UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2 Nomor 4 Tahun 2024 Halaman 483-488.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Comparison of Naïve Bayes and K-Nearest Neighbors Methods in Classifying Human Development Index by Districts/City Indonesia in 2022

Rudi Anggara, Tessy Octavia Mukhti*, Yenni Kurniawati dan Dina Fitria
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: tessyoctaviam@fmipa.unp.ac.id

Submitted : 23 Oktober 2024
Revised : 11 November 2024
Accepted : 24 November 2024

ABSTRACT

The Human Development Index (HDI) is an indicator used to measure the success of efforts to improve the quality of human life in a particular region. Indonesia's HDI has increased every year, but the HDI in several districts/cities in Indonesia remains in the low category. The low HDI in these districts/cities is due to unequal development between regions in Indonesia. This disparity in development is influenced by HDI indicators as well as other factors. To address this issue, a decision system is needed to determine HDI categories using the Naive Bayes and KNN methods. Naive Bayes is applied with the assumption of Gaussian distribution, while KNN is implemented with the optimization of the nearest K value. Model performance evaluation is conducted to determine the best accuracy of the two methods using a confusion matrix. The analysis results show that the Naïve Bayes model outperforms the KNN algorithm in classifying the Human Development Index (HDI) by district/city in Indonesia for the year 2022, with Naïve Bayes achieving an accuracy of 93%. Therefore, the Naïve Bayes algorithm show good performance in terms of accuracy.

Keywords: *Confusion Matrix, K-Nearest Neighbors , Naive Bayes*

I. PENDAHULUAN

Indeks Pembangunan Manusia (IPM) adalah suatu indikator yang digunakan untuk mengukur keberhasilan dalam upaya membangun kualitas hidup manusia di suatu daerah. IPM dibangun dari tiga aspek dasar, yaitu usia yang panjang dan hidup sehat, pendidikan, serta standar hidup yang layak (BPS, 2022). Menurut BPS (2022), ada empat kategori klasifikasi IPM, kategori rendah (IPM < 60), kategori sedang (60 ≤ IPM < 70), kategori tinggi (70 ≤ IPM 80) , dan kategori sangat tinggi (IPM ≥ 80). Pada tahun 2015 hingga tahun 2022, angka IPM Indonesia terus mengalami peningkatan setiap tahun. Di tahun 2015, IPM Indonesia yang hanya sebesar 69,55 poin terus meningkat hingga pada tahun 2022 angka IPM Indonesia berada pada angka 72,91 poin. Walaupun terjadi peningkatan setiap tahun, IPM Indonesia masih ditemukan permasalahan yaitu tidak meratanya IPM di setiap kabupaten/kota di Indonesia. Tidak meratanya pembangunan yang terjadi dipengaruhi oleh berbagai faktor, termasuk faktor-faktor dari indikator pembangun IPM seperti umur harapan hidup (UHH), rata-rata lama sekolah, harapan lama sekolah, serta pendapatan per kapita, maupun faktor-faktor lainnya.

Untuk menyikapi masalah tersebut, diperlukan sebuah sistem keputusan yang dapat dengan cepat dan akurat dalam menentukan kategori IPM di setiap kabupaten/kota, guna meningkatkan efektivitas kinerja pemerintah dalam menganalisis kategori IPM di setiap wilayah. Penentuan kategori IPM dapat dipermudah dengan penerapan metode klasifikasi. Menurut Han & Kamber (2006), klasifikasi yaitu langkah yang digunakan untuk mengidentifikasi model dan membedakan kelas data, yang bertujuan untuk memprediksi kelas dari objek yang label kelasnya belum diketahui. Klasifikasi yang akan dilakukan menggunakan metode *Naïve Bayes* dan *K-Nearest Neighbors* (KNN). *Naïve Bayes* melakukan klasifikasi statistik untuk memprediksi probabilitas keanggotaan kelas tertentu (Hang & Kamber, 2006). Sedangkan KNN yaitu metode klasifikasi yang berbasis jarak. KNN dilakukan dengan mengidentifikasi K objek terdekat (paling mirip) dalam data *training* yang sesuai dengan objek dalam data *testing* (Wu & Kumar, 2009). Beberapa penelitian terdahulu mengenai *Naïve Bayes* dan KNN antara lain, Putri, dkk (2022) menerapkan klasifikasi *Naïve Bayes* dan KNN pada analisis data penyakit jantung. Hasil penelitian menunjukkan bahwa metode *Naïve Bayes* mencapai akurasi sebesar 86,17%, sedangkan metode KNN dengan parameter K=9 memperoleh akurasi sebesar 85,11%.

