JAMBURA JOURNAL OF MATHEMATICS
Jambura J. Math. Vol. 5, No. 2, pp. 404–418, August 2023
Journal Homepage: http://ejurnal.ung.ac.id/index.php/jjom
DOI: https://doi.org/10.34312/jjom.v5i2.21103

Perbandingan Metode KNN, Naive Bayes, dan Regresi Logistik Binomial dalam Pengklasifikasian Status Ekonomi Negara

N. K. Kutha Ardana[1], Ruhiyat[1,*], Nurfatimah Amany[1], Teofilus Kevin Irawan[1], Raymond[1], Rizalius Karunia[1], Syifa Fauzia[1]

[1]*Departemen Matematika, Fakultas MIPA, Institut Pertanian Bogor, Bogor 16680, Indonesia*
[*]*Corresponding author. Email: ruhiyat-mat@apps.ipb.ac.id*

ABSTRAK

Klasifikasi status ekonomi suatu negara sebagai maju atau berkembang sering kali melibatkan faktor angka harapan hidup dan peubah-peubah yang mendasarinya. Penelitian ini bertujuan untuk melakukan perbandingan performa tiga algoritma pembelajaran mesin, yaitu KNN (*K-Nearest Neighbors*), *naive Bayes*, dan regresi logistik binomial, dalam mengklasifikasi status ekonomi negara sebagai maju atau berkembang. Data yang digunakan pada penelitian ini adalah "Life Expectancy (WHO) Fixed" diperoleh dari situs Kaggle. Analisis statistika yang pertama dilakukan adalah melakukan analisis komponen utama (AKU) dengan 16 peubah prediktor. AKU menghasilkan tiga komponen utama yang mampu menjelaskan keragaman sebesar 71,41% yang selanjutnya digunakan pada metode KNN, *naive Bayes*, dan regresi logistik binomial. Hasil analisis dari metode KNN, *naive Bayes*, dan regresi logistik binomial masing-masing menghasilkan F1-*score* sebesar 100%, 98,19%, dan 97,36%.

Kata Kunci:
Angka Harapan Hidup; Klasifikasi Negara; KNN; *Naive Bayes*; Regresi Logistik Binomial

ABSTRACT

The classification of a country's economic status as developed or developing often involves factors such as life expectancy and its underlying variables. This research aims to compare the performance of three machine learning algorithms, namely KNN (K-Nearest Neighbors), naive Bayes, and binomial logistic regression, in classifying the economic status of countries as developed or developing. The data used in this study is "Life Expectancy (WHO) Fixed," obtained from the Kaggle website. The first statistical analysis conducted was Principal Component Analysis (PCA) using 16 predictor variables. PCA resulted in three principal components capable of explaining 71.41% of the variance, which were subsequently used in the KNN, naive Bayes, and binomial logistic regression methods. The analysis results from the KNN, naive Bayes, and binomial logistic regression methods produced F1-scores of 100%, 98.19%, and 97.36%, respectively.

Keywords:
Life Expectancy; Countries Classification; KNN; Naive Bayes; Binomial Logistic Regression

Style Sitasi:

N. K. K. Ardana, *et al.*, "Perbandingan Metode KNN, Naive Bayes, dan Regresi Logistik Binomial dalam Pengklasifikasian Status Ekonomi Negara", *Jambura J. Math.*, vol. 5, No. 2, pp. 404–418, 2023, doi: https://doi.org/10.34312/jjom.v5i2.21103

e-ISSN: 2656-1344
Received: 14-07-2023 | Revised: 30-07-2023 | Accepted: 03-08-2023| Published: 05-08-2023

1. Pendahuluan

Perbedaan kondisi perekonomian suatu negara saat ini masih membuat perbedaan beban penyakit atau *Burden of Disease* (BoD) yang memengaruhi angka harapan hidup (AHH) [1]. Harapan hidup adalah indikator kesehatan dan kesejahteraan suatu negara yang digambarkan dengan rata-rata usia yang dicapai seseorang dalam ruang lingkungan tertentu [2]. AHH WHO dihitung sebagai rata-rata lama hidup suatu populasi dan harapan hidup dipengaruhi oleh beberapa faktor, seperti mortalitas bayi di bawah satu tahun, mortalitas bayi di bawah lima tahun, kematian orang dewasa, penyakit menular, penyakit tidak menular, kecelakaan lalu lintas, pembunuhan, konsumsi alkohol, kondisi lingkungan, serta indeks massa tubuh [1]. Harapan hidup juga ditentukan oleh berbagai faktor seperti perkembangan ekonomi, fasilitas kesehatan, pendidikan, dan laju kematian.

Khan, *et al*. [3] menyebutkan bahwa AHH di negara berkembang lebih rendah dibandingkan dengan AHH di negara maju. Riset tersebut menjelaskan bahwa faktor-faktor yang membuat AHH negara maju lebih tinggi adalah sumber daya finansial, standar dan fasilitas kesehatan, dan kondisi lingkungan sekitar. Menurut Freeman, *et al*. [4], berdasarkan perbandingan tiga negara, faktor yang memengaruhi AHH adalah sistem kesejahteraan, partisipasi politik, serta kekuatan dan aksesibilitas warga sipil kepada pekerjaan, perumahan, air bersih, lingkungan sehat, dan pendidikan. Menurut Miladinov [5], berdasarkan hasil penelitian pada lima negara, pendapatan per kapita yang tinggi dan kematian bayi yang rendah membuat tingginya AHH.

