E-ISSN 2527-9378

Jurnal Statistika Industri dan Komputasi
Volume 08, No. 1, Januari 2023, pp. 58-67

PERBANDINGAN METODE *K-NEAREST NEIGHBOR* DAN *RANDOM FOREST* PADA KLASIFIKASI INDEKS PEMBANGUNAN MANUSIA DI KABUPATEN/KOTA SELURUH INDONESIA

Nurafidah[1], Kris Suryowati[2], Maria Titah Jatipaningrum[3*]
Jurusan Statistika, Fakultas Sains Terapan, Institut Sains & Teknologi AKPRIND Yogyakarta
Email: nurafidahfidah2@gmail.com
**corresponding author*

Abstract

The Human Development Index in Indonesian regencies/cities varies, this is due to the uneven development in Indonesia. HDI in Indonesia is used as one of the allocators for determining the General Allocation Fund and to measure the performance of the government. The purpose of this study is to compare the K-Nearest Neighbor and Random Forest methods in the HDI classification. In the K-Nearest Neighbor method in classifying using $k = 10$ *and in the Random Forest method in classifying using* $ntree = 100$ *and* $mtry = 2$*. The variables used in this study include the dependent variable, namely HDI with low, medium, high and very high HDI categories, while the independent variables include HLS, RLS, UHH and PPKD. Classification using the K-Nearest Neighbor method produces 124 correct classification data and 5 misclassified data, while the Random Forest method produces 119 correctly classified data and 10 misclassified data. The classification results also show that the K-Nearest Neighbor method is the best method to use in classifying HDI in regencies/cities throughout Indonesia, because it produces an accuracy value and an average AUC of 96.12% and 0.9618, which is greater than the Random Forest method, which is 92.25. % and 0.9538 and the K-Nearest Neighbor method produces a smaller error rate of 3.88% compared to the Random Forest method of 7.75%. The results of this study are expected to be information to continue to increase the Human Development Index in regencies/cities throughout Indonesia in 2022.*

Keywords: *Human Development Index, K-Nearest Neighbor, Random Forest*

Abstrak

Indeks Pembangunan Manusia yang ada di Kabupaten/Kota Indonesia beragam, hal ini sebabkan oleh pembangunan yang ada di Indonesia masih tidak merata. IPM di Indonesia digunakan sebagai salah satu alokator penentuan Dana Alokasi Umum dan untuk mengukur kinerja dari pemerintah. Tujuan dari penelitian ini yaitu membandingkan metode *K-Nearest Neighbor* dan *Random Forest* pada klasifikasi IPM. Pada metode *K-Nearest Neighbor* dalam melakukan klasifikasi menggunakan $k = 10$ dan pada metode *Random Forest* dalam melakukan klasifikasi menggunakan $ntree = 100$ dan $mtry = 2$. Variabel-variabel yang digunakan pada penelitian ini diantaranya variabel dependent terdapat IPM dengan kategori IPM rendah, sedang, tinggi dan sangat tinggi, sedangkan pada variabel independen terdapat HLS, RLS, UHH dan PPKD. Klasifikasi dengan menggunakan metode *K-Nearest Neighbor* menghasilkan 124 data benar klasifikasi dan 5 data kesalahan klasifikasi sedangkan metode *Random Forest* menghasilkan 119 data benar lasifikasi dan 10 data kesalahan klasifikasi. Hasil klasifikasi juga menunjukkkan bahwa metode *K-Nearest Neighbor* adalah metode yang terbaik untuk digunakan dalam melakukan klasifikasi IPM di Kabupaten/Kota seluruh Indonesia, karena menghasilkan nilai akurasi dan rata-rata *AUC* sebesar 96.12% dan 0.9618 lebih besar dibandingkan metode *Random Forest* yaitu sebesar 92.25% dan 0.9538 serta metode *K-Nearest Neighbor* menghasilkan nilai *error rate* yang lebih kecil yaitu sebesar 3.88% dibandingkan metode *Random Forest* sebesar 7.75%. Hasil dari penelitian ini diharapkan dapat menjadi informasi untuk terus meningkat Indeks Pembangunan Manusia di Kabupaten/Kota seluruh Indonesia pada tahun 2022.

Kata kunci: *Indeks Pembangunan Manusia, K-Nearest Neighbor, Random Forest*

1. Pendahuluan

Indeks Pembangunan Manusia merupakan salah satu alat yang digunakan untuk mengukur capaian pembangunan manusia berbasis sejumlah komponen dasar kualitas hidup di suatu wilayah atau negara. Sebagai alat ukuran kualitas hidup, IPM dibangun melalui pendekatan tiga dimensi dasar. Dimensi pertama yaitu umur panjang dan hidup sehat, dimana diukur menggunakan indikator umur harapan hidup saat lahir, dimensi kedua adalah pengetahuan diukur menggunakan harapan lama sekolah dan rata-rata lama sekolah, dan dimensi ketiga yaitu standar hidup layak diukur dengan pengeluaran riil per kapita disesuaikan [1].

