Published online on: https://journal.uii.ac.id/jurnalsnati/
DOI: 10.20885/snati.v4.i1.4
Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi
Vol. 4 No. 1 (2025) 26 - 33 ISSN Media Electronic: 2807-5935

Perbandingan Algoritma Machine Learning dalam Klasifikasi Kab/Kota di Indonesia Menurut Indeks Pembangunan Manusia (IPM) Tahun 2021

Comparison of Machine Learning Algorithms in Classifying Districts/Cities in Indonesia According to the Human Development Index (HDI) in 2021

Ni Kadek Ayu Purnami Sari Dewi[1], Arie Wahyu Wijayanto[2], Joko Ade Nursiyono[3]
[1,2]Program Studi Statistika, Politeknik Statistika STIS, Jakarta Timur, Indonesia
[3]Badan Pusat Statistika Provinsi Jawa Timur, Jawa Timur, Indonesia
[1]211911183@stis.ac.id, [2]ariewahyu@stis.ac.id, [3]joko.ade@bps.go.id

Abstract

The human development index (HDI) is one of the measuring tools for achieving the quality of life of a region or even a country, including Indonesia. There are 3 basic components of the HDI, namely the dimensions of health, knowledge, and decent living. Development in Indonesia is uneven as indicated by the Human Development Index (HDI) of districts/cities in 2021 which varies greatly. The purpose of this study is to compare several machine learning algorithms to classify districts/cities in Indonesia according to the Human Development Index (HDI) in 2021. There are six machine learning algorithms used in this study, namely Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearset Neighbor (K-NN), Random Forest, Decision Tree, and Naive Bayes. The k-Fold Cross Validation method is applied to form the training set and testing set, with 10 folds and 1 repetition. The results of the study showed that the classification results of the SVM algorithm using the Radial Basis Function (RBF) kernel parameters with sigma = 0.4864648 and C = 1 were the best among the other five algorithms with an average accuracy of 76.08% and a maximum accuracy of 88.24%.

Keywords: Data Mining; Confusion Matrix; Human Development Index (HDI); Classification; Machine Learning

Abstrak

Indeks pembangunan manusia (IPM) merupakan salah satu alat ukur pencapaian kualitas hidup suatu wilayah bahkan negara, termasuk Indonesia. Terdapat 3 komponen dasar penyusun IPM yaitu dimensi kesehatan, pengetahuan, dan hidup layak. Pembangunan di Indonesia tidak merata sebagaimana ditunjukkan oleh Indeks Pembangunan Manusia (IPM) kabupaten/kota tahun 2021 yang sangat bervariasi. Tujuan penelitian ini adalah membandingkan beberapa algoritma machine learning untuk mengklasifikasikan kab/kota di Indonesia menurut Indeks Pembangunan Manusia (IPM) tahun 2021. Ada enam algoritma machine learning yang digunakan dalam penelitian ini yaitu Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearset Neighbor (K-NN), Random Forest, Decision Tree, dan Naive Bayes. Metode k-Fold Cross Validation diterapkan untuk membentuk training set dan testing set, dengan 10 lipatan dan 1 pengulangan. Hasil penelitian menunjukan bahwa hasil klasifikasi algoritma SVM menggunakan parameter kernel Radial Basis Function (RBF) dengan sigma = 0,4864648 and C = 1 adalah yang paling baik diatara lima algoritma lainnya dengan rata-rata akurasi sebesar 76,08% dan akurasi maksimal 88,24%.

Kata kunci: *Data Mining*; *Confusion Matrix*; Indeks Pembangunan Manusia (IPM); Klasifikasi; *Machine Learning*

1. Pendahuluan

Manusia merupakan kekayaan sejati suatu negara. Sementara pembangunan manusia merupakan tahapan memperluas pilihan manusia itu sendiri. Oleh sebab itu, dalam suatu wilayah atau negara, manusia sebagai target utama pembangunan memerlukan lingkungan yang mampu memberikan manfaat atau fasilitas sekaligus untuk menikmati periode hidup yang panjang, sehat, dan produktif. Oleh karena itu, kebijakan pembangunan daerah yang tidak memerhatikan pembangunan manusia akan tertinggal dibandingkan daerah lain [1].