Penelitian-penelitian tersebut mengindikasikan bahwa suatu status negara, yakni maju atau berkembang, dapat ditentukan melalui AHH dan peubah-peubah yang mendasarinya. Oleh karena itu, dibuatlah model prediktif untuk mengklasifikasikan negara maju atau berkembang berdasarkan AHH dan peubah-peubah penyusunnya. Tiga algoritma pembelajaran mesin yang digunakan untuk menentukan kategori negara maju dan berkembang pada penelitian ini adalah *K-Nearest Neighbors* (KNN), *naive Bayes*, dan regresi logistik. Metode KNN dipilih karena data yang digunakan merupakan data sekunder dan tujuan utama dari metode ini adalah mengklasifikasi objek baru berdasarkan atribut lama dan *training sample* [6]. Algoritma KNN ini memprediksi kategori suatu data baru yang ditambahkan, dalam hal ini negara, dan menentukan letaknya lebih dekat ke kategori maju atau berkembang. Metode KNN ini memiliki tingkat efektivitas yang tinggi, produktif, dan *effortless* untuk mengklasifikasikan data [7]. Metode ini mudah untuk diterapkan, efektif pada data besar, dan cepat dalam memproses data latih [8]. *Naive Bayes* mengklasifikasi data menggunakan pendugaan fungsi kepekatan peluang dengan asumsi prediktor bebas [9]. *Naive Bayes* dapat diterapkan di berbagai aplikasi pada dunia nyata [10]. Metode ini memiliki keunggulan jika diterapkan pada data besar yang memiliki *noise*, data tidak lengkap, dan kurang relevan [11]. *Naive Bayes* juga memiliki kecepatan yang baik untuk *training* dan baik saat menghadapi tipe data baru [12]. Hal ini menjadi alasan pemilihan metode *naive Bayes*. Alasan penggunaan regresi logistik adalah peubah respons yang digunakan bersifat *dummy* yang mengindikasikan penerimaan atau penolakan kategori yang ditetapkan [13] yang dalam hal ini adalah negara maju atau berkembang. Keuntungan utama dari metode ini adalah dapat menghindari efek pengganggu dengan menganalisis hubungan semua variabel secara bersamaan [14]. Regresi logistik memiliki kemiripan dengan analisis diskriminan dalam hal fungsinya untuk

menentukan apakah peluang peubah respons termasuk pada kategori tertentu dapat diprediksi oleh peubah-peubah prediktor. Pada penelitian ini, kategori peubah respons hanya ada dua, yaitu negara maju atau berkembang sehingga metode regresi logistik yang digunakan adalah regresi logistik binomial [15].

Penelitian ini bertujuan untuk membandingkan performa tiga algoritma pembelajaran mesin, yaitu KNN, *naive Bayes*, dan regresi logistik binomial dalam mengklasifikasikan negara maju dan berkembang berdasarkan AHH dan peubah-peubah yang mendasarinya. Penelitian ini juga menunjukkan faktor-faktor yang signifikan yang menjadi perhatian utama dalam penentuan klasifikasi negara maju dan berkembang oleh WHO. Klasifikasi negara-negara ke dalam kategori maju dan berkembang ini diharapkan dapat memiliki implikasi yang baik untuk pembuatan kebijakan dan alokasi sumber daya dan sumber dana pemerintah negara terkait serta menjadi acuan bagi negara-negara berkembang untuk mengejar ketertinggalan atau mengatasi permasalahan di negaranya.

2. Metode

2.1. Data dan Tahapan Penelitian

Data yang digunakan pada penelitian ini adalah gugus data (*dataset*) yang berjudul "Life Expectancy (WHO) Fixed" yang diperoleh dari situs Kaggle. Gugus data ini berisi angka harapan hidup, kesehatan, imunisasi, ekonomi, dan informasi demografi pada 179 negara dari tahun 2000 hingga tahun 2015. Gugus data ini memiliki 2864 observasi yang mana 85% data (2421 observasi) digunakan sebagai data latih dan 15% sisanya (443 observasi) digunakan sebagai data uji. Dari gugus data ini, kategori negara maju atau berkembang digunakan sebagai peubah respons (Y) dan 16 peubah seperti pada Tabel 1 dan Tabel 2 digunakan sebagai peubah-peubah prediktor. Beberapa kolom yang tidak diperlukan, yakni nama negara, daerah, dan tahun, dieliminasi.

Tabel 1. Peubah-peubah prediktor yang digunakan

Peubah	Keterangan
Kematian bayi baru lahir (X_1)	Banyaknya kematian bayi baru lahir per 1000 populasi
Kematian anak di bawah usia 5 tahun (X_2)	Banyaknya kematian anak di bawah 5 tahun per 1000 populasi
Kematian orang dewasa (X_3)	Banyaknya kematian orang dewasa per 1000 populasi
Konsumsi alkohol (X_4)	Konsumsi alkohol per orang dengan usia di atas 15 tahun (dalam liter)
Hepatitis B (X_5)	Persentase anak usia 1 tahun dengan imunisasi Hepatitis B (HepB3)
Campak (X_6)	Persentase anak usia 1 tahun dengan vaksin campak dosis pertama (MCV1)
BMI (X_7)	Indeks massa tubuh/*body mass index* (BMI) yang dihitung dengan membagi massa tubuh terhadap kuadrat tinggi (kg/m^2)
Polio (X_8)	Persentasi anak usia 1 tahun dengan imunisasi Polio (Pol3)
Difteri (X_9)	Persentase anak usia 1 tahun dengan imunisasi *Diphtheria, tetanus toxoid, and pertussis* (DTP3)
HIV (X_{10})	Kejadian HIV (*Human Immunodeficiency Virus*) per 1000 populasi pada usia 15-59 tahun
PDB (X_{11})	Produk Domestik Bruto (PDB) per kapita dalam US Dolar

Tabel 2. Peubah-peubah prediktor yang digunakan (Lanjutan)

Peubah	Keterangan
Populasi (X_{12})	Total populasi (dalam juta jiwa)
Kekurusan usia 10-19 tahun (X_{13})	Prevalensi kekurusan 10-19 tahun dengan simpangan baku BMI < -2 di bawah median
Kekurusan usia 5-9 tahun (X_{14})	Prevalensi kekurusan 5-9 tahun dengan simpangan baku BMI < -2 di bawah median
Pendidikan (X_{15})	Rataan lama sekolah formal (dalam tahun) untuk orang-orang dengan usia di atas 25 tahun
AHH (X_{16})	Rataan AHH pada tahun 2010 sampai 2015

Tahapan yang dilakukan pada penelitian ini adalah sebagai berikut:

1. Mereduksi dimensi dengan analisis komponen utama (AKU).
2. Mengklasifikasikan status ekonomi negara menggunakan komponen utama yang terpilih pada AKU dengan metode KNN.
3. Mengklasifikasikan status ekonomi negara menggunakan komponen utama yang terpilih pada AKU dengan metode *naive Bayes*.
4. Mengklasifikasikan status ekonomi negara menggunakan komponen utama yang terpilih pada AKU dengan metode regresi logistik binomial.
5. Membandingkan performa dari ketiga metode yang digunakan.