Indeks pembangunan Manusia di Indonesia pada tahun 2021 mencapai 72.29, dimana meningkat sebesar 0.35 poin (0.49%) dari Indeks pembangunan Manusia di Indonesia pada tahun 2020 yang sebesar 71.94. Peningkatan IPM pada tahun 2021 terjadi pada semua dimensi diantaranya dimensi umur panjang dan hidup sehat, pengetahuan dan standar hidup layak. Pada dimensi hidup layak yang diukur berdasarkan rata-rata pengeluaran riil per kapita yang disesuaikan meningkat sebesar 1.30%. Serta pada dimensi umur panjang dan hidup sehat pada tahun 2021 bayi yang lahir memiliki harapan untuk dapat hidup hingga 71.57 tahun, dimana lebih lama 0.10 tahun dari tahun 2020. Sedangkan pada dimensi pendidikan, penduduk yang berusia 7 tahun memiliki harapan lama sekolah selama 13.08 tahun, dimana meningkat sebesar 0.10 tahun dibandingkat tahun 2020 dan rata-rata lama untuk penduduk yang berusia 25 tahun meningkat sebesar 0.06 tahun, dari 8.48 tahun menjadi 8.54 tahun [2].

Klasifikasi Indeks Pembangunan Manusia di Indonesia dilakukan karena IPM di Kabupaten/Kota beragam yang disebabkan oleh pembangunan yang ada di Indonesia tidak merata. IPM juga menjadi salah satu alokator penentuan Dana Alokasi Umum dan untuk mengukur kinerja dari pemerintah. Indeks Pembagunan Manusia dikategorikan menjadi 4 yaitu Indeks Pembangunan Manusia rendah apabila $IPM < 60$, sedang jika $60 \leq IPM < 70$, tinggi jika $70 \leq IPM < 80$ dan sangat tinggi bila $IPM \geq 80$ [1]. Oleh karena itu klasifikasi Indeks Pembangunan Manusia di Kabupaten/Kota seluruh Indonesia menggunakan metode *K-Nearest Neighbor* dan *Random Forest*.

Beberapa penelitian terdahulu yang telah dilakukan berkaitan dengan klasifikasi Indeks Pembangunan Manusia dengan menggunakan metode *k-nearest neighbor* dan *random forest* diantaranya:

a. Darsyah (2017) melakukan klasifikasi Indeks Pembangunan Manusia (IPM) dengan pendekatan *K-Nearest Neighbor* (K-NN), dimana dalam penelitian ini menggunakan jumlah *k* atau tetangga terdekat lebih dari satu. Penelitian ini menghasilkan nilai akurasi tertinggi dengan menggunakan nilai $k = 5$ dan 10 dengan tingkat akurasi mencapai 91.43%, dengan sensitivitas 100% dan spesivitas 83.33%.
b. Fauzi (2017) yang membandingkan metode *K-Nearest Neighbor* (K-NN) dan *Support Vector Machine* (SVM) untuk Klasifikasi Indeks Pembangunan Manusia Provinsi Jawa Tengah, dimana pada penelitian ini menggunakan empat nilai *k* yang berbeda untuk metode *K-Nearest Neighbor* dan pada metode *Support Vector Machine* nilai parameter gamma dan C yang digunakan untuk menghitung kernel *Radial Basis Function* masing-masing tiga. Hasil perbandingan akurasi kedua metode menunjukkan bahwa metode terbaik adalah SVM dengan parameter $gamma = 1, C = 1, 10, 100$ dan nilai akurasi sebesar 95.36%
c. Mauludiyah (2020) menggunakan metode *Random Forest* pada klasifikasi Indeks Pembangunan Manusia kabupaten/kota di Indonesia, pada penelitian ini variabel *dependent* yang digunakan lebih dari dua kategori dan jumlah mtry ada tiga dan ntree ada lima. Penelitian ini menghasilkan nilai akurasi klasifikasi sebesar 93.69%.
d. Rachmi (2020) menggunakan metode klasifikasi *Random Forest* dan *Extreme Gradient Boosting* untuk memprediksi *churn* pelanggan dengan data *churn* telekomunikasi di distrik Columbia, pada penelitian ini nilai akurasi yang dihasilkan sangat baik yaitu lebih dari 90%. Hasil analisis klasifikasi menunjukkan bahwa metode *XGBoost* lebih unggul dibandingkan metode *random forest*, hal ini ditunjukkan oleh nilai akurasi dan AUC

metode *XGboost* sebesar 95.6% dan 0.876, sedangkan metode *random forest* diperoleh nilai akurasi dan AUC sebesar 93.5% dan 0.799.

