Factors* (VIF), dimana variabel dikatakan tidak ada multikolinearitas jika nilai VIF < 10. Berikut hasil pemeriksaan multikolinearitas.

Tabel 2. Pemeriksaan Multikolinearitas

Variabel	VIF
X_1	2,497
X_2	2,598
X_3	1,492
X_4	1,294
X_5	1,342
X_6	1,353
X_7	1,347
X_8	1,419
X_9	1,397

Berdasarkan Tabel 2, dapat diketahui bahwa antar variabel prediktor telah memenuhi kriteria yang telah ditentukan yaitu nilai VIF < 10. Hal ini menunjukkan bahwa antar variabel APS jenjang SD/Sederajat, APS jenjang SMP/Sederajat, APS jenjang SMA/Sederajat, persentase penduduk mempunyai keluhan kesehatan, persentase PNS perempuan, Tingkat Partisipasi Angkatan Kerja, sumbangan pendapatan perempuan, rasio jenis kelamin, dan persentase penduduk miskin tidak terjadi multikolinearitas.

Estimasi Parameter

Estimasi parameter pada model regresi logistik yaitu menggunakan pendekatan *Maximum Likelihood Estimator* (MLE). Tujuan dari pendekatan ini yaitu untuk memaksimumkan fungsi *likelihood*nya dalam menduga parameter β. Berikut ini adalah hasil estimasi parameter model pada data *training*.

Tabel 3. Koefisien Parameter Model

Parameter	Koefisien	Std.Error
(*Intercept*)	3,848	7,567
X_1	-0,025	0,070
X_2	-0,060	0,043
X_3	0,069	0,018
X_4	0,038	0,017
X_5	0,083	0,020
X_6	0,005	0,024
X_7	0,042	0,020
X_8	-0,072	0,034
X_9	-0,038	0,022

Berdasarkan Tabel 3, model regresi logistik biner yang terbentuk adalah sebagai berikut.

$$\pi(x) = \frac{\exp(g(x))}{1+\exp(g(x))}$$

Dimana

$$g(x) = 3{,}848 - 0{,}025x_1 - 0{,}060x_2 + 0{,}069x_3 + 0{,}038x_4 + 0{,}084x_5 + 0{,}005x_6 + 0{,}042x_7 - 0{,}072x_8 - 0{,}038x_9$$

Uji Signifikansi Parameter

Pengujian signifikansi parameter model dilakukan baik secara serentak maupun secara parsial untuk mengetahui faktor-faktor yang diduga berpengaruh terhadap Indeks Pembangunan Gender. Uji signifikansi parameter secara serentak digunakan untuk mengetahui pengaruh variabel prediktor terhadap variabel respon secara serentak. Pengujian signifikansi parameter secara serentak menggunakan statistik uji G (*Likelihood Ratio Test*) Berikut hasil uji signifikansi parameter secara serentak.

Tabel 4. Hasil Uji Serentak

	χ^2	df	$\chi^2_{(0.05,9)}$	*P-value*
Likelihood Ratio Test	110,5	9	16,92	$2,2 \times 10^{-16}$

Berdasarkan uji *Likelihood Ratio Test* pada Tabel 4, diketahui bahwa nilai chi-square hitung yaitu 110,5 lebih besar dari nilai chi-square tabel 16,92 atau p-value $2,2 \times 10^{-16}$ kurang dari alpha (0,05). Hal ini artinya variabel prediktor berpengaruh secara serentak terhadap variabel respon atau minimal terdapat satu variabel prediktor yang berpengaruh secara signifikan terhadap Indeks Pembangunan Gender.

Uji parsial dilakukan untuk mengetahui adanya pengaruh antara variabel prediktor terhadap variabel respon secara parsial atau individu. Uji parsial dapat dilakukan dengan menggunakan uji *Wald*. Berikut hasil uji signifikansi parameter secara parsial.

Tabel 5. Hasil Uji Secara Parsial

Parameter	W	*P-Value*
X_1	-0,357	0,717
X_2	-1,395	0,159
X_3	**3,833**	**0,0001**
X_4	**2,235**	**0,028**
X_5	**4,15**	**$2,57 \times 10^{-5}$**
X_6	0,208	0,844
X_7	**2,1**	**0,035**
X_8	**-2,118**	**0,034**
X_9	-1,723	0,0815

Berdasarkan uji *Wald* pada Tabel 5 dapat diketahui bahwa variabel yang berpengaruh signifikan terhadap Indeks Pembangunan Gender adalah variabel Angka Partisipasi Sekolah SMA (X_3), persentase penduduk mempunyai keluhan kesehatan (X_4), persentase Pegawai Negeri Sipil perempuan (X_5), sumbangan perempuan (X_7), dan rasio jenis kelamin (X_8). Hal ini karena nilai uji *Wald* pada masing-masing variabel tersebut lebih besar dari z_{hitung} (1,96) atau *p-value* masing-masing variabel lebih kecil dari alpha (0,05).