2.2. *Analisis Komponen Utama*

Tahapan pertama yang dilakukan pada penelitian ini adalah mereduksi dimensi dengan analisis komponen utama (AKU). Menurut Everitt dan Dunn [16] dan Jolliffe [17], AKU merupakan suatu pendekatan yang digunakan untuk mereduksi dimensi data yang memiliki jumlah peubah yang banyak. Tujuan utama dari AKU adalah untuk mengurangi kompleksitas data dengan memperkecil dimensi, mengungkapkan pola atau struktur yang terdapat pada data, serta menghilangkan korelasi antara peubah-peubah yang ada. Beberapa hal yang dilakukan saat menerapkan AKU adalah sebagai berikut:

2.2.1. Menormalisasi data

Data diolah menggunakan pendekatan *Z-Score* dengan formula seperti pada Persamaan (1):

$$Z_{i,j} = \frac{X_{i,j} - \overline{X}_j}{\sigma_j}, \quad i = 1,\ 2,\ 3, \ldots,\ n, \quad j = 1,\ 2,\ 3, \ldots,\ p \tag{1}$$

dengan $X_{i,j}$ adalah data pada baris (observasi) ke-i dan kolom (peubah) ke-j, $\overline{X}_j$ adalah rata-rata data pada kolom ke-j, σ_j adalah simpangan baku data pada kolom ke-j, n adalah banyaknya observasi, dan p adalah banyaknya peubah.

2.2.2. Menentukan matriks koragam

Matriks koragam diperlukan dalam menentukan nilai dan vektor eigen. Formula yang digunakan untuk menghitung koragam antarpeubah adalah persamaan (2):

$$S_{j,k} = Cov\left(X_j, X_k\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{i,j} - \overline{X}_j\right)\left(X_{i,k} - \overline{X}_k\right), \;\; j,\, k = 1,\, 2,\, 3,\, \ldots,\, p. \tag{2}$$

Selanjutnya, masing-masing koragam yang telah diperoleh menggunakan Persamaan (2) disusun ke sebuah matriks sehingga diperoleh matriks koragam seperti pada Persamaan (3)

$$S = \begin{bmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,p} \\ S_{2,1} & S_{2,2} & \cdots & S_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p,1} & S_{p,2} & \cdots & S_{p,p} \end{bmatrix}_{p\times p}. \tag{3}$$

2.2.3. Menentukan nilai dan vektor eigen

Untuk memperoleh nilai eigen, digunakan formula seperti pada Persamaan (4):

$$\left|S - \lambda I_p\right| = 0. \tag{4}$$

Setelah mendapatkan nilai eigen, vektor eigen dapat diperoleh menggunakan formula seperti pada Persamaan (5):

$$\left(S - \lambda I_p\right) a = 0. \tag{5}$$

2.2.4. Menentukan komponen utama

Komponen utama diperoleh dari kombinasi linear matriks vektor eigen dan juga matriks data seperti pada persamaan (6):

$$\begin{aligned} KU_1 &= \sum_{j=1}^{p} a_{j,1}X_j = a_{1,1}X_1 + a_{2,1}X_2 + \cdots + a_{p,1}X_p \\ KU_2 &= \sum_{j=1}^{p} a_{j,2}X_j = a_{1,2}X_1 + a_{2,2}X_2 + \cdots + a_{p,2}X_p \\ &\vdots \\ KU_p &= \sum_{j=1}^{p} a_{j,p}X_j = a_{1,p}X_1 + a_{2,p}X_2 + \cdots + a_{p,p}X_p. \end{aligned} \tag{6}$$

Persamaan (6) dapat dibuat dalam bentuk perkalian matriks seperti pada Persamaan (7):

$$\begin{bmatrix} KU_1 \\ KU_2 \\ \vdots \\ KU_p \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,p} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p,1} & a_{p,2} & \cdots & a_{p,p} \end{bmatrix}_{p\times p} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}_{p\times 1} \tag{7}$$

dengan syarat komponen utama yang pertama memiliki ragam sampel yang paling tinggi. Untuk itu, diperlukan penyelesaian dari masalah seperti pada Persamaan (8):

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} X_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1. \tag{8}$$

Selanjutnya, diperoleh Persamaan (9):

$$KU_{p\times 1} = a^T X_{p\times 1}^T \tag{9}$$

dengan a adalah elemen *loading* atau koefisien dalam membentuk kombinasi linear dari peubah-peubah yang ada.

2.3. K-Nearest Neighbors (KNN)

KNN merupakan suatu metode untuk mengklasifikasi suatu data ke dalam satu dari beberapa kelas menggunakan sejumlah tetangga terdekat dari titik tersebut [18]. Metode ini disebut sebagai metode pembelajaran mesin yang "malas" [19] karena metode ini tidak menggunakan data latih untuk membentuk sebuah model melainkan hanya mencocokkan data latih dengan data uji. Jarak Euclid antara titik data yang memiliki koordinat (x_1, y_1) dengan titik data yang memiliki koordinat (x_2, y_2) dapat dihitung menggunakan formula pada Persamaan (10) [20]:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \tag{10}$$

Apabila data yang ingin diklasifikasikan memiliki dimensi lebih dari 2, jarak Euclid antartitik data dapat dihitung menggunakan formula seperti pada Persamaan (11):

$$d(x_{i_1}, x_{i_2}) = \sqrt{\sum_{j=1}^{p} (x_{i_1,j} - x_{i_2,j})^2}. \tag{11}$$