DOI: https://doi.org/10.24036/ujsds/vol2-iss4/319

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Kemudian penelitian dari Lubis (2021), melakukan klasifikasi *Decision Tree* dan *Naïve Bayes* pada Indeks Pembangunan Manusia di Provinsi Sumatera Utara. Hasil dari penelitian ini menunjukkan bahwa metode *Naïve Bayes* mencapai akurasi sebesar 91%, sedangkan metode *Decision Tree* memperoleh akurasi sebesar 88%.

Berdasarkan uraian permasalahan di atas, maka pada artikel ini akan membahas perbandingan antara metode *Naïve Bayes* dan KNN dalam mengklasifikasikan Indeks Pembangunan Manusia menurut Kabupaten/Kota di Indonesia . Tujuan dari penelitian ini adalah untuk menentukan tingkat akurasi yang diperoleh dari kedua metode, yaitu *Naïve Bayes* dan KNN.

II. METODE PENELITIAN

Penelitian yang dilakukan menggunakan data sekunder, yaitu data yang diambil dari data Badan Pusat Statistik (BPS) tahun 2022. Jumlah data yang digunakan sesuai dengan kabupaten dan kota di Indonesia tahun 2022 sebanyak 514 data. Variabel penelitian yang akan digunakan dilihat pada Tabel 1.

Tabel 1. Variabel Penelitian

Variabel	Keterangan	Satuan	Skala Data
Y	Indeks Pembangunan Manusia (IPM)	Indeks	Ordinal
X_1	Umur Harapan Hidup saat Lahir (UHH)	Tahun	Rasio
X_2	Rata-rata Lama Sekolah (RLS)	Tahun	Rasio
X_3	Harapan Lama Sekolah (HLS)	Tahun	Rasio
X_4	Pengeluaran per Kapita Disesuaikan	Rupiah	Rasio

Data diatas akan di analisis menggunakan metode *Naïve Bayes* dan KNN. Langkah-langkah yang dilakukan dalam melakukan analisis data sebagai berikut.

1. Menginput dan melakukan *pre processing* data.
2. Membagi data *training* dan *testing*
   Pada penelitian ini data dibagi menjadi data *training* sebesar 80% dan data *testing* sebesar 20%.
3. Melakukan pengklasifikasian menggunakan metode *Naïve Bayes* dan KNN.
   a. *Naïve Bayes*
   1) Menghitung kelas *probability* untuk mencari nilai *prior probability* dari data *training* menggunakan Persamaan (1).

$$P(H) = \frac{x_i}{n} \tag{1}$$

Dimana :
$P(H)$: Peluang terjadinya H
x_i : Nilai variabel ke-i
n : Total sampel

2) Mencari nilai *mean* dan standar deviasi setiap variabel pada Persamaan (2) dan Persamaan (3).

