JAMBURA JOURNAL OF PROBABILITY AND STATISTICS
Volume 4 Nomor 2, November 2023
Journal Homepage: http://ejurnal.ung.ac.id/index.php/jps
DOI: https://doi.org/10.37905/jjps.v4i2.18460

IMPLEMENTASI ALGORITMA RANDOM FOREST DENGAN FORWARD SELECTION UNTUK KLASIFIKASI INDEKS PEMBANGUNAN MANUSIA

Tiara Posangi[1], Lailany Yahya[2], Djihad Wungguli[3]
[1] Program Studi Statistika, Jurusan Matematika, Fakultas MIPA, Universitas Negeri Gorontalo
[2,3] Program Studi Matematika, Jurusan Matematika, Fakultas MIPA, Universitas Negeri Gorontalo

e-mail: djihad@ung.ac.id

Abstrak
Pembangunan pada hakikatnya merupakan proses perubahan secara terus-menerus yang dilakukan dalam mencapai suatu kondisi hidup yang lebih baik. Sehingga tolak ukur dari keberhasilan suatu pembangunan dilihat pada pembangunan manusianya. Tiga dimensi dasar pembentuk pembangunan manusia ialah umur panjang dan sehat, pengetahuan, dan kehidupan yang layak. Indikator-indikator yang mempresentasikan ketiga dimensi itu terangkum dalam satu nilai tunggal yaitu Indeks Pembangunan Manusia (IPM). Pada tahun 2021 angka IPM di Indonesia sebesar 72.29 artinya tergolong tinggi. Namun karena letak geografis daerah di indonesia yang beragam maka hal tersebut turut mempengaruhi angka IPM pada masing-masing daerah di indonesia sehingga pada penelitian ini menggunakan Algoritma Random Forest untuk mendapatkan hasil akurasi dari klasifikasi IPM dan menggunakan Forward Selection untuk menentukan fitur yang berpengaruh dalam pengklasifikasian. Adapun hasil penelitian menunjukan bahwa fitur yang berpengaruh dalam pengklasifikasian adalah pengeluaran perkapita, harapan lama sekolah, angka harapan hidup, dan rata-rata lama sekolah, dan mendapatkan hasil akurasi akhir sebesar 80%.
Kata Kunci: Indeks Pembangunan Manusia, Random Forest, Forward Selection

Abstract
Development is essentially a process of continuous change carried out to achieve better living condition. So that the benchmark for the success of a development is seen in its human development. 3 The basic dimensions that form human development are long and healthy life, knowledge, and a decent life. The indicators that represent the three dimensions are summarized in a single value, namely the Human Development Index (IPM). In 2021 the HDI figure in Indonesia is 72.29, which means it is high. However, due to the diverse geographical location of regions in Indonesia, this also influences the HDI rate in each region in Indonesia, so this study uses the Random Forest Algorithm to obtain accurate results from the HDI classification and uses Forward Selection to determine features that influence the classification. The results of the study show that the features that influence the classification are per capita spending, expected length of schooling, life expectancy, and average length of schooling, and get a final accuracy of 80%.
Keywords: *Human Development Index, Random Forest, Forward Selection*

1. PENDAHULUAN

Hakekatnya, proses perubahan untuk mencapai kehidupan yang lebih baik secara terus-menerus disebut dengan Pembangunan (Dewi, Yusuf and Iyan, 2017). Tolak ukur keberhasilan suatu pembangunan tidak hanya dilihat pada pertumbuhan ekonominya saja, namun dilihat pada berhasil tidaknya pembangunan manusia. Pembangunan manusia dibentuk dengan ukuran kinerja pembangunan secara keseluruhan dengan menggunakan ukuran tiga dimensi yaitu Umur panjang dan sehat, kehidupan yang layak, serta pengetahuan. Ketiga ukuran tersebut terangkum dalam satu nilai tunggal, yakni Indeks Pembangunan Manusia (IPM) (Mauludiyah, 2020). IPM merupakan indikator penting untuk

Received: 2023-01-17 | Accepted: 2023-11-11 | Available Online: 2023-11-30

mengukur keberhasilan dalam upaya membangun kualitas hidup,secara global standar ukuran tersebut digunakan sebagai salah satu penentu negara atau daerah tersebut dikatakan maju atau berkembang (Fajri, 2021). Sejalan juga dengan pendapat Durand (Durand and Valla, 2008) yang menjelaskan bahwa IPM merupakan keadaan dimana penduduk memiliki kemampuan untuk mengakses Pendidikan, kesehatan dll.