2.4. Naive Bayes

Naive Bayes merupakan metode klasifikasi yang membutuhkan berbagai petunjuk dalam menentukan kelas yang tepat bagi data sampel [21]. Metode ini memerlukan penghitungan peluang dari setiap nilai atribut untuk membangun model *naive Bayes*. Metode ini merupakan metode klasifikasi yang sederhana dan setiap atributnya memiliki sifat yang saling bebas sehingga setiap atributnya dimungkinkan untuk memiliki kontribusi terhadap keputusan akhir [22]. Metode ini cocok digunakan saat data yang diteliti terdiri atas atribut kategori karena metode ini dapat menghitung peluang berdasarkan sebaran kategori yang ada dalam data tersebut. Hal ini bermanfaat dalam mengklasifikasikan data seperti analisis sentimen, klasifikasi teks, atau sistem rekomendasi. Selain itu, metode ini juga dapat menghasilkan hasil yang memuaskan meskipun jumlah sampel dalam gugus data terbatas. Meskipun metode ini memiliki asumsi kebebasan yang kuat, metode ini tetap stabil dan dapat diandalkan. Selain itu, metode ini juga dapat memberikan kinerja yang baik pada gugus data

dengan fitur-fitur yang tidak relevan atau dimensi yang tinggi. Dengan mengasumsikan kebebasan antarfitur, metode ini dapat mengabaikan fitur-fitur yang tidak berpengaruh pada saat pengklasifikasian data. Persamaan (12) digunakan pada saat mengklasifikasikan data, yaitu:

$$\Pr\left(C \left| F_1, F_2, \ldots, F_p \right.\right) = \Pr\left(C\right) \prod_{j=1}^{p} \Pr\left(F_j \right| C\left.\right) \tag{12}$$

dengan C adalah peubah yang mewakili kelas dan $F_1, F_2, \ldots, F_p$ mewakili petunjuk-petunjuk atau karakteristik-karakteristik yang dibutuhkan dalam pengklasifikasian.

2.5. Regresi Logistik Binomial

Regresi logistik binomial merupakan suatu metode klasifikasi data yang hasil pengklasifikasiannya hanya ada dua kemungkinan. Metode ini menggunakan konsep peluang. Regresi logistik memodelkan peluang posterior dari dua kelas melalui fungsi linear dari peubah-peubah prediktor serta memastikan bahwa semua peluang tersebut terjumlah menjadi satu dan masing-masing tetap berada dalam selang [0, 1]. Mengacu pada [9], model regresi logistik untuk p peubah prediktor diberikan oleh Persamaan (13):

$$\log\left(\frac{p\left(X\right)}{1-p\left(X\right)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \tag{13}$$

Ruas kiri pada Persamaan (13) disebut dengan logaritma *odds* (logit), sedangkan ruas kanan merupakan fungsi linear dari peubah-peubah prediktor. Peluang $p\left(X\right)$ dan peubah X berhubungan secara sigmoid, sedangkan fungsi logit dan peubah X berhubungan secara linear [23]. Hal ini berakibat pada setiap perubahan satu satuan peubah X, mengakibatkan perubahan peluang sebesar logit tersebut. Dari fungsi tersebut, diambil dengan *cut-off* sebesar 0.5 untuk pengklasifikasian, artinya jika $p\left(X\right) \geq 0.5$, maka data dikategorikan sebagai negara maju dan jika $p\left(X\right) < 0.5$, maka data dikategorikan sebagai negara berkembang.

2.6. Confusion Matrix

Tabel yang biasanya digunakan untuk mengevaluasi kinerja hasil klasifikasi suatu metode adalah *confusion matrix*. Matriks ini menyajikan informasi tentang seberapa baik suatu metode dapat mengklasifikasikan data dengan cara membandingkan label hasil prediksi yang telah dilakukan terhadap label aktual. Setiap baris menyajikan label aktual dari data, sedangkan setiap kolom menyajikan label hasil prediksi. Skema matriks ini terdapat pada Tabel 3.

Tabel 3. Skema *confusion matrix*

	Predicted Positive	*Predicted Negative*
Actual Positive	TP (*True Positive*)	FN (*False Negative*)
Actual Negative	FP (*False Positive*)	TN (*True Negative*)

Suryadewiansyah dan Tju [24] menjelaskan bahwa *confusion matrix* adalah sebuah

konstruksi yang terdiri atas empat kategori, yaitu *true positive* (kasus aktual positif yang diprediksi secara benar, yaitu positif juga), *false negative* (kasus aktual positif yang salah diprediksi, yaitu sebagai negatif), *false positive* (kasus aktual negatif yang salah diprediksi, yaitu sebagai positif), dan *true negative* (kasus aktual negatif yang diprediksi secara benar, yaitu negatif juga). Terdapat juga *rule of thumb* bagi *confusion matrix*, yaitu acuan praktis yang digunakan untuk memberikan pemahaman awal tentang kinerja metode klasifikasi berdasarkan sebaran nilai dalam *confusion matrix*. Beberapa acuan ini meliputi:

1. Jika jumlah TP dan TN tinggi, serta jumlah FP dan FN rendah, maka model dianggap memiliki kinerja yang baik.
2. Jika jumlah FP sangat tinggi, model cenderung menyajikan lebih banyak prediksi positif palsu, atau mengklasifikasikan terlalu banyak label sebagai positif (*over-prediction* bagi model).
3. Jika jumlah FN sangat tinggi, model cenderung menyajikan lebih banyak prediksi negatif palsu, atau gagal mendeteksi banyak label yang sebenarnya positif (*under-prediction* bagi model).
4. Jika jumlah TP rendah, model memiliki masalah dalam mengidentifikasi label positif dengan benar, sehingga perlu penyesuaian atau perbaikan lagi.
5. Jika jumlah TN rendah, model memiliki masalah dalam mengidentifikasi label negatif dengan benar, sehingga perlu penyesuaian atau perbaikan lagi.