Dari 17 Tujuan Pembangunan Berkelanjutan (SDGs), terdapat beberapa tujuan yang erat kaitannya dengan pembangunan manusia. Melalui beberapa tujuan tersebut, target pembangunan manusia dikembangkan lebih lanjut. Sebagai ukuran pembangunan manusia dan kualitas kehidupan sebuah negara, United Nations Development Programme (UNDP) dan Badan Pusat Statistik (BPS) menggunakan tiga dimensi dasar. Dimensi tersebut mencakup umur panjang atau hidup

Received: 25-12-2022 | Revised: 30-09-2023 | Accepted: 25-01-2024 | Published: 15-01-2025

sehat, pengetahuan, dan standar hidup yang layak. Umur panjang merupakan ukuran dari dimensi kesehatan, pengetahuan menjadi ukuran dimensi pendidikan, serta standar hidup yang layak sebagai ukuran dimensi ekonomi.

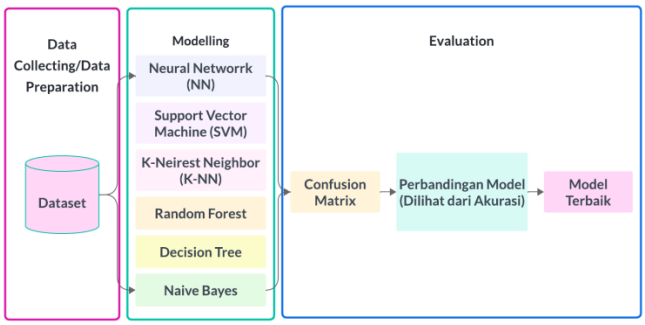
Ketidakmerataan pembangunan manusia di Indonesia sebagaimana ditunjukkan oleh Indeks Pembangunan Manusia (IPM) Kabupaten/Kota tahun 2021 terlihat bervariasi. Secara umum, daerah dengan IPM yang tinggi hanya terpusat di kabupaten/kota besar di Indonesia sebab mempunyai pelayanan kesehatan, pendidikan dan fasilitas yang memadai [2]. Nilai IPM berkisar antara 0 sampai dengan 100. Nilai IPM memberikan gambaran menyeluruh mengenai capaian yang telah dinikmati manusia hasil pembangunan suatu negara atau daerah. Menurut Badan Pusat Statistik (2021), pengkategorian IPM terbagi dalam 4 menurut besarnya, yaitu IPM rendah (< 60), IPM sedang (60 ≤ IPM < 70), IPM tinggi (70 ≤ IPM < 80), dan IPM sangat tinggi (≥ 80). Semakin tinggi nilai IPM suatu negara atau daerah, maka semakin baik kinerja pembangunan manusianya.

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan pembelajaran mesin (machine learning) untuk mengekstraksi dan mengidentifikasi informasi yang berguna dan informasi terkait dari berbagai basis data besar [3]. Data mining memiliki algoritma klasifikasi untuk memisahkan beberapa entitas ke dalam kelas yang berbeda [4], [5]. Tujuan dari penelitian ini adalah membandingkan beberapa algoritma machine learning dalam mengkasifikasikan kabupaten/kota di Indonesia menurut Indeks Pembangunan Manusia (IPM) Tahun 2021 untuk menentukan metode pengklasifikasian terbaik.

Penelitian [6] mengklasifikasikan IPM Provinsi Jawa Tengah dengan 2 metode, yaitu Support Vector Machine (SVM) serta k-Nearest Neighbour (k-NN) menghasilkan nilai akurasi klasifikasi sebesar 91,64% untuk k-NN dan 95,36% untuk metode SVM. Dari hasil penelitian tersebut, diperoleh kesimpulan metode SVM sebagai metode yang tepat untuk klasifikasi IPM. Penelitian [7] mengklasifikasikan set data Multiclass IPM Pulau Sumatera diperoleh hasil bahwa nilai akurasi untuk metode kecerdasan buatan atau Artificial Neural Network (ANN) mencapai 97,4%, atau lebih tinggi dibanding metode SVM yang akurasinya hanya sebesar 53,25%. Oleh karena itu, disimpulkan bahwa metode klasifikasi ANN relatif lebih relevan digunakan untuk mengklasifikasikan IPM di Pulau Sumatera. Klasifikasi IPM dengan K-Nearest Neighbor menghasilkan akurasi yang sangat baik [8]. Selain itu, penelitian dengan tujuan membandingkan performa metode Bagging dan Non-Ensemble Machine Learning untuk klasifikasi IPM kewilayahan [9] menunjukan bahwa algoritma Random Forest merupakan metode terbaik dengan akurasi sebesar 95,14% dan nilai Kappa sebesar 0,8346 dibandingkan dengan metode KNN, Decision Tree C4.5, dan Naïve Bayes. Namun dari sejumlah penelitian terdahulu tersebut, belum ada satupun penelitian tentang topik klasifikasi Indeks Pembangunan Manusia (IPM) di seluruh Kab/Kota Indonesia yang mengkombinasikan metode Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearset Neighbor (K-NN), Random Forest, Decision Tree, dan Naive Bayes sekaligus. Oleh karena itu, penelitian ini berbeda dari peneltian-penelitian terdahulu dan sangat penting untuk dilakukan. Pada penelitian ini, peneliti ingin membandingkan beberapa metode machine learning yaitu Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearset Neighbor (K-NN), Random Forest, Decision Tree, dan Naive Bayes untuk mengklasifikasikan Kab/Kota di Indonesia menurut Indeks Pembangunan Manusia (IPM) Tahun 2021.