Menurut Badan Pusat Statistik (Badan Pusat Statistik, 2014), terdapat 4 kategori dari Indeks Pembangunan Manusia (IPM) diantaranya kategori rendah dengan nilai IPM < 60, kategori sedang dengan nilai IPM 60 ≤ IPM < 70, kategori tinggi dengan nilai 70 ≤ IPM < 80, dan kategori sangat tinggi dengan nilai IPM ≥ 80. Di Indonesia, nilai IPM pada tahun 2021 sebesar 72,29 yang artinya termasuk dalam kategori tinggi. Namun akibat letak geografis daerah di Indonesia yang beragam, maka hal tersebut turut mempengaruhi angka IPM di masing-masing daerah di Indonesia. Oleh karena itu, digunakan metode klasifikasi untuk melihat dan memprediksi angka IPM ke dalam empat kategori tersebut pada tiap daerah provinsi di Indonesia.

Klasifikasi menjadi salah satu teknik yang ada dalam data mining untuk melakukan pengelompokan data yang sesuai dengan keterikatan data terhadap data sampel (Utomo and Mesran, 2020). Penelitian terkait klasifikasi sudah pernah dilakukan oleh Isran dkk. (Hasan, Resmawan and Ibrahim, 2022) yang mengklasifikasi lama studi mahasiswa dengan membandingkan hasil dari dua algoritma klasifikasi yaitu Algoritma K-Nearest Neighbor (KNN) dan Random Forest dan didapatkan tingkat akurasi Random Forest lebih lebih besar dari algoritma KNN yakni untuk Random Forest sebesar 100% dan KNN sebesar 86,67% sehingga Random Forest dapat dikatakan bekerja lebih baik dibandingkan dengan algoritma KNN. Penelitian lainnya yaitu oleh (Zailani and Hanun, 2020) yaitu menerapkan algoritma klasifikasi Random Forest dalam menentukan kelayakan pemberian kredit, untuk menganalisis kredit yang bermasalah dan debitur tidak bermasalah dan mendapatkan hasil tingkat akurasi 87,88%. Dimana, algoritma Random Forest mampu meningkatkan akurasi dalam menganalisis kelayakan kredit yang diajukan calon debitur melalui pohon keputusan. Dari beberapa penelitian sebelumnya dapat dikatakan bahwa klasifikasi dengan metode Random Forest memiliki tingkat akurasi yang lebih baik dibanding metode klasifikasi lainnya. Scornet dkk. (Scornet, Biau and Vert, 2015) juga menjelaskan bahwa Random Forest merupakan salah satu metode yang ada dalam klasifikasi yang tujuannya untuk membangun sejumlah pohon keputusan secara acak. Random Forest merupakan pengembangan lanjutan dari Classification and Regression Tree (CART) yang mulanya berasal dari metode Random feature selection dan juga Bagging (bootstrap aggregating) (Sandag, 2020)

Dengan demikian, penelitian ini akan menerapkan Algoritma Random Forest dalam melakukan klasifikasi indeks pembangunan manusia pada tiap daerah provinsi di Indonesia. Serta untuk menghilangkan tingkat kompleksitas dari suatu algoritma klasifikasi maka digunakan Forward Selection untuk menyeleksi fitur-fitur yang tidak berpengaruh sehingga dapat mengoptimalkan akurasi dari algoritma klasifikasi tersebut. Seperti pada penelitian sebelumnya oleh Zeniarja dkk. (Zeniarja, Widia and Sani, 2020) yaitu menggunakan Forward Selection untuk mengurangi fitur/atribut yang tidak berpengaruh dan terbukti dapat meningkatkan nilai akurasi, dimana hasil pengujian tanpa Forward Selection mempunyai tingkat akurasi sebesar 83,33% sedangkan kinerja dengan menggunakan Forward Selection mengalami peningkatan sebesar 2, 67% menjadi 86, 00%.