Namun, perlu diingat bahwa *rule of thumb* dalam *confusion matrix* hanya memberikan pemahaman awal yang kasar tentang performa metode klasifikasi. Evaluasi yang lebih mendalam dan akurat melibatkan ukuran evaluasi yang lebih rinci dan mempertimbangkan konteks serta tujuan penggunaan metode tersebut [25]. Pada penelitian ini disajikan empat bentuk ukuran ketepatan masing-masing metode dalam mengklasifikasikan data, yaitu melalui akurasi, presisi, *recall* (sensitivitas), dan *F1-Score*.

Akurasi digunakan untuk membandingkan seberapa sering model benar mengklasifikasikan terhadap total keseluruhan data dan dihitung menggunakan formula seperti pada Persamaan (14):

$$\text{Akurasi} = \frac{\text{TP} + \text{TN}}{\text{Total}}. \tag{14}$$

Presisi digunakan untuk membandingkan seberapa sering hasil prediksi positif itu benar saat model memprediksi positif dan dihitung menggunakan formula seperti pada Persamaan (15):

$$\text{Presisi} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{15}$$

Recall digunakan untuk membandingkan seberapa sering model memprediksi positif dari semua contoh yang sebenarnya positif dan dihitung menggunakan formula seperti pada Persamaan (16):

$$\textit{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{16}$$

F1-Score digunakan untuk mengukur kinerja model klasifikasi pada gugus data dengan

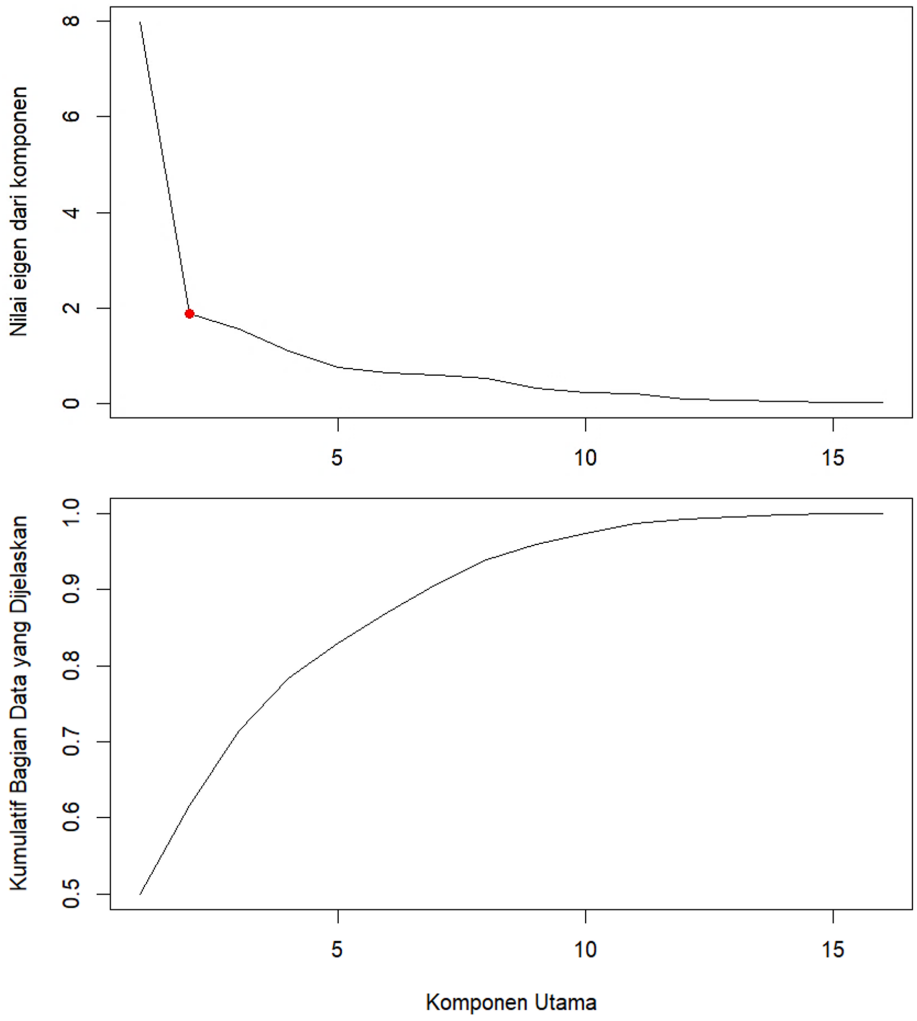
ketidakseimbangan antara contoh positif dan negatif. Presisi dan *recall* dapat menghasilkan informasi yang kurang akurat dalam kasus ini, sedangkan *F1-Score* dapat memberikan pengukuran yang lebih seimbang antara keduanya. *F1-Score* dihitung menggunakan rataan harmonik dari presisi dan *recall* dengan formula seperti pada Persamaan (17):

$$F1 - Score = 2\frac{\text{presisi} \times recall}{\text{presisi} + recall}. \tag{17}$$

3. Hasil dan Pembahasan

3.1. Reduksi Dimensi dengan Analisis Komponen Utama

Banyaknya peubah prediktor direduksi menggunakan analisis komponen utama (AKU) agar proses pengklasifikasian data menjadi lebih efisien. AKU menghasilkan 16 komponen utama sebagai kombinasi linear dari 16 peubah prediktor awal. *Scree plot* dan nilai eigen kumulatif yang diperoleh setelah AKU diterapkan pada data latih ditampilkan pada Gambar 1.



Gambar 1. Nilai eigen dan kumulatif bagian data yang dijelaskan

Everitt dan Dunn [16] menyarankan untuk mengambil komponen utama yang mampu menjelaskan keragaman hingga 70%-90%. Apabila hanya digunakan komponen utama pertama, proporsi keragaman yang dapat dijelaskan hanya sebesar 49,93%. Apabila digunakan dua komponen utama, proporsi keragaman yang dapat dijelaskan belum mencapai 70%, yakni sebesar 61,67%. Apabila digunakan tiga komponen utama, proporsi keragaman yang dapat dijelaskan adalah sebesar 71,41% dan nilai ini sudah berada pada interval yang disarankan oleh Everitt dan Dunn [16]. Oleh sebab itu, pada kasus ini dipilih tiga komponen utama. Dengan mengambil tiga komponen utama, didapat formulasi baru untuk setiap komponen seperti pada Tabel 4. Formulasi ini kemudian dijadikan landasan untuk membuat nilai-nilai baru pada setiap amatan yang akan digunakan dalam membuat model prediktif. Tiga komponen utama ini diharapkan sudah cukup untuk mewakili seluruh peubah prediktor awal untuk memprediksi status perekonomian suatu negara.