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} \tag{3}$$

Dimana :
μ : Rata-rata (*mean*)
σ : Standar deviasi
x_i : Variabel ke-i
n : Total sampel

DOI: https://doi.org/10.24036/ujsds/vol2-iss4/319

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

3) Menghitung nilai *Densitas Gauss* menggunakan Persamaan (4).

$$P\left((X_i = x_i | Y = y_j\right) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i-\pi_{ij})^2}{2\sigma_{ij}^2}} \quad (4)$$

Dimana :
P : Peluang
X_i : Variabel ke- i
x_i : Nilai variabel ke- i
Y : Kelas yang akan ditentukan
y_i : Sub kelas Y yang akan ditentukan
π : Rata-rata (*mean*)
σ : Standar deviasi

4) Menghitung nilai probabilitas akhir (*prior probability*)
Langkah terakhir dalam metode *Naïve Bayes* yaitu memprediksi label kelas, dengan menghitung *posterior probability*. *Posterior probability* dapat dihitung pada Persamaan (5).

$$P(H|X_1, X_2, \dots, X_n) = P(H) \,.P(X_1|H).P(X_2|H).\dots.P(X_n|H) \quad (5)$$

b. *K-Nearest Neighbors* (KNN)
Menurut Suntoro (2019), cara menghitung KNN dengan menggunakan *euclidean distance* sebagai berikut.
1) Menentukan nilai parameter K.. Parameter K yang digunakan adalah K=1, K=3, K=5, dan K=7
2) Mengukur jarak antara data *training* dan data *testing* menggunakan *euclidean distance*. *Euclidean distance* menggunakan persamaan (6).

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (6)$$

Dimana :
$D(x, y)$: Jarak amatan x ke amatan y
n : Total amatan
x_i : Data *testing* ke- i
y_i : Data *training* ke- i

3) Mengurutkan data yang mempunyai jarak terkecil hingga terbesar.
4) Menetapkan kelas. Kelas akan yang akan dipilih merupakan kelas dengan jumlah nilai K terbanyak dalam data *testing*.

4. Menentukan tingkat akurasi metode *Naïve Bayes* dan KNN menggunakan *confusion matrix*.
Confusion matrix adalah suatu metode yang digunakan untuk menilai dan mengevaluasi kinerja suatu metode klasifikasi. Beberapa nilai yang dihitung pada *confusion matrix*, yaitu akurasi, *specificity*, dan *recall*. *Confusion matrix* dapat dilihat pada Tabel 2.

Tabel 2. Tabel *Confusion Matrix*

Klasifikasi		**Kelas Prediksi**	
		Kelas = 1	Kelas = 0
Kelas Asli	Kelas = 1	*True Positive* (TP)	*False Negative* (FN)
	Kelas = 0	*False Positive* (FP)	*True Negative* (TN)

(Sumber: Gorunescu, F. 2011)

Keterangan :
1. *True Positive* (TP) : Kelas asli benar dengan hasil kelas prediksi positif

DOI: https://doi.org/10.24036/ujsds/vol2-iss4/319

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

2. *True Negative* (TN) : Kelas asli benar dengan hasil kelas prediksi negatif.
3. *False Positive* (FP) : Kelas asli salah dengan hasil kelas prediksi positif.
4. *False Negative* (FN) : Kelas asli salah dengan hasil kelas prediksi negatif.

Evaluasi hasil kinerja klasifikasi dengan *confusion matrix* mendapatkan nilai *accuracy, specificity* dan *sensitivity* (Rokarch dan Maimon, 2015)

$$Accuracy\ (\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (7)$$

$$Sensitivity(\%) = \frac{TP}{TP + FN} \times 100\% \quad (8)$$

$$Specificity\ (\%) = \frac{TN}{FP + FN} \times 100\% \quad (9)$$

5. Membandingkan metode *Naïve Bayes* dan KNN berdasarkan hasil akurasi

III. HASIL DAN PEMBAHASAN

A. *Preprocessing* Data

Preprocessing yang akan dilakukan yaitu menstandarisasi data. Standarisasi digunakan untuk menyesuaikan skala data, terutama jika distribusi data tidak normal. Standarisasi dilakukan dengan merubah data menjadi distribusi dengan rata-rata (*mean*) 0 dan standar deviasi 1.