Tujuan dari penelitian ini melakukan klasifikasi Indeks Pembangunan Manusia di Kabupaten/Kota seluruh Indonesia pada tahun 2021 dan membandingkan nilai akurasi, *error rate* dan *AUC* dari metode *k-nearest neighbor* dan *random forest* untuk mengetahui metode terbaik dalam melakukan klasifikasi.

2. Metode

Penelitian ini menggunakan data sekunder yang diperoleh dari Badan Pusat Statistik yang dipublikasikan pada tahun 2022, dimana data diambil di *website* https://www.bps.go.id. Data yang digunakan ini adalah data Indeks Pembangunan Manusia di Kabupaten/Kota seluruh Indonesia pada tahun 2021.

A. Variabel

Variabel yang digunakan dalam penelitian ini diantaranya variabel *dependent* terdapat Indeks Pembangunan Manusia (IPM) dengan kategori IPM rendah, sedang, tinggi dan sangat tinggi, sedangkan pada variabel *independen* terdapat Harapan Lama Sekolah (HLS), Rata-rata Lama Sekolah (RLS), Umur Harapan Hidup Saat Lahir (UHH) dan Pengeluaran per Kapita Disesuaikan (PPKD).

B. Klasifikasi

Klasifikasi merupakan bagian dari prediksi, dimana nilai yang diprediksi berupa label (Fitriani, Aryanti, Saepudin, & Ardiansyah, 2020). Klasifikasi dapat diartikan sebagai suatu proses untuk menemukan model yang menggambarkan dan dapat memisah kelas data satu dengan yang lainnya, untuk digunakan dalam memprediksi data lain yang belum memiliki kelas data [10]

Langkah-langkah dalam proses klasifikasi adalah sebagai berikut [8]:

1. Membangun model dari data training dengan nilai label kelas diketahui. Algoritma klasifikasi digunakan untuk membuat model dari dataset training.
2. Melakukan pemeriksaan akurasi model dari data train, jika kebenaran model memuaskan maka model digunakan untuk mengklasifikasi data dengan label kelas yang tidak diketahui.

C. K-Nearest Neighbor

K-Nearest Neighbor merupakan salah satu metode yang digunakan untuk melakukan klasifikasi terhadap suatu objek. Metode ini bekerja dengan cara mencari k buah data latih yang jaraknya paling dekat dengan objek tersebut. Pada algoritma k-nearest neighbor nilai k merupakan banyaknya tetangga yang memiliki jarak terdekat dengan objek yang akan digunakan sebagai titik untuk melakukan klasifikasi.

Langkah-langkah algoritma k-nearest neighbor sebagai berikut:

1. Menentukan jumlah k (tetangga terdekat).
2. Menghitung jarak euclidean objek terhadap data training.
3. Mengurutkan hasil pada no 2 secara berurutan dari nilai terendah ke tertinggi.
4. Mengumpulkan kategori variabel dependent (Y) untuk melakukan klasifikasi nearest neighbor berdasarkan nilai k.
5. Menggunakan kategori nearest neighbor paling mayoritas untuk memprediksi objek yang baru.

Menghitung jarak antar objek pada algoritma k-nearest neighbor dapat dilakukan dengan menggunakan beberapa cara, diantaranya *euclidean distance* dan *mahattan distance*. *Euclidean distance* merupakan cara yang sering digunakan dalam mengukur kedekatan atau jarak antar objek [3].

Rumus dari jarak Euclidean sebagai berikut:

$$Jarak(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \quad i = 1,2,\dots,n \tag{1}$$

Keterangan:
$Jarak(X_1, X_2)$: Jarak antar objek x_{1i} dan x_{2i}
x_{1i} : Data *testing*
x_{2i} : Data *training*
n : Banyaknya variabel *independent*

D. Bootstrap

Bootstrap pertama kali diperkenalkan pada tahun 1979 oleh Efron. Bootstrap adalah metode resampling atau pengambilan n sampel dengan pengembalian terhadap n data asli dan dilakukan berkali-kali untuk mencari distribusi sampling dari suatu penduga parameter. Ukuran sampel dalam resampling bootstrap boleh sama dengan sampel data asli atau lebih kecil. Metode bootstrap lebih baik digunakan untuk sampel yang berukuran kecil [9]. Metode *resampling bootstrap* ini akan digunakan untuk pengambilan *n* data sampel dari sebuah *dataset* pada algoritma *random forest*. Metode *Bootstrap* ini digunakan karena dapat mengurangi varians pada saat klasifikasi, dimana varians yang tinggi dapat menyebabkan kesalahan klasifikasi.

E. Random Forest

Metode *random forest* merupakan metode klasifikasi yang terdiri dari gabungan sejumlah pohon klasifikasi dan menciptakan sebuah hutan. *Random forest* adalah salah satu dari sekian banyak metode ensemble yang memiliki tujuan untuk mencari dan meningkatkan hasil akurasi pada suatu klasifikasi data dari sebuah pemilah tunggal yang tidak stabil melalui kombinasi banyak pemilah dari suatu metode yang sama dengan proses *majority voting* untuk memperoleh prediksi pada klasifikasi akhir [7].