2. Metodologi Penelitian

Data yang digunakan dalam penelitian ini yaitu data sekunder yang berasal dari Badan Pusat Statistik (BPS) dan Badan Perencanaan Pembangunan Nasional (Bappenas). Data set tersebut akan dilakukan *preprocessing* dan selanjutnya akan dibagi ke dalam *data training* dan *data testing* dengan memanfaatkan teknik *k-Fold Cross Validation* dengan 10 lipatan dan 1 pengulangan atau dengan kata lain, perbandingan *training data* dan *testing data* yang digunakan adalah 90:10. Selanjutnya dilakukan enam macam algoritma klasifikasi dalam *machine learning* yaitu Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearset Neighbor (K-NN), Random Forest, Decision Tree, dan Naive Bayes. Klasifikasi dilakukan dengan menggunakan *software* R studio versi 4.0.2. Verifikasi hasil akan direpresentasikan oleh *Confusion Matrix* untuk menghitung rata-rata akurasi dan akurasi maksimal yang bisa dihasilkan oleh model secara keseluruhan. Nilai akurasi dari keenam algoritma tersebut akan dibandingkan untuk mendapatkan metode terbaik. Tahapan penelitian ini ditunjukkan pada gambar 1 berikut.



Gambar 1. Metodologi Penelitian
(Sumber: diolah)

2.1 *Data Colecting*

Data penelitian ini merupakan data sekunder yang berasal dari Badan Pusat Statistik (BPS) dan Badan

Perencanaan Pembangunan Nasional (Bappenas) dengan unit observasi sebanyak 514 kabupaten/kota Indonesia dan menggunakan tahun 2021. Adapun beberapa variabel yang digunakan terdiri atas Status Indeks Pembangunan Manusia (Status_IPM), Indeks Pemberdayaan Gender (IDG), Indeks Keparahan Kemiskinan (IKK), Tingkat Pengagguran Terbuka (TPT), Pengeluaran Asli Daerah (PAD), dan PDRB per kapita pada tahun 2021 menurut kabupaten/kota di Indonesia. Dengan struktur data terdapat pada tabel 1 di bawah ini.

Tabel 1. Struktur Data
(Sumber: BPS, Bappenas, 2021)

Atribut	Definisi	Kategori	Skala
Kab_Kota	Kabupaten/Kota di Indonesia	-	Nominal
Status_IPM	Indeks Pembangunan Manusia (IPM)	Rendah = IPM < 70 Tinggi = IPM ≥ 70	Ordinal
IDG	Indeks Pemberdayaan Gender	-	Rasio
IKK	Indeks Keparahan Kemiskinan	-	Rasio
TPT	Tingkat Pengagguran Terbuka	-	Rasio
PAD	Pengeluaran Asli Daerah (Juta Rupiah)	-	Rasio
PDRB_Per kapita	PDRB Per Kapita Atas Dasar Harga Berlaku (Ribu Rupiah)	-	Rasio

2.2 Penyiapan data

Penyiapan data atau yang biasa diistilahkan sebagai *data preparation* adalah suatu proses mempersiapkan data sedemikian rupa sehingga siap untuk diolah lebih lanjut menggunakan metode analitik *data mining* [10]. Persiapan data tersebut meliputi proses *feature selection, cleaning, coding,* atau *transformation* data ke dalam format sesuai kebutuahn. *Data mining* terdiri dari dari objek data yang menjadi sampel atau diistilahkan contoh atau *tuple* serta data *field* atau atribut yang mewakili karakteristik atau fitur dari objek tersebut. Berikutnya, dalam analisis klasifikasi atribut khusus yang menjadi target disebut dengan variabel target.