2. METODE PENELITIAN

Data yang digunakan dalam penelitian ini merupakan data sekunder, dimana data tersebut diperoleh dari website Badan Pusat Statistik (BPS) Republik Indonesia. Adapun variabel-variabel yang digunakan terdiri dari variabel dependen (Y) yaitu Indeks Pembangunan Manusia, dan 8 variabel independen yaitu (X1) Pengeluaran per Kapita, (X2) Harapan Lama Sekolah, (X3) Angka Harapan Hidup, (X4) Rata-rata Lama Sekolah, (X5) Jumlah Penduduk Miskin, (X6) Tingkat Pengangguran Terbuka, (X7) Tingkat Partisipasi Angkatan Kerja, dan (X8) Produk Domestik Regional Bruto. Teknik pengambilan sampel dalam penelitian ini menggunakan sampling jenuh dimana seluruh anggota populasi dijadikan sebagai sampel. Selanjutnya akan dilakukan pengklasifikasian menggunakan algoritma Random Forest dengan Forward Selection dan menggunakan bantuan software Python dan R Studio. Berikut merupakan langkah-langkah dalam melakukan penelitian:

1. Melakukan pengumpulan data yaitu data Indeks Pembangunan Manusia pada 34 Provinsi di Indonesia tahun 2021.
2. Menentukan atribut terbaik dengan menggunakan seleksi fitur Forward Selection.
3. Membagi data ke dalam dua bagian, yakni data training dan data testing.
4. Model Klasifikasi Random Forest menggunakan data training.
5. Mengukur tingkat akurasi menggunakan data testing.

3. HASIL DAN PEMBAHASAN

3.1 Forward Selection

Sebelum melakukan klasifikasi, maka terlebih dahulu dilakukan seleksi fitur hal ini berguna untuk mengurangi tingkat kompleksitas dari algoritma klasifikasi serta dapat mengetahui tingkat pengaruh dari suatu fitur (Supriyanti and Puspitasari, 2018). Cara kerja metode ini dimulai dari nol peubah (*empty model*), setelah itu satu-persatu peubah dimasukkan hingga memenuhi kategori tertentu (Nugroho and Wibowo, 2017). Fitur yang dianggap berpengaruh dilakukan dengan membandingkan hubungan ekivalensi yang dihasilkan oleh set fitur kemudian menghapus fitur dengan memberikan kualitas analisis yang masih lebih baik. Jadi pada penelitian ini didapatkan model akhir dengan fitur terbaik:

$$\hat{y} = \beta_0 + \beta_1 X_a + \beta_2 X_b + \beta_3 X_c + \beta_4 X_d \quad (1)$$

dengan,

Xa : Fitur Pengeluaran per kapita
Xb : Fitur Harapan lama sekolah
Xc : Fitur Angka harapan hidup
Xd : Fitur Rata-rata lama sekolah

3.2 *Preprocessing* Data

Preprocessing sangat dibutuhkan dalam proses kinerja algoritma klasifikasi. Karena tidak semua data yang digunakan dalam proses mining memiliki kondisi yang ideal untuk diproses. Oleh sebab itu pada penelitian ini hanya variabel Y akan dikonversi karena variabel Y memiliki data awal yang berbentuk rasio dan hal tersebut dapat menghambat proses mining sehingga akan diubah ke bentuk kategorik sehingga siap untuk di mining.

Setelah melewati proses *preprocessing* selanjutnya data akan dibagi, pembagian data bertujuan untuk mendapatkan data *training* dan data *testing* (Sasongko, 2016). Pada penelitian ini data *training* digunakan untuk direpresentasikan dalam pembentukan model klasifikasi, selanjutnya data *testing* digunakan dalam memperkirakan akurasi dari klasifikasi. Pembagian data ini dilakukan dengan pemilihan secara *random* dengan nilai 70% untuk data

training dan 30% data *testing*. Sehingga didapatkan data *training* sebanyak 24 data dan data *testing* sebanyak 10 data.