Metode *random forest* untuk membangun *tree* menggunakan *informasi gain*. *Informasi gain* digunakan untuk mengukur pemilihan atribut yang akan digunakan sebagai pemisah (split) *node* pada sebuah *tree* [5]. Atribut yang akan digunakan sebagai pemisah (*split*) *node* adalah atribut yang memiliki nilai *informasi gain* terbesar.

Rumus untuk menghitung *information gain* sebagai berikut [5]:

$$Gain(A) = Info(D) - Info_A(D) \tag{2}$$

Keterangan:
$Gain(A)$: Perbedaan antara informasi asli yang dibutuhkan dengan jumlah informasi baru yang didapatkan dari partisi A
$Info(D)$: Rata-rata dari informasi yang dibutuhkan untuk mengetahui label kelas dari tupel D
$Info_A(D)$: Informasi harapan yang dibutuhkan untuk mengklasifikasi suatu tupel dari D berdasarkan partisi dari atribut A

Nilai $Info(D)$ dan $Info_A(D)$ dapat dicari menggunakan persamaan 3 dan 4 dibawah ini:

$$Info(D) = -\sum_{i=1}^{m} p_i * log_2(p_i) \quad i = 1,2,\dots,m \tag{3}$$

Keterangan:
m : Jumlah kelas target
p_i : Probabilitas munculnya kelas ke-*i* pada partisi *D*

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D_j) \quad j = 1,2,\dots,v \tag{4}$$

Keterangan:
v : Jumlah partisi
D_j : Total partisi ke-*j*
D : Jumlah tupel pada semua partisi

Langkah-langkah algoritma Random Forest sebagai berikut [6]:

1. Mengambil n data sampel dari dataset awal dengan menggunakan teknik *resampling bootstrap* dengan pengembalian.
2. Menyusun pohon klasifikasi dari setiap *dataset* hasil *resampling bootstrap*, dengan penentuan pemilah terbaik didasarkan pada variabel prediktor yang diambil secara acak(*mtry*) dan menentukan jumlah pohon (*ntree*). Jumlah *mtry* dapat ditentukan melalui perhitungan $mtry = \frac{1}{2}\sqrt{m}$ atau $mtry = \sqrt{m}$ atau $mtry = 2\sqrt{m}$ dimana m adalah banyak variabel prediktor.
3. Melakukan prediksi klasifikasi data sampel berdasarkan pohon klasifikasi yang terbentuk.
4. Mengulangi langkah 1-3 hingga diperoleh sejumlah pohon klasifikasi yang diiinginkan. Perulangan dilakukan sebanyak k kali.
5. Melakukan prediksi klasifikasi data sampel akhir dengan mengkombinasikan hasil prediksi pohon klasifikasi yang diperoleh.

F. Confusion Matrix

Confusion matrix merupakan suatu alat ukur yang dapat digunakan untuk mengukur seberapa baik model dalam melakukan klasifikasi kelas dari data *testing*.

Tabel 1 Matriks *confusion multiclass*

Prediksi	Aktual			
	M	N	O	P
M	TP_M	TP_{MN}	TP_{MO}	TP_{MP}
N	TP_{NM}	TP_N	TP_{NO}	TP_{NP}
O	TP_{OM}	TP_{ON}	TP_O	TP_{OP}
P	TP_{PM}	TP_{PN}	TP_{PO}	TP_P

Tabel 1 di atas terdapat empat kelas prediksi dengan variabel M, N, O dan P. Matriks confusion multiclass merupakan perkembangan dari matriks confusion binary dimana sebelumnya terdapat FP (False Positive), TP (True Positive), FN (False Negative), dan TN (True Negative). Pada matriks confusion multiclass hanya terdapat TP, untuk penentuan FP adalah kasus-kasus dimana data aktualnya Tidak dan diprediksi Ya, penentuan FN adalah kasus-kasus dimana data aktualnya Ya dan diprediksi Tidak, penentuan TN adalah kasus-kasus dimana data aktualnya Tidak dan prediksinya Tidak dan penentuan TP adalah kasus-kasus dimana data aktualnya Ya dan prediksinya Ya.