Tabel 3. Standarisasi Data

Amatan	AHH	HLS	RLS	PK
1	-1.30669	0.72737	0.760159	-1.18988
2	-0.66936	0.085006	0.959472	-0.59973
3	-1.5534	0.208537	1.227777	-0.8328
⋮	⋮	⋮	⋮	⋮
512	-1.17453	-3.26887	-4.15366	-1.82512
513	-1.25383	-3.26887	-2.49017	-2.12183
514	0.244053	1.968861	1.496083	1.652894

B. Data *Training* dan Data *Testing*

Sebelum melakukan analisis klasifikasi tehadap IPM menurut kab/kota di Indonesia tahun 2022. Data yang diperoleh terdiri dari 514 amatan, yang dibagi menjadi 411 amatan sebagai data *training* dan 103 amatan sebagai data *testing*.

C. *Naïve Bayes*

Kinerja metode *Naïve Bayes* dalam klasifikasi yang berisi *confusion matrix* hasil prediksi, dilihat pada Tabel 4.

Tabel 4. *Confusion Matrix Naïve Bayes*

Klasifikasi		Kelas Prediksi			
		Sangat Tinggi	Tinggi	Sedang	Rendah
Kelas Asli	Sangat Tinggi	7	0	0	0
	Tinggi	2	37	4	0
	Sedang	0	0	44	0
	Rendah	0	0	1	8

Berdasarkan Tabel 4, hasil dari *confusion matrix* metode *Naïve Bayes* sebagai berikut :

Pada kategori **sangat tinggi**, *True Positive* (TP) = 7, artinya kelas asli adalah sangat tinggi, dan prediksi sangat tinggi. *False Positive* (FP) = 2, artinya kelas asli adalah tinggi (2), sedang (0), rendah (0), tapi diprediksi sebagai sangat tinggi. *False Negative* (FN) = 0, artinya kelas asli adalah sangat tinggi, tapi diprediksi sebagai tinggi (0), sedang (0),

DOI: https://doi.org/10.24036/ujsds/vol2-iss4/319

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

rendah (0). *True Negative* (TN) = 94, artinya 94 amatan diprediksi dengan benar bahwa amatan tersebut tidak termasuk kedalam kategori sangat tinggi (berada pada kategori tinggi (37,0,0), sedang (4,44,1), rendah (0,0,8).

Pada kategori **tinggi**, *True Positive* (TP) = 37, artinya kelas asli adalah tinggi, dan prediksi tinggi. FP = 0, artinya kelas asli adalah sangat tinggi (0), sedang (0), rendah (0), tapi diprediksi sebagai tinggi. FN = 6, artinya kelas asli adalah tinggi, tapi diprediksi sebagai sangat tinggi (2), sedang (4), rendah (0). TN = 60, artinya 60 amatan diprediksi dengan benar bahwa amatan tersebut tidak termasuk kedalam kategori tinggi (berada pada kategori sangat tinggi (7,0,0), sedang (0,44,1), rendah (0,0,8).

Pada kategori **sedang**, *True Positive* (TP) = 44, artinya kelas asli adalah sedang, dan prediksi sedang. FP = 5, artinya kelas asli adalah sangat tinggi (0), tinggi (4), rendah (1), tapi diprediksi sebagai sedang. FN = 0, artinya kelas asli adalah sedang, tapi diprediksi sebagai sangat tinggi (0), tinggi (0), rendah (0). TN = 54, artinya 54 amatan diprediksi dengan benar bahwa amatan tersebut tidak termasuk kedalam kategori sedang (berada pada kategori sangat tinggi (7,2,0), tinggi (0,37,0), rendah (0,0,8).

Pada kategori **rendah**, *True Positive* (TP) = 8, artinya kelas asli adalah rendah, dan prediksi rendah. FP = 0, artinya kelas asli adalah sangat tinggi (0), tinggi (0), sedang (0), tapi diprediksi sebagai rendah. FN = 1, artinya kelas asli adalah rendah, tapi diprediksi sebagai sangat tinggi (0), tinggi (0), sedang (1). TN = 94, artinya 94 amatan diprediksi dengan benar bahwa amatan tersebut tidak termasuk kedalam kategori rendah (berada pada kategori sangat tinggi (7,2,0), tinggi (0,37,0), sedang (0,4,44).