2.3 *Modelling*

Proses ini melibatkan algoritma klasifikasi dari *data preparation*. Pada langkah ini akan dibentuk model yang dapat untuk mengelompokkan kelas-kelas data. Data tersebut kemudian dipecah menjadi dalam dua jenis himpunan, yakni data latih atau *training data* dan data uji atau *testing data*. *Training data* adalah sehimpun data yang dimanfaatkan dalam pembentukan model, sedangkan *testing data* merupakan sehimpun data yang digunakan untuk menghitung kinerja model yang terbentuk. Kinerja model tersebut kemudian diukur melalui perbandingan antara label data hasil prediksi dengan data berlabel yang sebenarnya [11] Pembentukan *training data* dan *testing data* tersebut dilakukan dengan menggunakan metode *k-Fold Cross Validation* dengan 10 lipatan dan 1 ulangan yaitu bentuk dari *resampling* dengan mengambil beberapa sampel dari keseluruhan observasi dan menjadikannya sebagai *training data* untuk model. Dari 514 kabupaten/kota, sebanyak 90% (462,6 data) digunakan sebagai training data dan sebanyak 10% (51,4 data) digunakan sebagai *testing data*. Setelah dua jenis himpunan data tersebut terbentuk, selanjutnya dilakukan pengklasifikasian dengan menggunakan beberapa algoritma *machine learning,* yaitu dengan ANN, SVM, K-NN, Random Forest, Decision Tree, serta Naive Bayes. Algoritma tersebut merupakan algoritma dalam data mining yang umum digunakan untuk melakukan klasifikasi [12].

1) Artificial Neural Network (ANN)

Artificial Neural Network (ANN) adalah salah satu algoritma klasifikasi pada *machine learning* dengan jenis jaringan saraf tiruan sebagaimana pemodelan sistem kerja jaringan dalam saraf otak manusia untuk menjalankan tugas pengenalan pola klasifikasi [13]. Berikut langkah-langkah algoritma *backpropagation* NN:

1. Menginisialisasi semua bobot dan tentukan fungsi aktivasi serta tingkat kemampuan belajar mesin;
2. Mengaktivasi *network* dengan cara mengimplementasikan masukan dan keluaran;
3. Meng-*update* bobot dengan cara pembobotan terkoreksi ketika kesalahan (*error*) untuk kemudian kembali sesuai dengan arah kembalinya sinyal keluaran;
4. Melakukan pengulangan sampai diperoleh nilai kesalahan (*error)* untuk pembobot yang kurang dari nilai harapan.

2) Support Vector Machine (SVM)

Support Vector Machine (SVM) merupakan salah satu algoritma klasifikasi pada *machine learning* sekaligus metode klasifikasi bersifat diskriminatif dengan memanfaatkan garis *hyperplane* sebagai batas pemisah. SVM bekerja dengan cara menganalisis *hyperplane* yang paling optimal untuk mengklasifikasian kelompok-kelompok secara jelas dan tepat [5], [6]. Untuk dapat memaksimalkan perfoma SVM, maka perlu dilakukan *tuning parameter* dengan menggunakan beberapa parameter yaitu:

1. *Kernel* yaitu fungsi yang digunakan untuk membentuk fungsi *hyperplane* atau pemisah;
2. C yaitu parameter yang menunjukan seberapa besar kemungkinan menghindari misklasifikasi;
3. *Gamma* merupakan parameter yang menjelaskan seberapa dekat himpunan data yang terjangkau sebagai *support vector* dengan *hyperplane* nya.