Berdasarkan tabel konfusi matriks dapat dihitung beberapa nilai yang dapat digunakan untuk melihat kinerja klasifikasi. Nilai-nilai tersebut diantaranya:

1. *Accurancy* digunakan untuk melihat tingkat kebenaran klasifikasi dalam memprediksi data terhadap data yang sebenarnya. Perhitungan akurasi dapat dilihat di bawah ini:
$$Akurasi = \frac{\sum Data\ uji\ benar\ klasifikasi}{\sum Data\ uji}\ x\ 100\% \quad (5)$$
2. *Error rate* adalah tingkat kesalahan klasifikasi dalam melakukan prediksi data terhadap data yang sebenarnya, dengan perhitungan persamaan sebagai berikut:
$$Error\ rate = \frac{\sum Data\ uji\ salah\ klasifikasi}{\sum Data\ uji}\ x\ 100\% \quad (6)$$
3. *Area Under the Curve* (*AUC*) adalah alat untuk menghitung *Under the ROC Curve*, dimana kurva *ROC* digunakan untuk menilai hasil prediskisi [8].
Berikut perhitungan untuk mencari nilai *AUC*:
$$AUC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (7)$$

Nilai *AUC* untuk empat kelas prediksi dapat menggunakan nilai rata-rata *AUC* dari keempat kelas tersebut. Perhitungan rata-rata *AUC* sebagai berikut:
$$Rata - rata\ AUC = \frac{AUC_M + AUC_N + AUC_O + AUC_P}{4} \quad (8)$$

Metode klasifikasi terbaik ditunjukkan dengan nilai akurasi dan *AUC* yang lebih tinggi serta nilai *error rate* yang lebih kecil.

3. Analisis dan Pembahasan

A. Klasifikasi Metode *K-Nearest Neighbor*

Pengklasifikasian Indeks Pembangunan Manusia dengan menggunakan metode *k-nearest neighbor* adalah sebagai berikut:

1. **Penentuan Jumlah *k***

Klasifikasi dengan metode *k-nearest neighbor* terlebih dahulu menentukan jumlah *k* (tetangga terdekat). Nilai kesalahan klasifikasi pada jumlah *k* dapat dilihat di bawah ini.

Tabel 2 Nilai kesalahan klasifikasi pada jumlah *k*

k	**Kesalahan klasifikasi**
1	0.0753
2	0.0753
3	0.0753
4	0.0753
5	0.0727
6	0.0675
7	0.0623
8	0.0597
9	0.0571
10	0.0545
11	0.0571

Berdasarkan Tabel 2 di atas jumlah k yang digunakan untuk melakukan klasifikasi dengan metode *k-nearest neighbor* adalah $k = 10$. Jumlah $k = 10$ ini digunakan karena memiliki nilai kesalahan klasifikasi paling kecil yaitu sebesar 0.0545.

2. **Hasil Prediksi Klasifikasi**

Hasil prediksi klasifikasi menggunakan metode *k-nearest neighbor* disajikan dalam tabel *confusion matrix*. Tabel *confusion matrix* dapat dilihat di bawah ini.

Tabel 3 *Confusion matrix* metode *k-nearest neighbor*

Prediksi	**Aktual**			
	Rendah	Sedang	Tinggi	Sangat tinggi
Rendah	5	0	0	0
Sedang	1	61	2	0
Tinggi	0	1	48	0
Sangat tinggi	0	0	1	10

Berdasarkan Tabel 3 di atas diperoleh dari data aktualnya rendah dan prediksinya rendah sebanyak 5 data, data aktualnya rendah dan prediksinya sedang sebanyak 1 data, data aktualnya sedang dan prediksinya sedang sebanyak 61 data, data aktualnya sedang dan prediksinya tinggi sebanyak 1 data, data aktualnya tinggi dan prediksinya sedang sebanyak 2 data, data aktualnya tinggi dan prediksinya tinggi sebanyak 48 data, data aktualnya tinggi dan prediksinya sangat tinggi sebanyak 1 data dan data aktualnya sangat tinggi dan prediksinya sangat tinggi sebanyak 10 data.

Klasifikasi dengan menggunakan metode *k-nearest neighbor* memberikan kesalahan klasifikasi sebanyak 5 diantaranya ada Kabupaten Manokwari Selatan yang data aktualnya rendah tetapi hasil prediksinya sedang, Kabupaten Aceh Jaya yang data aktualnya sedang tetapi hasil prediksinya tinggi, Kabupaten Tapanuli Selatan dan Kabupaten Maros yang data aktualnya tinggi tetapi hasil prediksinya sedang dan Kota Kupang yang data aktualnya tinggi tetapi hasil prediksinya sangat tinggi.

Berdasarkan nilai pada *confusion matrix* di atas dapat dihitung nilai akurasi, *error rate* dan *AUC*. Perhitungan nilai akurasi klasifikasi dan *error rate* sebagai berikut:

1. $Akurasi = \frac{\sum Data\ uji\ benar\ klasifikasi}{\sum Data\ uji}\ x\ 100\% = \frac{124}{129}\ x\ 100\% = 96.12\%$
2. $Error\ rate = \frac{\sum Data\ uji\ salah\ klasifikasi}{\sum Data\ uji}\ x\ 100\% = \frac{5}{129}\ x\ 100\% = 3.88\%$
3. Rata-rata nilai *AUC* empat kelas prediksi, dimana M = Rendah, N = Sedang, O = Tinggi dan P = Sangat tinggi dapat dilihat pada tabel berikut ini.