D. *K-Nearest Neighbors*

Klasifikasi dilakukan menggunakan metode KNN dengan mencari nilai parameter K=1, 3, 5, dan 7 untuk memperoleh akurasi terbaik, dapat ditampilkan pada Tabel 5.

Tabel 5. Hasil Akurasi Parameter KNN

Parameter K	Akurasi (%)
1	0,89
3	0,91
5	0,90
7	0,92

Berdasarkan Tabel 5, akurasi tertinggi terdapat pada parameter K=7 dengan nilai akurasi sebesar 92%. Untuk melihat hasil pengujian atau performa metode KNN dalam klasifikasi, dapat dilihat melalui *confusion matrix* yang terdapat pada Tabel 6.

Tabel 6. *Confusion Matrix* KNN

Klasifikasi		Kelas Prediksi			
		Sangat Tinggi	Tinggi	Sedang	Rendah
Kelas Asli	Sangat Tinggi	7	0	0	0
	Tinggi	0	39	4	0
	Sedang	0	1	43	0
	Rendah	0	0	3	6

Berdasarkan Tabel 6, hasil dari *confusion matrix* metode KNN sebagai berikut :

Pada kategori **sangat tinggi**, *True Positive* (TP) = 7, artinya kelas asli adalah sangat tinggi, dan prediksi sangat tinggi. *False Positive* (FP) = 0, artinya kelas asli adalah tinggi (0), sedang (0), rendah (0), tapi diprediksi sebagai sangat tinggi. *False Negative* (FN) = 0, artinya kelas asli adalah sangat tinggi, tapi diprediksi sebagai tinggi (0), sedang (0), rendah (0). *True Negative* (TN) = 96, artinya 96 amatan diprediksi dengan benar bahwa amatan tersebut tidak termasuk kedalam kategori sangat tinggi (berada pada kategori tinggi (39,1,0), sedang (4,43,3), rendah (0,0,6).

Pada kategori **tinggi**, *True Positive* (TP) = 39, artinya kelas asli adalah tinggi, dan prediksi tinggi. FP = 1, artinya kelas asli adalah sangat tinggi (0), sedang (1), rendah (0), tapi diprediksi sebagai tinggi. FN = 4, artinya kelas asli adalah tinggi, tapi diprediksi sebagai sangat tinggi (0), sedang (4), rendah (0). TN = 59, artinya 59 amatan diprediksi dengan benar bahwa amatan tersebut tidak termasuk kedalam kategori tinggi (berada pada kategori sangat tinggi (7,0,0), sedang (0,43,3), rendah (0,0,6).

Pada kategori **sedang**, *True Positive* (TP) = 43, artinya kelas asli adalah sedang, dan prediksi sedang. FP = 7, artinya kelas asli adalah sangat tinggi (0), tinggi (4), rendah (3), tapi diprediksi sebagai sedang. FN = 1, artinya kelas

DOI: https://doi.org/10.24036/ujsds/vol2-iss4/319

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

asli adalah sedang, tapi diprediksi sebagai sangat tinggi (0), tinggi (1), rendah (0). TN = 52, artinya 52 amatan diprediksi dengan benar bahwa amatan tersebut tidak termasuk kedalam kategori sedang (berada pada kategori sangat tinggi (7,0,0), tinggi (0,39,0), rendah (0,0,6).

Pada kategori **rendah**, *True Positive* (TP) = 6, artinya kelas asli adalah rendah, dan prediksi rendah. FP = 0, artinya kelas asli adalah sangat tinggi (0), tinggi (0), sedang (0), tapi diprediksi sebagai rendah. FN = 3, artinya kelas asli adalah rendah, tapi diprediksi sebagai sangat tinggi (0), tinggi (0), sedang (3). TN = 94, artinya 94 amatan diprediksi dengan benar bahwa amatan tersebut tidak termasuk kedalam kategori rendah (berada pada kategori sangat tinggi (7,0,0), tinggi (0,39,1), sedang (0,4,43).