Tabel 4. Formulasi tiga komponen utama

Peubah prediktor awal	*Loading* komponen utama 1	*Loading* komponen utama 2	*Loading* komponen utama 3
X_1	−0,33253*	−0,09661	0,050959
X_2	−0,3283	−0,12574	0,063354
X_3	−0,28683	−0,1446	0,371128
X_4	0,185364	−0,25233	0,059882
X_5	0,202822	0,284212	0,406251
X_6	0,221634	0,018411	0,187224
X_7	0,255479	−0,17872	0,044906
X_8	0,276618	0,276482	0,294765
X_9	0,271705	0,291326	0,306142
X_{10}	−0,141	−0,12018	0,546306*
X_{11}	0,211019	−0,13966	−0,16587
X_{12}	−0,0328	0,325205	−0,27972
X_{13}	−0,22449	0,473662*	−0,04393
X_{14}	−0,22333	0,473437	−0,04572
X_{15}	0,297396	−0,12937	−0,01428
X_{16}	0,323946	0,104719	−0,24682

*Nilai mutlak terbesar

Nilai *loading* dalam AKU merujuk pada bobot atau kontribusi setiap peubah prediktor awal dalam pembentukan komponen utama. Hal ini mengindikasikan sejauh mana peubah prediktor awal berperan dalam menjelaskan keragaman dalam gugus data. Setiap peubah prediktor awal diberikan *loading* untuk setiap komponen utama yang dihasilkan dalam AKU. *Loading* menggambarkan pentingnya peubah tersebut dalam membentuk komponen utama. Nilai *loading* berkisar antara -1 hingga 1 dengan tanda menunjukkan arah hubungan antara peubah prediktor awal dan komponen utama. *Loading* yang tinggi menandakan kontribusi yang signifikan, sedangkan *loading* yang mendekati nol menunjukkan kontribusi yang rendah atau tidak signifikan. Jika suatu peubah memiliki *loading* yang tinggi dan positif pada komponen utama pertama, maka peubah tersebut memiliki pengaruh yang kuat dalam menjelaskan keragaman pada komponen utama pertama. Sebaliknya, jika suatu peubah memiliki *loading* yang tinggi dan negatif, maka peubah tersebut memiliki pengaruh yang berlawanan dengan komponen utama tersebut. Dengan melihat *loading*, peubah yang paling penting dalam menjelaskan keragaman dalam gugus data dapat diidentifikasi. Peubah dengan *loading* tinggi dianggap memiliki pengaruh besar dalam membentuk pola dan struktur data.

Nilai *loading* pada setiap komponen utama menjadi koefisien bagi setiap peubah prediktor awal yang bersesuaian. Nilai koefisien yang tinggi menandakan nilai *loading* yang diberikan tinggi, begitu pula sebaliknya. Semakin besar nilai koefisien yang dimiliki, semakin besar pula pengaruhnya dalam pengklasifikasian ini. Pembuatan nilai KU_1, KU_2, dan KU_3 untuk setiap titik data berturut-turut diberikan oleh Persamaan (18), (19), dan (20):

$$\begin{aligned} KU_1 = &-0,3325X_1 - 0,3283X_2 - 0,2868X_3 + 0,1854X_4 + 0,2028X_5 + 0,2216X_6 + \\ &0.2555X_7 + 0,2766X_8 + 0,2717X_9 - 0,1410X_{10} + 0,2110X_{11} - 0,0328X_{12} - \\ &0,2245X_{13} - 0,2233X_{14} - 0,2974X_{15} + 0,3240X_{16} \end{aligned} \tag{18}$$

$$\begin{aligned} KU_2 = &-0,0966X_1 - 0,1257X_2 - 0,1446X_3 - 0,2523X_4 + 0,2842X_5 + 0,0184X_6 - \\ &0,1787X_7 + 0,2765X_8 + 0,2913X_9 - 0,1202X_{10} - 0,1397X_{11} + 0,3252X_{12} + \\ &0,4737X_{13} - 0,4734X_{14} - 0,1294X_{15} + 0,1047X_{16} \end{aligned} \tag{19}$$

$$\begin{aligned} KU_3 = &\,0,0510X_1 + 0,0634X_2 + 0,3711X_3 + 0,0599X_4 + 0,4063X_5 + 0,1872X_6 + \\ &0,0449X_7 + 0,2948X_8 + 0,3061X_9 + 0,5463X_{10} - 0,1659X_{11} - 0,2797X_{12} - \\ &0,0439X_{13} - 0,0457X_{14} - 0,0143X_{15} - 0,2468X_{16}. \end{aligned} \tag{20}$$

Dapat dilihat bahwa pada KU_1, peubah X_1 memilki pengaruh paling besar dengan koefisien negatif sebesar $-0{,}3325$, sedangkan pada KU_2, peubah X_{13} memberikan pengaruh terbesar dengan koefisien positif sebesar 0,4670. Terakhir, pada KU_3, peubah X_{10} adalah peubah yang paling berpengaruh dengan koefisien positif sebesar 0,5463. Oleh sebab itu, peubah yang paling memengaruhi dalam proses klasifikasi dengan tiga komponen utama adalah kematian bayi baru lahir, kekurusan usia 10-19 tahun, serta kejadian HIV pada usia 15-59 tahun.

3.2. Klasifikasi dengan KNN

Metode KNN diterapkan sebagai model prediktif pertama untuk melakukan klasifikasi terhadap data uji. Tiga komponen utama pertama hasil AKU digunakan sebagai peubah-peubah bebas pada metode ini. Tiga komponen utama tersebut diyakini sudah cukup untuk mewakili 16 peubah prediktor awal. Validasi silang (*cross validation*) dilakukan terhadap metode KNN dan teknik validasi silang yang digunakan adalah teknik *k-fold validation* sebanyak 5 *folds*. Hasil klasifikasinya disajikan pada Tabel 5.