Uji Kesesuaian Model

Uji kesesuaian model digunakan untuk mengetahui apakah model yang telah terbentuk sudah sesuai atau tidak. Model dikatakan telah sesuai apabila tidak terdapat perbedaan antara hasil observasi dengan kemungkinan hasil prediksi model. Uji kesesuaian model menggunakan uji Hosmer dan Lemeshow. Berikut hasil uji kesesuaian model.

Tabel 6. Hasil Uji Kesesuaian Model

	$\hat{C}$	Df	$\chi^2_{(0.05,8)}$	*P-value*
Hosmer Lemeshow	8,8579	8	18,31	0,3544

Berdasarkan uji *Hosmer Lemeshow* pada Tabel 6 menunjukkan bahwa dengan menggunakan alpha 5% diperoleh keputusan gagal tolak H_0, karena nilai $\hat{C}$ (8,8579) lebih kecil dari chi-square tabel (18,31) atau p-value (0,3544) lebih besar dari alpha (0,05). Sehingga dapat disimpulkan bahwa model yang terbentuk telah sesuai, atau dapat dikatakan bahwa tidak terdapat perbedaan antara hasil observasi dengan kemungkinan hasil prediksi model.

Ketepatan Klasifikasi

Model regresi logistik biner yang telah didapatkan pada data *training* akan digunakan untuk klasifikasi pada data testing. Hal tersebut dilakukan dengan cara mensubtitusikan variabel prediktor pada data testing ke dalam model, sehingga dapat diperoleh nilai peluangnya. Apabila nilai peluang lebih dari 0,5, maka kabupaten/kota tersebut terklasifikasikan ke dalam kelas IPG tinggi. Namun, jika probabilitasnya kurang dari 0,5 maka akan terklasifikasikan ke dalam kelas IPG rendah. Berikut contoh perhitungan klasifikasi kabupaten atau kota untuk salah satu data testing yaitu kabupaten Simeulue.

$$g(x) = 3,848 - 0,025(99,3) - 0,060(98,53) + 0,069(88,64) + 0,038(17,62) + 0,084(47,44) + 0,005(70,37) + 0,042(24,3) - 0,072(105) - 0,038(18,49)$$

$$g(x) = -0,77033$$

$$\pi(x) = \frac{\exp(-0,77033)}{1+\exp(-0,77033)} = \frac{0,462858}{1+0,462858} = 0,316407$$

Diperoleh nilai probabilitas $\pi(x) < 0{,}5$, maka dapat diprediksi bahwa kabupaten Simeulue masuk ke dalam kelas 0 yaitu kabupaten/kota dengan status IPG rendah. Hasil klasifikasi regresi logistik biner menggunakan data *testing* dapat disajikan pada tabel confussion matrix berikut.

Tabel 7. Confusion Matrix Regresi Logistik Biner

Observasi	Prediksi		Total
	IPG Rendah	IPG Tinggi	
IPG Rendah	35	23	58
IPG Tinggi	16	54	70
Total	51	77	128

Berdasarkan tabel 7, menunjukkan bahwa metode regresi logistik biner akan mengklasifikasikan secara tepat atau benar 89 sampel dari total 128 sampel pada data *testing*. Dari tabel tersebut dapat dihitung nilai akurasi, sensitivitas, spesifisitas sebagai berikut.