3) K-Nearset Neighbor (K-NN)

Metode k-Nearset Neighbor (k-NN) adalah salah satu metode klasifikasi dalam *data mining*. Metode k-NN dilakukan dengan mencari k objek dalam *training data* yang paling dekat atau mirip dengan objek pada *data testing* [14]. Cara untuk mengukur jarak antara data baru dengan data yang lama dapat dilakukan dengan menghitung *Euclidean distance, mahattan distanc, hamming distance,* atau *minkowski distance*. Ukuran yang paling seing digunakan untuk mengukur kedekatan antar data adalah *euclidean distance* [15] yaitu:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2}$$

dimana $a = a_1, a_2, \ldots, a_n$ *.dan* $b = b_1, b_2, \ldots, b_n$ mewakili n nilai atribut dari dua *record.* Langkah-langkah dari algoritma KNN meliputi:

1. Menentukan bilangan *k* yang merupakan bilangan bulat positif sebagai *threshold* jumlah jangkauan tetangga terdekat yang digunakan amatan. Adapun nilai dari *threshold* tersebut direkomendasikan bilangan ganjil.
2. Mengkalkulasi jarak yang dikuadratkan misal dengan menggunakan *euclidien distance* pada *data training* yang digunakan.
3. Mensortir urut amatan dari amatan berjarak terkecil hingga amatan berjarak terbesar.
4. Memanfaatkan pengkategorian "ketetanggaan terdekat" yang paling dominan sebagai bekal memprediksi kategori amatan.

4) Random Forest

Random Forest adalah teknik penggabungan pohon keputusan atau hasil pengembangan metode *Classification and Regression Tree* (CART) serta mengandalkan sejumlah pohon keputusan untuk dapat diterapkan pada data *nonlinier*. Proses klasifikasi Random Forest dimulai dengan membagi secara acak data contoh ke dalam sejumlah pohon keputusan, berikutnya pemberlakuan *voting* bagi setiap kelas kemudian menghimpun *vote* setiap kelas sebagai bahan kesimpulan akhir berdasarkan *vote* terbanyak. Pembentukan sejumlah pohon putusan Random Forest diterapkan melalui seleksi fitur secara acak sehingga meminimalisir kesalahan. Langkah-langkah algoritma tersebut adalah sebagai berikut:

1. Memilih "k" fitur dari total "m" fitur secara acak, di mana $k \leq m$;
2. Menghitung jumlah node "d" di antara "k" fitur menggunakan titik pisah terbaik;
3. Memisahkan *node* menjadi *node* anak menurut *hyperplan* terbaik;
4. Mengulangi langkah 1 – 3 sempai diperoleh jumlah "l" dari *node* terpenuhi;
5. Membangun *forest* dengan mengulangi langkah 1 – 4 bagi sebanyak "n" kali untuk membentuk sebanyak "n" pohon.

5) Decision Tree

Algoritma klasifikasi ini digambarkan berstruktur pohon atau hierarki. Decision Tree memiliki keuntungan yaitu memiliki kemampuan dalam memecah proses pengambilan keputusan yang kompleks menjadi lebih sederhana serta mampu mengeliminasi informasi atau perhitungan yang tidak penting. Istilah lain dari algoritma ini adalah *Classification and Regression Tree (CART)*, metode ini menggambungkan dua macam pohon tahap-tahap berikut:

1. Memilih basis *attribute value*;
2. Untuk setiap *attribute value* katakanlah A, selanjutnya mencari *gaint rasio* informasi ternormalisasi hasil pemisahan A;
3. Membuat simpul dari keputusan yang terbentuk;
4. Mengulang kembali daftar hassil putusan yang diperoleh melalui pemisahan a_best kemudian ditambahkan dengan simpul tersebut sebagai anak-anak simpul.

6) Naive Bayes

Naïve Bayes diusulkan ilmuwan Inggris sebagai produk pemikiran mengenai kombinasi konsep statistika dan peluang. Penerapan metode klasifikasi untuk kebutuhan prediksi ini dilakukan dengan memerhatikan kemungkinan masa depan yang didasari oleh pengalaman masa lalu, kemudian konsep ini menjadi sebuah teorema yang disebut *Bayes Theorem*. Pernyataan tersebut digabungkan dengan pernyataan Naïve yang mengasumsikan bahwa keadaan di antara atribut bersifat independen. Klasifikasi Naive Bayes mengasumsikan ada atau tidaknya sebuah fitur tertentu dalam sebuah kelas tidak berhubungan dengan fitur kelas lain. Model teorema Naive Bayes yang digunakan dalam proses klasifikasi adalah sebagai berikut.