Tabel 5. *Confusion matrix* hasil KNN

Aktual	Hasil prediksi		Jumlah baris
	Berkembang	Maju	
Berkembang	361	0	361
Maju	0	82	82
Jumlah kolom	361	82	443

Tabel 5 menunjukkan bahwa hasil prediksi semua data sesuai dengan kategori

aktualnya. Dalam kasus ini, terlihat bahwa metode KNN berhasil memprediksi semua data uji dengan tepat.

3.3. Klasifikasi dengan Naive Bayes

Metode *naive Bayes* diterapkan sebagai model prediktif kedua untuk melakukan klasifikasi terhadap data uji. Sama seperti halnya pada metode KNN, pada metode ini juga digunakan tiga komponen utama pertama sebagai peubah-peubah bebasnya. Validasi silang yang dilakukan terhadap metode ini adalah teknik *k-fold validation* sebanyak 10 *folds*. Hasil klasifikasinya disajikan pada Tabel 6.

Tabel 6. *Confusion matrix* hasil *naive Bayes*

Aktual	Hasil prediksi		Jumlah baris
	Berkembang	Maju	
Berkembang	352	9	361
Maju	4	78	82
Jumlah kolom	356	87	443

Tabel 6 menunjukkan bahwa ada 352 data yang diprediksi masuk kategori negara berkembang dan sesuai dengan kategori aktualnya, yaitu negara berkembang juga. Ada 9 data yang kategori aktualnya negara berkembang tetapi diprediksi sebagai negara maju. Ada 4 data yang diprediksi sebagai negara berkembang, padahal kategori aktualnya adalah negara maju. Ada juga sebanyak 78 data yang diprediksi sebagai negara maju dan sesuai dengan kategori aktualnya, yaitu negara maju juga.

3.4. Klasifikasi dengan Regresi Logistik Binomial

Metode regresi logistik binomial diterapkan sebagai model prediktif ketiga untuk melakukan klasifikasi terhadap data uji. Sama seperti metode KNN dan *naive Bayes*, metode ini juga menggunakan komponen utama pertama, kedua, dan ketiga sebagai peubah-peubah bebasnya. Validasi silang yang dilakukan terhadap metode ini adalah teknik *k-fold validation* sebanyak 10 *folds*. Hasil klasifikasinya disajikan pada Tabel 7.

Tabel 7. *Confusion matrix* hasil regresi logistik binomial

Aktual	Hasil prediksi		Jumlah baris
	Berkembang	Maju	
Berkembang	350	11	361
Maju	8	74	82
Jumlah kolom	358	85	443

Tabel 7 menunjukkan bahwa ada 350 data yang diprediksi masuk kategori negara berkembang dan sesuai dengan kategori aktualnya, yaitu negara berkembang juga. Sementara itu, ada 8 data yang kategori aktualnya negara maju tetapi diprediksi sebagai negara berkembang. Terdapat 11 data yang diprediksi sebagai negara maju, padahal kategori aktualnya adalah negara berkembang. Sisanya, sebanyak 74 data yang diprediksi sebagai negara maju dan sesuai dengan kategori aktualnya, yaitu negara maju juga.

3.5. Perbandingan Performa Ketiga Metode

Setelah diperoleh hasil klasifikasi dengan KNN, *naive Bayes*, dan regresi logistik binomial, ditentukan metode yang paling baik dalam melakukan klasifikasi negara berkembang dan negara maju yang dipublikasi oleh WHO. Dari *confusion matrices* yang telah diperoleh, elemen-elemennya diambil untuk menghitung empat ukuran ketepatan yang digunakan dalam melakukan perbandingan, yaitu akurasi, presisi, *recall*, dan *F1-Score*. Perbandingan keempat ukuran tersebut untuk tiga metode yang digunakan disajikan pada Tabel 8.

Tabel 8. Perbandingan ukuran ketepatan

Ukuran	KNN	*Naive Bayes*	Regresi logistik binomial
Akurasi	100%*	97,07%	95,71%
Presisi	100%*	98,88%	97,77%
Recall	100%*	97,51%	96,95%
F1-Score	100%*	98,19%	97,36%

*Nilai terbesar untuk masing-masing ukuran ketepatan

Berdasarkan Tabel 8, nilai akurasi terbesar dihasilkan oleh metode KNN yang mengindikasikan 100% data berhasil (benar) diklasifikasikan sebagai negara maju maupun berkembang berdasarkan data aktualnya. Metode KNN juga menghasilkan 100% untuk presisi, *recall*, dan *F1-Score* karena metode ini berhasil memprediksi semua data uji dengan benar. Hal ini secara jelas menunjukkan bahwa metode KNN merupakan metode yang paling baik.

4. Kesimpulan

Penelitian ini berhasil menentukan algoritma pembelajaran mesin terbaik dalam mengklasifikasikan status ekonomi negara-negara di dunia berdasarkan angka harapan hidup dan peubah-peubah yang mendasarinya berdasarkan data yang dipublikasi oleh WHO. Pada penelitian ini, 16 peubah prediktor awal direduksi menjadi tiga komponen utama dengan proporsi keragaman yang dapat dijelaskan sudah melebihi 70%. Untuk mengklasifikasikan data uji, metode KNN merupakan metode terbaik apabila dibandingkan dengan *naive Bayes* dan regresi logistik binomial. Metode KNN memiliki F1-*score* terbesar dengan nilai sempurna (100%), sedangkan metode *naive Bayes* dan regresi logistik binomial berturut-turut memiliki F1-*score* sebesar 98,19% dan 97,36%.

Referensi

[1] WHO, *World health statistics 2022: monitoring health for the SDGs, sustainable development goals*. Geneva: World Health Organization, 2022, [Online] Available: https://pesquisa.bvsalud.org/portal/resource/pt/who-356584.