Tabel 4 Nilai *AUC* metode *k-nearest neighbor*

Prediksi	***AUC***
Rendah	0.917
Sedang	0.970
Tinggi	0.964
Sangat tinggi	0.996
Rata-rata	**0.9618**

Berdasarkan Tabel 4 di atas diperoleh nilai rata-rata *AUC* klasifikasi dengan menggunakan metode *k-nearest neighbor* sebesar 0.9618.

Perhitungan di atas menunjukkan bahwa klasifikasi dengan metode *k-nearest neighbor* memiliki nilai akurasi sebesar 96.12%, *error rate* sebesar 3.88% dan rata-rata *AUC* sebesar 0.9618.

B. Klasifikasi Metode *Random Forest*

Pengklasifikasian Indeks Pembangunan Manusia dengan menggunakan metode *random forest* adalah sebagai berikut:

1. Penentuan Jumlah *mtry* dan *ntree*

Klasifikasi dengan menggunakan metode *random forest* terlebih dahulu harus menentukan jumlah *mtry* dan *ntree* yang akan digunakan. Penentuan jumlah *mtry* dan *ntree* yang akan digunakan dapat dilihat di bawah ini.

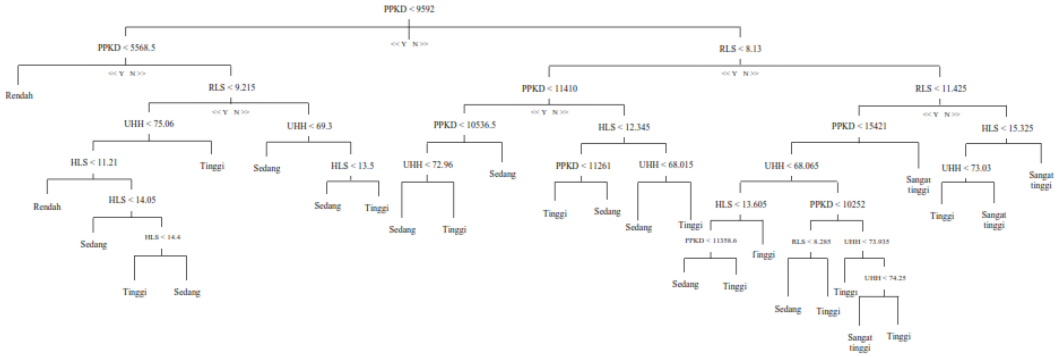
Tabel 5 Nilai akurasi dari *mtry* dan *ntree*

mtry	***ntree***	**Akurasi**
2	50	0.9225
	100	0.9225
4	50	0.9147
	100	0.9147

Berdasarkan Tabel 5 di atas menunjukkan bahwa jumlah *mtry* dan *ntree* yang digunakan dalam klasifikasi menggunakan algoritma *random forest* adalah $mtry = \sqrt{m} = \sqrt{4} = 2$ dan $ntree = 100$. Jumlah *mtry* dan *ntree* ini digunakan karena memiliki nilai akurasi yang lebih tinggi.

2. Pohon Klasifikasi

Pohon klasifikasi pada metode random forest digunakan untuk memprediksi data testing. Pohon klasifikasi dapat dilihat pada gambar di bawah ini.



Gambar 1. Pohon Klasifikasi

Berdasarkan Gambar 1 di atas diperoleh 28 keputusan yang dapat digunakan untuk memprediksi data *testing*.

3. Hasil Prediksi Klasifikasi

Hasil prediksi klasifikasi menggunakan metode *random forest* disajikan dalam tabel *confusion matrix*. Tabel *confusion matrix* dapat dilihat di bawah ini.

Tabel 6 *Confusion matrix* metode *random forest*

Prediksi	**Aktual**			
	Rendah	Sedang	Tinggi	Sangat tinggi
Rendah	6	1	0	0
Sedang	0	60	5	0
Tinggi	0	1	43	0
Sangat tinggi	0	0	3	10

Tabel 6 di atas menunjukkan bahwa dari data aktualnya rendah dan prediksinya rendah sebanyak 6 data, data aktualnya sedang dan prediksinya rendah sebanyak 1 data, data aktualnya sedang dan prediksinya sedang sebanyak 60 data, data aktualnya sedang dan prediksinya tinggi sebanyak 1 data, data aktualnya tinggi dan prediksinya sedang sebanyak 5 data, data aktualnya tinggi dan prediksinya tinggi sebanyak 43 data, data aktualnya tinggi dan prediksinya sangat tinggi sebanyak 3 data dan data aktualnya sangat tinggi dan prediksinya sangat tinggi sebanyak 10 data.