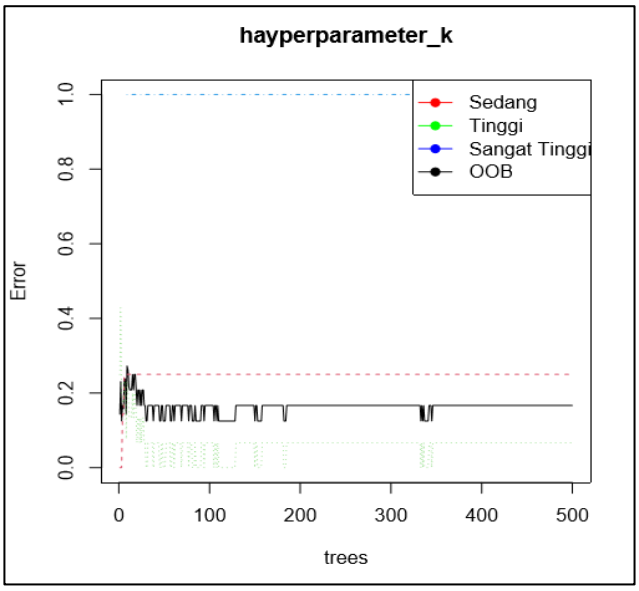
3.3 Klasifikasi dengan Random Forest

Random forest merupakan suatu metode klasifikasi yang terdiri dari sejumlah pohon keputusan di berbagai subset dari dataset dan mengambil rata-rata dalam meningkatkan akurasi prediksi dari dataset tersebut. *Random forest* memprediksi berdasarkan suara terbanyak dari setiap pohon.

Pembentukan Model

Setelah pembagian data dan didapatkan data *training*, selanjutnya sepertiga sampel dari data *training set* digunakan sebagai data *out of bag* (OOB) . data OOB digunakan untuk menghitung error dan menentukan *variable importance* yaitu variabel yang digunakan untuk pemisahan (split) terbaik ditentukan secara acak. Namun, sebelum pembentukan model *Random Forest* maka terlebih dahulu menentukan hyperparameter yang akan digunakan dalam pembentukan model.

Adapun hyperparameter yang diperlukan adalah hyperparameter k, dimana menunjukkan jumlah *tree* yang akan digunakan untuk pembentukan model. Ditampilkan pada Gambar 1.



Gambar 1. Hyperparameter K optimal

Berdasarkan Gambar 1 dengan bantuan RStudio terlihat tingkat misklasifikasi dari ke empat kategori relatif semakin stabil pada pohon dengan jumlah 400 hingga 500. Untuk itu dapat ditetapkan bahwa k optimal dalam penelitian ini adalah 500, dalam tahapan selanjutnya, penggunaan jumlah pohon yang dibentuk adalah default sebesar 500 pohon.

Adapun hasil model akhir yang dihasilkan dengan menggunakan data *training* dengan jumlah k 500 disajikan pada Gambar 2.

```
 randomForest(formula = IPM_Y ~ Pengeluaran_Per_Kapita_X1 + Harapan_Lama_Sekol
ah_X2 +      Angka_Harapan_Hidup_X3 + Rata_Lama_Sekolah_X4, data = train,
mtry = 2, importance = T)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 12.5%
Confusion matrix:
              Sedang Tinggi Sangat Tinggi class.error
Sedang             6      2             0        0.25
Tinggi             0     15             0        0.00
Sangat Tinggi      0      1             0        1.00
>
>
```

Gambar 2. Model Random Forest

Pada Gambar 2 dapat dilihat bahwa tipe *Random Forest* yang terbentuk adalah klasifikasi. Dengan jumlah pohon yang dibentuk sebanyak 500 dan untuk variabel yang digunakan dalam setiap iterasi berjumlah 2 dengan tingkat OOB yang dihasilkan adalah sebesar 12.5%.