[2] R. Muda, R. A. Koleangan, and J. B. Kalangi, "Pengaruh angka harapan hidup, tingkat pendidikan dan pengeluaran perkapita terhadap pertumbuhan ekonomi di sulawesi utara pada tahun 2003-2017," *Jurnal Berkala Ilmiah Efisiensi*, vol. 19, no. 1, pp. 44–55, 2019, [Online] Available: https://ejournal.unsrat.ac.id/index.php/jbie/article/view/22368.

[3] A. Khan, S. Khan, and M. Khan, "Factors effecting life expectancy in developed and developing countries of the world (an approach to available literature)," *International Journal of Yoga, Physiotherapy and Physical Education*, vol. 1, no. 1, pp. 31–33, 2016.

[4] T. Freeman, H. A. Gesesew, C. Bambra, E. R. J. Giugliani, J. Popay, D. Sanders, J. Macinko, C. Musolino, and F. Baum, "Why do some countries do better or worse in life expectancy relative to income? an analysis of brazil, ethiopia, and the united states of america,"

International Journal for Equity in Health, vol. 19, no. 1, p. 202, 2020, doi: 10.1186/s12939-020-01315-z.

[5] G. Miladinov, "Socioeconomic development and life expectancy relationship: evidence from the eu accession candidate countries," *Genus*, vol. 76, no. 1, pp. 1–20, 2020, doi: 10.1186/s41118-019-0071-0.

[6] Y. A. Setianto, K. Kusrini, and H. Henderi, "Penerapan algoritma k-nearest neighbour dalam menentukan pembinaan koperasi kabupaten kotawaringin timur," *Creative Information Technology Journal*, vol. 5, no. 3, pp. 232–241, 2019, doi: 10.24076/citec.2018v5i3.179.

[7] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE, 2019, pp. 1255–1260, doi: 10.1109/ICCS45141.2019.9065747.

[8] N. Bhatia and Vandana, "Survey of nearest neighbor technique," *International Journal of Computer Science and Information Security*, vol. 8, no. 2, pp. 302–305, 2010.

[9] G. James, D. Witten, T. Hastie, R. Tibshirani *et al.*, *An introduction to statistical learning*, 2nd ed. New York: Springer, 2013, vol. 112.

[10] I. Wickramasinghe and H. Kalutarage, "Naive bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," *Soft Computing*, vol. 25, no. 3, pp. 2277–2293, 2021, doi: 10.1007/s00500-020-05297-6.

[11] H. Muhamad, C. A. Prasojo, N. A. Sugianto, L. Surtiningsih, and I. Cholissodin, "Optimasi naïve bayes classifier dengan menggunakan particle swarm optimization pada data iris," *J. Teknol. Inf. dan Ilmu Komput*, vol. 4, no. 3, pp. 180–184, 2017.

[12] C. Zhang, D. Jia, L. Wang, W. Wang, F. Liu, and A. Yang, "Comparative research on network intrusion detection methods based on machine learning," *Computers & Security*, vol. 121, p. 102861, 2022, doi: 10.1016/j.cose.2022.102861.

[13] Y. Tampil, H. Komaliq, and Y. Langi, "Analisis regresi logistik untuk menentukan faktor-faktor yang mempengaruhi indeks prestasi kumulatif (ipk) mahasiswa fmipa universitas sam ratulangi manado," *d'CARTESIAN*, vol. 6, no. 2, pp. 56–62, 2017, doi: 10.35799/dc.6.2.2017.17023.

[14] S. Sperandei, "Understanding logistic regression analysis," *Biochemia Medica*, vol. 24, no. 1, pp. 12–18, 2014, doi: 10.11613/BM.2014.003.

[15] T. Abedin, Z. Chowdhury, A. Afzal, F. Yeasmin, and T. Turin, "Application of binary logistic regression in clinical research," *Journal of National Heart Foundation of Bangladesh*, vol. 5, no. 1, pp. 8–11, 2016.

[16] B. Everitt, G. Dunn *et al.*, *Applied multivariate data analysis*, 2nd ed. London: Wiley Online Library, 2001, vol. 2.

[17] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer-Verlag, 2002, doi: 10.1007/b98835.

[18] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer New York, 2009, doi: 10.1007/978-0-387-84858-7.

[19] L. Farokhah, "Implementasi k-nearest neighbor untuk klasifikasi bunga dengan ekstraksi fitur warna rgb," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 6, pp. 1129–1136, 2020, doi: 10.25126/jtiik.2020722608.

[20] W. Wahyono, I. N. P. Trisna, S. L. Sariwening, M. Fajar, and D. Wijayanto, "Comparison of distance measurement on k-nearest neighbour in textual data classification," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 1, pp. 54–58, 2020, doi: 10.14710/jtsiskom.8.1.2020.54-58.

[21] A. W. Syaputri, E. Irwandi, and M. Mustakim, "Naïve bayes algorithm for classification of student major's specialization," *Journal of Intelligent Computing & Health Informatics*, vol. 1, no. 1, pp. 17–21, 2020, doi: 10.26714/jichi.v1i1.5570.

[22] M. Hasan, "Prediksi tingkat kelancaran pembayaran kredit bank menggunakan algoritma naïve bayes berbasis forward selection," *ILKOM Jurnal Ilmiah*, vol. 9, no. 3, pp. 317–324, 2017, doi: 10.33096/ilkom.v9i3.163.317-324.

[23] D. H. Ismunarti, "Regresi logistik binomial, model untuk toksisitas logam berat timbal pb terhadap larva udang vannamae," *Buletin Oseanografi Marina*, vol. 1, no. 5, pp. 47–52, 2012.

[24] M. K. Suryadewiansyah and T. E. E. Tju, "Naïve bayes dan confusion matrix untuk efisiensi analisa intrusion detection system alert," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 8, no. 2, pp. 81–88, 2022, doi: 10.25077/TEKNOSI.v8i2.2022.81-88.

[25] I. W. Saputro and B. W. Sari, "Uji performa algoritma naïve bayes untuk prediksi masa studi mahasiswa," *Creative Information Technology Journal*, vol. 6, no. 1, pp. 1–11, 2020, doi: 10.24076/citec.2019v6i1.178.