Klasifikasi dengan menggunakan metode *random forest* memberikan kesalahan klasifikasi sebanyak 10 diantaranya ada Kabupaten Sampang yang data aktualnya sedang tetapi hasil prediksinya rendah, Kabupaten Lampung Timur yang data aktualnya sedang tetapi hasil prediksinya tinggi, Kabupaten Tapanuli Selatan, Kabupaten Samosir, Kabupaten Pesisir Selatan, Kabupaten Bungo dan Kabupaten Gowa yang data aktualnya tinggi tetapi hasil prediksinya sedang dan Kota Pangkal Pinang, Kota Tanjung Pinang dan Kota Kupang yang data aktualnya tinggi tetapi hasil prediksinya sangat tinggi.

Berdasarkan nilai pada *confusion matrix* di atas dapat dihitung nilai akurasi, *error rate* dan *AUC*. Perhitungan nilai akurasi klasifikasi dan *error rate* sebagai berikut:

1. $Akurasi = \frac{\sum Data\ uji\ benar\ klasifikasi}{\sum Total\ data\ uji}\ x\ 100\% = \frac{119}{129}\ x\ 100\% = 92.25\%$
2. $Error\ rate \frac{\sum Data\ uji\ salah\ klasifikasi}{\sum Data\ uji}\ x\ 100\% = \frac{10}{129}\ x\ 100\% = 7.75\%$
3. Rata-rata nilai *AUC* empat kelas prediksi, dimana M = Rendah, N = Sedang, O = Tinggi dan P = Sangat tinggi dapat dilihat pada tabel berikut ini.

Tabel 7 Nilai *AUC* metode *random forest*

Prediksi	***AUC***
Rendah	0.996

Sedang	0.947
Tinggi	0.915
Sangat tinggi	0.987
Rata-rata	**0.9538**

Berdasarkan Tabel 7 di atas diperoleh nilai rata-rata *AUC* klasifikasi dengan menggunakan metode *random forest* sebesar 0.9538.

Hasil perhitungan di atas menunjukkan bahwa metode *random forest* memperoleh nilai akurasi klasifikasi sebesar 92.25%, *error rate* sebesar 7.75% dan rata-rata *AUC* sebesar 0.9538.

C. Perbandingan Metode *K-Nearest Neighbor* dan *Random Forest*

Nilai akurasi, error rate dan rata-rata AUC dari metode metode *k-nearest neighbor* dan *random forest* akan dibandingkan untuk mengetahui metode terbaik dalam melakukan klasifikasi Indeks Pembangunan Manusia di Kabupaten/Kota seluruh Indonesia. Nilai akurasi kedua metode dapat di lihat pada tabel di bawah ini:

Tabel 8 Nilai akurasi, *error rate* dan rata-rata *AUC* metode klasifikasi

Metode	**Akurasi**	***Error Rate***	***Rata-rata AUC***
K-Nearest Neighbor	96.12%	3.88%	0.9618
Random Forest	92.25%	7.75%	0.9538

Berdasarkan Tabel 8 diketahui bahwa klasifikasi dengan menggunakan metode *k-nearest neighbor* memperoleh nilai akurasi sebesar 96.12% dan rata-rata AUC sebesar 0.9618 dimana lebih besar dibandingkan dengan metode *random forest* yang memiliki nilai akurasi sebesar 92.25% dan rata-rata AUC sebesar 0.9538 serta metode *k-nearest neighbor* memperoleh nilai error rate sebesar 3.88% dimana lebih kecil dibandingkan dengan dengan metode *random forest* yang memiliki nilai error rate sebesar 7.75%. Maka metode *k-nearest neighbor* menjadi metode yang terbaik untuk digunakan dalam klasifikasi Indeks Pembangunan Manusia di Kabupaten/Kota seluruh Indonesia.

4. Kesimpulan

Berdasarkan hasil dari analisis dan pembahasan yang telah dilakukan, maka dapat diambil beberapa kesimpulan antara lain:

1. Indeks Pembangunan Manusia pada tahun 2021 memiliki rata-rata sebesar 69.927, standar deviasi sebesar 6.497, minimum sebesar 32.84, median sebesar 69.61 dan maksimum sebesar 87.17. Pada tahun 2021 ada sebanyak 22 Kabupaten/Kota memiliki IPM rendah, sebanyak 250 Kabupaten/Kota memiliki IPM sedang, sebanyak 204 Kabupaten/Kota memiliki IPM tinggi dan sebanyak 38 Kabupaten/Kota memiliki IPM sangat tinggi.
2. Klasifikasi Indeks Pembangunan Manusia menggunakan metode *k-nearest neighbor* menghasilkan tingkat akurasi klasifikasi sebesar 96.12%, *error rate* sebesar 3.88% dan rata-rata *AUC* sebesar 0.9618. Hasil klasifikasi IPM dengan menggunakan metode *k-nearest neighbor* ini diperoleh sebanyak 5 Kabupaten/Kota yang memiliki IPM rendah, sebanyak 61 Kabupaten/Kota yang memiliki IPM sedang, sebanyak 48 Kabupaten/Kota yang memiliki IPM tinggi dan sebanyak 10 Kabupaten/Kota yang memiliki IPM sangat tinggi serta terdapat 5 Kabupaten/Kota yang memiliki kesalahan klasifikasi dari 129 Kabupaten/Kota..
3. Klasifikasi Indeks Pembangunan Manusia menggunakan metode *random forest* menghasilkan tingkat akurasi klasifikasi sebesar 92.25%, *error rate* sebesar 7.75% dan rata-rata *AUC* sebesar 0.9538. Hasil klasifikasi IPM dengan menggunakan metode *random forest* ini diperoleh sebanyak 6 Kabupaten/Kota yang memiliki IPM rendah, sebanyak 60