$$Akurasi = \frac{35+54}{35+16+23+54} = 0{,}6953$$

$$Sensitivitas = \frac{54}{23+54} = 0{,}7013$$

$$Spesifisitas = \frac{35}{35+16} = 0{,}6863$$

$$\text{AUC} = \frac{1}{2}(0{,}7013 + 0{,}6863) = 0{,}6938$$

Berdasarkan perhitungan tersebut didapatkan nilai akurasi dari klasifikasi regresi logistik biner adalah sebesar 0,6953, yang artinya metode tersebut memiliki ketepatan klasifikasi dalam memprediksi data Indeks Pembangunan Gender (IPG) di Indonesia tahun 2020 sebesar 69,53%. Nilai sensitivitas yang diperoleh yaitu sebesar 0,7013, nilai tersebut menunjukkan bahwa kabupaten/kota dengan status IPG tinggi yang terklasifikasikan secara tepat sebagai kabupaten/kota dengan status IPG tinggi adalah sebesar 70,13% atau sebanyak 54 kabupaten/kota. Sedangkan nilai spesifisitas yang dihasilkan adalah sebesar 0,6863, yang artinya bahwa kabupaten/kota dengan status IPG rendah yang diklasifikasikan secara tepat sebagai kabupaten/kota dengan status IPG rendah adalah sebesar 68,63% atau sebanyak 35 kabupaten/kota. Nilai AUC yang diperoleh yaitu sebesar 0,6938 yang menunjukkan bahwa hasil klasifikasi metode regresi logistik biner masuk dalam kategori *Fair Classification*.

C. Klasifikasi Menggunakan K-NN

Metode K-NN merupakan metode klasifikasi dengan menggunakan pendekatan *supervised machine learning*, sehingga membutuhkan data *training* yang sudah dilabeli. K-NN dilakukan dengan cara mencari nilai dari k buah data training yang memiliki jarak terdekat dengan data testing yang labelnya belum diketahui. Berikut contoh perhitungan metode K-NN dengan menggunakan dataset IPG.

Tabel 8. Dataset IPG

No	X_1	X_2	X_3	...	X_9	Y
1	99,58	97,41	84,89	...	20,2	0
2	99,68	98,46	83,47	...	13,21	1
3	99,98	97,53	88,51	...	15,08	1
4	99,69	98,15	78,68	...	18,34	0
5	99,96	98,57	83,08	...	13,84	1
6	99,30	98,53	88,64	...	18,49	0

Tabel 8 menjelaskan bahwa dataset terdiri dari 6 data, dimana 5 data awal digunakan sebagai data *training* dan data ke-6 digunakan sebagai data *testing*. Setelah membagi data menjadi data *training* dan data *testing*, langkah selanjutnya yaitu menentukan nilai k tetangga terdekat. Pada contoh kali ini, nilai k yang digunakan yaitu k=3. Tahap berikutnya yaitu menghitung jarak antara data *training* dengan data *testing*. Salah satu jarak yang sering digunakan adalah jarak e*uclidean.* Nilai jarak *euclidean* yang diperoleh yaitu sebagai berikut.

Jarak untuk data *training* 1 dengan data *testing*:

$$dist_{(1,6)} = \sqrt{(99{,}58-99{,}30)^2 + (97{,}41-98{,}53)^2 + (84{,}89-88{,}64)^2 + \cdots + (20{,}2-18{,}49)^2}$$
$$= 14{,}40011$$

Jarak untuk data *training* 2 dengan data *testing*:

$$dist_{(2,6)} = \sqrt{(99{,}68-99{,}30)^2 + (98{,}46-98{,}53)^2 + (83{,}47-88{,}64)^2 + \cdots + (13{,}21-18{,}49)^2}$$
$$= 17{,}02364$$

Jarak untuk data *training* 3 dengan data *testing*:

$$dist_{(3,6)} = \sqrt{(99{,}98 - 99{,}30)^2 + (97{,}53 - 98{,}53)^2 + (88{,}51 - 88{,}64)^2 + \cdots + (15{,}08 - 18{,}49)^2}$$
$$= 26{,}12592$$

Jarak untuk data *training* 4 dengan data *testing*:

$$dist_{(4,6)} = \sqrt{(99{,}69 - 99{,}30)^2 + (98{,}15 - 98{,}53)^2 + (78{,}68 - 88{,}64)^2 + \cdots + (18{,}34 - 18{,}49)^2}$$
$$= 23{,}51955$$

Jarak untuk data *training* 5 dengan data *testing*:

$$dist_{(5,6)} = \sqrt{(99{,}96 - 99{,}30)^2 + (98{,}57 - 98{,}53)^2 + (83{,}08 - 88{,}64)^2 + \cdots + (13{,}84 - 18{,}49)^2}$$
$$= 25{,}43176$$

Setelah didapatkan jarak *euclidean* dari data *testing* ke setiap data *training* yang ada, langkah selanjutnya yaitu mengurutkan setiap nilai jarak euclidean tersebut dari yang terkecil atau terdekat. Hasil pengurutan jarak *euclidean* dari terkecil ke terbesar adalah sebagai berikut.