Pengujian Akurasi Model

Setelah didapatkan model dengan menggunakan data *training* maka selanjutnya melakukan pengujian dengan menggunakan data *testing* untuk melihat akurasi dari model yang didapat. Dapat dilihat pada Tabel 1.

Tabel 1 Hasil Prediksi Klasifikasi Dengan Data *Testing*

Provinsi	Y	X1	X2	X3	X4	Prediksi Klasifikasi
Sumatera Selatan	Tinggi	10662	12,54	69,98	8,30	Tinggi
Lampung	Sedang	10038	12,73	70,73	8,08	Tinggi
Kep. Bangka Belitung	Tinggi	12819	12,17	70,73	8,08	Tinggi
Dki Jakarta	Sangat Tinggi	18520	13,07	73,01	11,17	Tinggi
Jawa Tengah	Tinggi	11034	12,77	74,47	7,75	Tinggi
Banten	Tinggi	12033	13,02	70,02	8,93	Tinggi
Bali	Tinggi	13820	13,40	72,24	9,06	Tinggi
Kalimantan Selatan	Tinggi	12143	12,81	68,83	8,34	Tinggi
Maluku	Sedang	8770	13,97	66,09	10,03	Sedang
Maluku Utara	Sedang	8140	13,68	68,45	9,09	Sedang

Tabel 1 merupakan hasil dari prediksi pada data *testing*. Terlihat bahwa hasil prediksi yang dihasilkan yaitu sejumlah sepuluh data yang terdiri dari delapan data terklasifikasi kedalam kelas tinggi dan dua data terklasifikasi ke dalam kelas sedang

Setelah mendapatkan hasil prediksi dari seluruh data *testing* dengan menggunakan model *Random Forest*, maka selanjutnya dapat dilakukan evaluasi hasil klasifikasi algoritma *random forest* dengan menggunakan confusion matrix. Hal ini sesuai dengan penelitian Septiana (Septiana, Susanto and Tukiyat, 2021) yang menjelaskan bahwa confusion matrix adalah tabel yang digunakan untuk mengevaluasi hasil yang diperoleh dari proses klasifikasi. Adapun hasil confusion matrix ditampilkan pada Tabel 2.

Tabel 2 Confusion Matrix

Prediksi Klasifikasi Indeks Pembangunan Manusia	Klasifikasi			Total
	Sedang	Tinggi	Sangat Tinggi	
Sedang	2	0	0	2
Tinggi	1*	6	1*	8
Sangat Tinggi	0	0	0	0
Total	3	6	1	10

*Miss Klasifikasi

dengan akurasi sebagai berikut :

$$akurasi = \frac{8}{10} x100 = 80\%$$

Berdasarkan Tabel 2 dapat diketahui bahwa Indeks Pembangunan Manusia pada setiap provinsi pada tahun 2021 dengan menggunakan algoritma *random forest* diperoleh hasil yakni 3 provinsi dengan kelas sedang namun hanya 2 provinsi yang dapat diklasifikasikan dengan benar sedangkan 1 provinsi lainnya terjadi misklasifikasi. Untuk 6 provinsi yang termasuk dalam kelas tinggi dapat diklasifikasikan dengan benar sedangkan 2 provinsi lainya tergolong mis klasifikasi dimana 1 digolongkan kelas sedang dan 1 digolongkan kelas sangat tinggi. Dan untuk kelas tinggi tidak dapat diklasifikasikan dengan benar. Oleh karena itu dengan terjadinya misklasifikasi maka didapatkan tingkat akurasi sebesar 80%.

4. KESIMPULAN

Berdasarkan hasil analisis dan pembahasan yang telah dilakukan maka dapat disimpulkan bahwa proses seleksi fitur dengan menggunakan *forward selection* didapatkan empat fitur terbaik dari delapan fitur yang tersedia. Adapun yang menjadi fitur terbaik yaitu pengeluaran perkapita, harapan lama sekolah, angka harapan hidup, dan rata-rata lama Sekolah. Dan untuk klasifikasi indeks pembangunan manusia dengan menggunakan Algoritma *Random Forest* didapatkan hasil akurasi sebesar 80%.