E. Pengukuran Kinerja Klasifikasi

Berdasarkan hasil klasifikasi menggunakan *confusion matrix* dengan metode *Naïve Bayes* dan KNN terhadap data IPM kabupaten/kota di Indonesia tahun 2022, hasil *confusion matrix* ditampilkan pada Tabel 4 dan Tabel 6. Hasil akurasi kedua metode dapat dilihat padaTabel 7.

Tabel 7. Hasil Ketepatan Klasifikasi *Naïve Bayes* dan KNN

Akurasi *Naïve Bayes*	Akurasi KNN
0,93	0,92

Berdasarkan Tabel 7, dalam memprediksi ketepatan kategori IPM di Indonesia menggunakan teknik klasifikasi *Naïve Bayes* dan KNN dengan parameter K=7, metode *Naïve Bayes* memberikan tingkat akurasi tertinggi sebesar 93%. Ini menunjukkan bahwa model *Naïve Bayes* memiliki keakuratan 93% dalam mengklasifikasikan data dengan benar.

IV. KESIMPULAN

Penelitian menggunakan metode *Naïve Bayes* dan KNN untuk memprediksi Indeks Pembangunan Manusia (IPM) menurut kabupaten/kota di Indonesia pada tahun 2022. Hasil analisis menunjukkan bahwa model *Naïve Bayes* unggul dibandingkan metode KNN dalam mengklasifikasikan IPM, dengan nilai akurasi *Naïve Bayes* mencapai 93%, yang menunjukkan performa akurasi yang baik. Diharapkan penelitian ini dapat memberikan masukan dan pertimbangan kepada pemerintah dalam merumuskan kebijakan yang berkaitan dengan kualitas hidup masyarakat dalam konteks IPM. Untuk penelitian mendatang, disarankan untuk menggunakan metode klasifikasi lain seperti *Neural Network*, dan *Support Vector Machine* dan *Decision Tree* dengan jumlah amatan yang lebih banyak untuk memperoleh hasil yang lebih baik.

DAFTAR PUSTAKA

Badan Pusat Statistik (BPS). 2022. “Indeks Pembangunan Manusia”, diakses pada 15 Januari 2023 pukul 08.00.

_____. 2022. “Metode Baru Indeks Pembangunan Manusia”, diakses pada 15 Januari 2023 pukul 09.00

Gorunescu, F. (2011). *Data Mining: Concepts and Techniques.* Berlin: Springer.

Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques, second edition.* California: Morgan Kaufman.

Lubis, R. M. (2021). Analisis Metode Decision Tree Algorithm Dan Metode Naive Bayes Mengidentifikasi Pertumbuhan Indeks Pembangunan Manusia (IPM) di Provinsi Sumatera Utara. *Jutisal (Jurnal Teknik Informatika Komputer Universal), 1*(2), 9-16.

Putri, R. W., Ristyawan, A., & Muzaki, M. N. (2022). Perbandingan Kinerja Algoritma K-NN dan NBC Untuk Klasifikasi Penyakit Jantung. *JTECS : Jurnal Sistem Telekomunikasi Elektronika Sistem Kontrol Power Sistem & Komputer, 2*(2), 143-154.

Rokarch, L., & Maimon, O. (2015). *Data Mining With Decision Trees Theory And Application 2nd Ed*. Singapore: World Scientific.

Suntoro, J. (2019). *Data Mining : Algoritma dan Implementasi dengan Pemograman PHP.* Jakarta: Elex Media Komputindo.

Wu, X., & Kumar, V. (2009). *The Top Ten Algorithms in Data Mining.* London: CRS Press Taylor & Francis Group.

DOI: https://doi.org/10.24036/ujsds/vol2-iss4/319