Tabel 9 Hasil Urutan Jarak

Data ke-	Kelas	Nilai Jarak	Urutan Jarak
1	0	14,40011	1
2	1	17,02364	2
3	1	26,12592	5
4	0	23,51955	3
5	1	25,43176	4

Berdasarkan Tabel 9, urutan jarak *euclidean* dari yang terkecil ke terbesar berturut-turut adalah data ke-1, 2, 4, 5, dan ke-3. Selanjutnya dilakukan pemeriksaan terhadap kelas sesuai dengan urutan jarak terdekat diperoleh urutan kelas yaitu 0, 1, 0, 1, dan 1. Kemudian menentukan label klasifikasi dari data *testing* tersebut berdasarkan kategori tetangga terdekat yang paling banyak. Pada contoh permasalahan ini digunakan nilai k = 3, maka akan didapatkan kelas hasil klasifikasi untuk data *testing* tersebut yaitu kelas 0.

Nilai k yang digunakan dalam penelitian ini yaitu menggunakan angka ganjil, hal ini dilakukan supaya tidak terdapat tetangga terdekat yang jumlahnya sama. Hasil akurasi dengan metode K-NN menggunakan nilai k ganjil yaitu 1, 3, 5, 7, 9, 11, 13, dan 15 adalah sebagai berikut.

Tabel 10 Ketepatan Klasifikasi Berdasarkan Nilai K Pada K-NN

Nilai K	Akurasi	Nilai K	Akurasi
K=1	0,6328	K=9	0,6953
K=3	0,6406	K=11	**0,7188**
K=5	0,6641	K=13	0,7031
K=7	0,6953	K=15	0,6797

Berdasarkan Tabel 10 diatas menunjukkan bahwa akurasi hasil prediksi K-NN paling tinggi yaitu menggunakan nilai k = 11, dengan tingkat akurasi mencapai 71,88%. Hasil klasifikasi K-NN dapat dibuat *confussion matrix* antara prediksi dengan aktual dari data *testing* yaitu sebagai berikut.

Tabel 10 Confussion Matrix K-NN dengan K = 11

Observasi	Prediksi		Total
	IPG Rendah	IPG Tinggi	
IPG Rendah	38	16	54
IPG Tinggi	20	54	74
Total	58	70	128

Berdasarkan Tabel 10, menunjukkan bahwa metode K-NN dapat melakukan klasifikasi secara tepat atau benar sebanyak 92 sampel dari total 128 sampel data *testing*. Dari tabel *confussion matrix* tersebut dapat dihitung nilai akurasi, sensitivitas, dan spesifisitas sebagai berikut.

$$Akurasi = \frac{38+54}{38+16+20+54} = 0{,}7188$$
$$Sensitivitas = \frac{38}{38+20} = 0{,}6552$$
$$Spesifisitas = \frac{54}{16+54} = 0{,}7714$$
$$\text{AUC} = \frac{1}{2}(0{,}6552 + 0{,}7714) = 0{,}7133$$

Berdasarkan perhitungan tersebut didapatkan nilai akurasi dari klasifikasi menggunakan K-NN adalah sebesar 0,7188, yang artinya metode tersebut memiliki ketepatan klasifikasi dalam memprediksi data Indeks Pembangunan Gender (IPG) di Indonesia tahun 2020 sebesar 71,88%. Nilai sensitivitas yang diperoleh yaitu sebesar 0,6552, nilai tersebut menunjukkan bahwa kabupaten/kota dengan status IPG tinggi yang terklasifikasikan secara tepat sebagai kabupaten/kota dengan status IPG tinggi adalah sebesar 65,52% atau sebanyak 54 kabupaten/kota. Sedangkan nilai spesifisitas yang dihasilkan adalah sebesar 0,7714, yang artinya bahwa kabupaten/kota dengan status IPG rendah yang diklasifikasikan secara tepat sebagai kabupaten/kota dengan status IPG rendah adalah sebesar 77,14% atau sebanyak 38 kabupaten/kota. Nilai AUC yang diperoleh yaitu sebesar 0,7133 yang menunjukkan bahwa hasil klasifikasi metode K-NN masuk dalam kategori *Fair Classification*.

D. Klasifikasi Menggunakan K-NN

Berdasarkan hasil yang diperoleh, maka perbandingan dari keseluruhan metode yang digunakan yaitu regresi logistik biner dan *K-Nearest Neighbor* yaitu sebagai berikut.