$$P(C|X_1, X_2, \ldots, X_n) = P(C) \prod_{i=1}^{n} P(X_i|C)$$

$$P(C|X) = P(x_1|c)P(x_2|c) \ldots P(x_n|c)P(C)$$

Keterangan :
x: Data kelas yang belum diketahui
c: Hipotesis data merupakan suatu kelas spesifik
P(c|x): Peluang hipotesis berdasar kondisi (peluang posteriori)
P(c): Peluang hipotesis (peluang prior)
P(x|c): Peluang berdasarkan kondisi pada hipotesis
P(x): Peluang c

Secara umum, langkah-langkah algoritma Naive Bayes adalah sebagai berikut:

1. Menghitung jumlah kelas
2. Menghitung banyak observasi untuk tiap kelas
3. Kalikan semua variabel kelas
4. Bandingkan hasil per kelas.

2.4 Evaluasi

Evaluasi merupakan tahapan pemilihan metode klasifikasi yang terbaik berdasarkan nilai akurasi paling tinggi, yakni dengan memerhatikan *Confusion Matrix*. Confusion matrix didefinisikan sebagai matriks dengan elemen-elemen numerik untuk mengetahui seberapa baik sebuah *classifier* dapat mengidentifikasi atau memprediksi kelas dari data [16]. *Confusion Matrix* biasanya berbentuk matriks berdimensi m, dengan m menunjukkan jumlah kelas. Bagian kolom diisi dengan label sebenarnya dari masing-masing kelas, sedangkan bagian baris diisi dengan label kelas yang diprediksi sebagaimana tertera pada tabel 2.

Tabel 2. Confusion Matrix

Prediction	Actual		Total
	Positive	Negative	
Positive	True Positive (TP)	False Negative (FN)	P’
Negative	False Positive (FP)	True Negative (TN)	N’
	P	N	P+N=P’+N’

Akurasi diterjemahkan sebagai persentase total *tuple* dalam *testing data* hasil pengklasifikasian dengan benar oleh *classifier* algoritma *machine learning*. Namun, ukuran akurasi ini hanya cocok pada saat perbandingan jumlah label data yang relatif sama (*balance*). Jika jumlah label data timpang atau *imbalance,* maka dapat menggunakan ukuran lain untuk mengevaluasi pengklasifikasian diantaranya yaitu *precision, recall*, dan *F1-score*. *Precision* merupakan ukuran persentase *tuple* sebagai hasil pelabelan bernilai positif dengan benar sesuai kondisi kenyataan. *Precision* kemudian dikenal pula sebagai ukuran kepastian. *Recall* adalah persentase *tuple* positif yang berhasil dilabeli positif dengan benar atau disebut pula sebagai ukuran kelengkapan. Sementara, *F1-score* merupakan kombinasi pengukuran dari *precision* dan *recall* menjadi suatu metrik.

$$Akurasi = \frac{TP+TN}{P+N} \quad (1)$$

$$Precision = \frac{TP}{TP+FN} = \frac{TP}{P\prime} \quad (2)$$

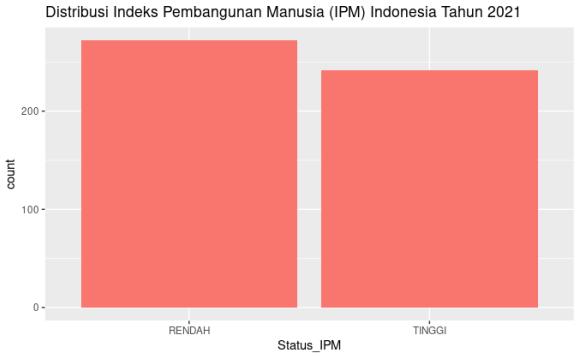
$$Recall = \frac{TP}{TP+FP} = \frac{TP}{P} \quad (3)$$

$$F1 - score = 2\frac{Precision\ x\ Recall}{Precision+Recall} \quad (4)$$

3. Hasil dan Pembahasan

3.1 Preprocessing Data

Tahapan *preprocessing data* dilakukan dengan menggunakan *software* R-Studio dan *Excel.* Tidak ada ditemukan *missing value* pada semua atribut sehingga penanganan *missing value* tidak perlu dilakukan. Pada variabel *Status_IPM* selanjutnya *direcode* menjadi dua kategori yaitu IPM rendah (< 70) dan IPM tinggi (≥ 70) untuk menghindari “*imbalance data*” sehingga didapatkan data yang sudah seimbang pada gambar 2 dan data bagus untuk dilakukan klasifikasi. Variabel PAD dengan satuan juta rupiah ditransformasi menjadi puluhan milyar rupiah dan variabel PDRB per kapita dengan satuan ribu rupiah ditransformasi ke satuan juta rupiah. Transformasi ini dilakukan agar *range* data antar variabel tidak jauh berbeda. Selanjutnya variabel respons *Status_IPM* yang sebelumnya bertipe *character* dikonversi menjadi data faktor yang dibagi menjadi 2 kelas, yaitu ‘*Rendah’* dan ‘*Tinggi’*.