Kabupaten/Kota yang memiliki IPM sedang, sebanyak 43 Kabupaten/Kota yang memiliki IPM tinggi dan sebanyak 10 Kabupaten/Kota yang memiliki IPM sangat tinggi serta terdapat 10 Kabupaten/Kota yang memiliki kesalahan klasifikasi dari 129 Kabupaten/Kota.

4. Metode *k-nearest neighbor* adalah metode terbaik dalam melakukan klasifikasi Indeks Pembangunan Manusia di Kabupaten/Kota seluruh Indonesia pada tahun 2021, karena memperoleh nilai akurasi sebesar 96.12% dan rata-rata *AUC* sebesar 0.9618 yang lebih besar dibandingkan dengan metode *random forest* yaitu sebesar 92.25 dan 0.9538 serta metode *k-nearest neighbor* memperoleh nilai *error rate* sebesar 3.88% yang lebih kecil dibandingkan dengan metode *random forest* sebesar 7.75%.

Ucapan Terima Kasih

Penulisan jurnal ini tidak terlepas dari bimbingan dan dukungan dari berbagai pihak, oleh karena itu penulis mengucapkan terima kasih yang setulus-tulusnya kepada:

1. Bapak Yudi Setyawan, MS., M.Sc selaku Ketua Jurusan Statstika Institut Sains & Teknologi AKPRIND Yogyakarta.
2. Ibu Kris Suryowati, S.Si, M.Si selaku Dosen Pembimbing I yang dengan sabar memberikan bimbingan serta saran demi kelancaran penyusunan ini.
3. Ibu Maria Titah Jatipaningrum, S.Si, M.Sc selaku Dosen Pembimbing II yang dengan sabar memberikan bimbingan serta saran demi kelancaran penyusunan ini.
4. Kedua orang tua yang telah memberikan do'a, kepercayaan, dukungan dan fasilitas yang tidak terhingga.

Daftar Pustaka

[1] BPS. (2015). Indeks Pembangunan Manusia 2014. Jakarta: Badan Pusat Statistik.

[2] BPS. (2022). Indeks Pembangunan Manusia 2021. Jakarta: Badan Pusat Statistik.

[3] Darsyah, M. Y. (2017). Klasifikasi Indeks Pembangunan Manusia (IPM) dengan Pendekatan *K-Nearset Neighbor* (K-NN). *Seminar Nasional Pendidikan, Sains dan Teknologi* (pp. 29-35). Semarang: Universitas Muhammadiyah Semarang.

[4] Fauzi, F. (2017). K-*Nearest Neighbor* (K-NN) dan *Support Vector Machine* (SVM) untuk Klasifikasi Indeks Pembangunan Manusia Provinsi Jawa Tengah. *Jurnal MIPA, 40(2)*, 118-124.

[5] Haristu, R. A. (2019). Penerapan Metode *Random Forest* untuk Prediksi *Win Ratio* Pemain *Player Unknown Battleground.* Yogyakarta: Universitas Sanata Dharma.

[6] Mauludiyah, K. (2020). Klasifikasi Indeks Pembangunan Manusia Kabupaten/Kota di Indonesia Menggunakan Metode *Random Forest.* Semarang: Universitas Muhammadiyah Semarang.

[7] Putra, M. I. (2019). Sistem Rekomendasi Kelayakan Kredit Menggunakan Metode *Random Forest* pada BRI Kantor Cabang Pelaihari. Surabaya: Universitas Islam Negeri Sunan Ampel.

[8] Rachmi, A. N. (2020). Implementasi Metode *Random Forest* dan *XGBoost* pada Klasifikasi *Customer Churn.* Yogyakarta: Universitas Islam Indonesia.

[9] Sundara, V. Y., Warti, R., & Mardia, A. (2019). Simulasi Metode *Resampling* dan Pendugaan Data Hilang Terbaik. *Jurnal Riset dan Aplikasi Matematika, 3(2)*, 101-108.

[10] Yunita, D. (2017). Perbandingan Algoritma *K-Nearest Neighbor* dan *Decision Tree* untuk Penentuan Risiko Kredit Kepemilikan Mobil. *Jurnal Informatika Universitas Pamulang, 2(2)*, 103-107.