Tabel 11. Perbandingan Evaluasi Performa Klasifikasi

Metode	Akurasi	Sensitivitas	Spesifisitas	AUC
Regresi Logistik Biner	0,6953	0,7013	0,6863	0,6938
K-NN	0,7188	0,6522	0,7714	**0,7133**

Berdasarkan hasil perbandingan kedua metode yang digunakan menunjukkan bahwa nilai akurasi dan AUC metode *K-Nearest Neighbor* lebih besar dari metode regresi logistik biner. Sehingga metode terbaik untuk mengklasifikasikan Indeks Pembangunan Gender Kabupaten/Kota di Indonesia tahun 2020 yaitu menggunakan *K-Nearest Neighbor* dengan k = 11. Tingkat akurasi, sensitivitas, spesifisitas yang didapat yaitu masing-masing sebesar 71,88%, 65,22% dan 77,14%. Nilai AUC sebesar 0,7133 menunjukkan bahwa hasil klasifikasi termasuk dalam tingkat klasifikasi yang baik.

IV. KESIMPULAN

Berdasarkan analisis data dan pembahasan di atas dapat disimpulkan bahwa jumlah kabupaten/kota di Indonesia yang memiliki nilai IPG rendah sebanyak 234 kabupaten/kota atau sebesar 45,5%. Sedangkan jumlah kabupaten/kota yang memiliki IPG tinggi sebanyak 280 kabupaten/kota atau sebesar 54,5%. Hasil pemodelan menggunakan Regresi logistik biner diperoleh 5 variabel yang berpengaruh secara signifikan terhadap IPG, yaitu Angka Partisipasi Sekolah jenjang SMA, persentase penduduk mempunyai keluhan kesehatan, persentase Pegawai Negeri Sipil perempuan, sumbangan pendapatan perempuan, rasio jenis kelamin. Diperoleh nilai akurasi, sensitivitas, spesifisitas, dan AUC masing-masing sebesar 69,53%, 70,13%, 68,63%, dan 69,38%. Hasil klasifikasi dengan metode K-NN didapatkan hasil terbaik dengan k=11. Nilai akurasi, sensitivitas, spesifisitas, dan AUC yang diperoleh masing-masing sebesar 71,88%, 65,52%, 77,14%, dan 71,33%. Hasil perbandingan ketiga metode yang digunakan menunjukkan bahwa metode terbaik untuk mengklasifikasikan Indeks Pembangunan Gender kabupaten/kota di Indonesia tahun 2020 yaitu menggunakan *K-Nearest Neighbor* (K-NN) dengan hasil klasifikasi termasuk dalam tingkat klasifikasi yang baik.

REFERENSI

[1] Kementerian Pemberdayaan Perempuan dan Perlindungan Anak, "Pembangunan Manusia Berbasis Gender 2020," KPPPA, Jakarta, 2020.

[2] Badan Pusat Statistik, "Sistem Rujukan Statistik," BPS, Jakarta, 2021.

[3] S. Yahyah, Klasifikasi Ketepatan Lama Studi Mahasiswa Menggunakan Metode Support Vector Machine Dan Random Forest. [Skripsi], Yogyakarta: Jurusan Statistika FMIPA. Universitas Islam Indonesia., 2018.

[4] A. Pratama, R. Wihandika and D. Ratnawati, "Implementasi Algoritme Support Vector Machine (SVM) untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 2, no. 4, 2018.

[5] Y. Hamamoto, S. Uchimura and S. Tomita, "A Bootstrap Technique for Nearest Neighbours Classifier Design," *Ieee Transactions On Pattern Analysis And Machine Intelligence,* vol. 19, no. 1, pp. 73-79., 1997.

[6] A. Fitarisca, Analisis Faktor-faktor yang Mempengaruhi Indeks Pembangunan Gender (IPG) dengan menggunakan Regresi Probit. [Skripsi]., Surabaya: Jurusan Statistika FMIPA, Institut Teknologi Sepuluh Nopember Surabaya, 2014.

[7] E. Pusporani, S. Qomariyah and Irhamah., "Klasifikasi Pasien Penderita Penyakit Liver dengan Pendekatan Machine Learning," *Inferensi,* vol. 2, no. 1, 2019.

[8] I. Utami, Fadjryani and D. Daniati, "Perbandingan Klasifikasi Status Pendonor Darah Dengan Menggunakan Regresi Logistik Dan K-Nearest Neighbor," *Jurnal Statistika & Komputasi Statistik,* vol. 12, no. 1, 2020.

[9] Badan Perencanaan Pembangunan Nasional, "Metadata Indikator Pilar Pembangunan Sosial Pelaksanaan Pencapaian Tujuan Pembangunan Berkelanjutan/Sustainable Development Goals (TPB/SDGs)," BPPN, Jakarta, 2020.