DAFTAR PUSTAKA

Badan Pusat Statistik (2014) *Indeks Pembangunan Manusia 2013*, *Badan Pusat Statistik*. Available at: https://www.bps.go.id/id/publication/2014/10/13/309e1d97d0691420ef1a0121/indeks-pembangunan-manusia-2013.html (Accessed: 15 January 2023).

Dewi, N., Yusuf, Y. and Iyan, R. Y. (2017) 'Pengaruh Kemiskinan Dan Pertumbuhan Ekonomi Terhadap Indeks Pembangunan Manusia Di Provinsi Riau', *Jurnal Online Mahasiswa (JOM) Bidang Ilmu Ekonomi*, 4(1).

Durand, F. and Valla, D. (2008) 'Assessment of Prognosis of Cirrhosis', *Seminars in Liver Disease*, 28(1), pp. 110–122. doi: 10.1055/s-2008-1040325.

Fajri, R. H. (2021) 'Analisis Faktor-Faktor Yang Mempengaruhi Indeks Pembangunan Manusia Di Provinsi Riau', *ECOUNTBIS: Economics, Accounting and Business Journal*, 1(1). Available at: https://jom.umri.ac.id/index.php/ecountbis/article/view/257.

Hasan, I. K., Resmawan, R. and Ibrahim, J. (2022) 'Perbandingan K-Nearest Neighbor dan

Random Forest dengan Seleksi Fitur Information Gain untuk Klasifikasi Lama Studi Mahasiswa', *Indonesian Journal of Applied Statistics*, 5(1), p. 58. doi: 10.13057/ijas.v5i1.58056.

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

Nugroho, M. F. and Wibowo, S. (2017) 'Fitur Seleksi Forward Selection Untuk Menetukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma Naive Bayes', *Jurnal Informatika Upgris*, 3(1). doi: 10.26877/jiu.v3i1.1669.

Sandag, G. A. (2020) 'Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest', *CogITo Smart Journal*, 6(2), pp. 167–178. doi: 10.31154/cogito.v6i2.270.167-178.

Sasongko, T. B. (2016) 'Komparasi dan Analisis Kinerja Model Algoritma SVM dan PSO-SVM (Studi Kasus Klasifikasi Jalur Minat SMA)', *Jurnal Teknik Informatika dan Sistem Informasi*, 2(2).

Scornet, E., Biau, G. and Vert, J.-P. (2015) 'Consistency of random forests', *The Annals of Statistics*, 43(4). doi: 10.1214/15-AOS1321.

Septiana, R. D., Susanto, A. B. and Tukiyat, T. (2021) 'Analisis Sentimen Vaksinasi Covid-19 Pada Twitter Menggunakan Naive Bayes Classifier Dengan Feature Selection Chi-Squared Statistic dan Particle Swarm Optimization', *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, 5(1), pp. 49–56. doi: 10.47970/siskom-kb.v5i1.228.

Supriyanti, W. and Puspitasari, N. (2018) 'Implementasi Teknik Seleksi Fitur Forward Selection Pada Algoritma Klasifikasi Data Mining untuk Prediksi Masa Studi Mahasiswa Politeknik Indonusa Surakarta', *Jurnal INFORMA*, 4(2).

Utomo, D. P. and Mesran, M.(2020) 'Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Dataset Penyakit Jantung', *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 4(2), p. 437. doi: 10.30865/mib.v4i2.2080.

Zailani, A. U. and Hanun, N. L. (2020) 'PENERAPAN ALGORITMA KLASIFIKASI RANDOM FOREST UNTUK PENENTUAN KELAYAKAN PEMBERIAN KREDIT DI KOPERASI MITRA SEJAHTERA', *Infotech: Journal of Technology Information*, 6(1), pp. 7–14. doi: 10.37365/jti.v6i1.61.

Zeniarja, J., Widia, K. and Sani, R. R. (2020) 'Penerapan Algoritma Naive Bayes dan Forward Selection dalam Pengklasifikasian Status Gizi Stunting pada Puskesmas Pandanaran Semarang', *JOINS (Journal of Information System)*, 5(1), pp. 1–9. doi: 10.33633/joins.v5i1.2745.