Gambar 2. Distribusi IPM di Kab/Kota Indonesia Tahun 2021 (Sumber: diolah dengan R Studio)

3.2 Analisis Deskriptif

Rerata IPM Indonesia di tahun 2021 yaitu sebesar 69,93 atau tercatata masuk dalam kategori rendah. IPM tertinggi sebesar 87,18 yang berada di Kota Yogyakarta, Provinsi Jawa Tengah. Sementara itu, IPM terendah sebesar 32,84 berada di Kabupaten Nduga, Provinsi Papua. Grafik yang tersaji berikut menunjukan jumlah klasifikasi kabupaten/kota di Indonesia.

Berdasarkan gambar 2, terlihat bahwa IPM kabupaten/kota di Indonesia tahun 2021 dominan yang tergolong ke dalam kategori rendah. Bila dirinci lebih lanjut, terdapat 272 kabupaten/kota dan 242 kabupaten/kota tergolong ke dalam kategori tinggi.

3.3 Pemilihan Model

Pengklasifikasian yang diterapkan pada penelitian ini memanfaatkan enam algoritma klasifikasi, yaitu Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearset Neighbor (K-NN), Random Forest, Decision Tree, serta Naive Bayes. Pengujian hasil klasifikasi diterapkan pula dengan konsep *k-fold cross validation* dan *threshold* k = 10 disertai perulangan sebanyak 1 kali. Pemilihan model terbaik masing-masing algoritma memperhatikan nilai akurasi yang paling tinggi. Berikut merupakan hasil *Confusion Matrix* dari masing-masing algoritma (Sel dalam *Confusion Matrix* merupakan rata-rata dari seluruh sampel k-fold cross validation dengan k = 10 dan 1 pengulangan)

1) Artificial Neural Network (ANN)

Tabel 3. *Confusion Matrix* ANN
(Sumber: diolah dari R Studio)

	True Rendah	*True* Tinggi
Pred. Rendah	52,9	46,1
Pred. Tinggi	0,0	1,0

Hasil analisis *training data* dari penggunaan metode Artificial Neural Network (ANN) didapat rerata akurasi sebesar 53,31%. Berdasar pada *Confusion Matrix* tabel 3, terlihat bahwa dari 514 kabupaten/kota di Indonesia, rata-rata kabupaten/kota yang dilakukan prediksi secara tepat masuk ke dalam kategori IPM rendah sekitar 52,9. Rata-rata kabupaten/kota yang diprediksi secara tepat masuk ke dalam kategori IPM tinggi sekitar 1,0. Di samping itu, terdapat 46,1 kabupaten/kota hasil prediksi IPM rendah namun pada kenyataannya justru masuk ke dalam IPM tinggi, serta 0,0 kabupaten/kota diprediksikan IPM tinggi namun pada kenyataannya justru masuk ke dalam IPM rendah. Algoritma ANN mampu mengklasifikasikan kab/ota di Indonsia dengan akurasi maksimal hanya 58,82%, dihasilkan pada saat *size* = 3. Masih rendahnya nilai akurasi dengan algoritma ini dikarenakan jumlah set data pada penelitian ini belum cukup besar. Seperti yang disebutkan [17] bahwa salah satu kelemahan algoritma neural network adalah membutuhkan *training data* dengan dimensi yang relatif besar.

2) Support Vector Machine (SVM)

Parameter yang digunakan pada metode ini yaitu kernel *Radial Basis Function* (RBF). Adapun hasil *Confusion Matrix* ditunjukkan pada tabel 4.

Tabel 4. *Confusion Matrix* SVM
(Sumber: diolah dari R Studio)

	True Rendah	*True* Tinggi
Pred. Rendah	41,6	12,6
Pred. Tinggi	11,3	34,4

Hasil analisis *training data* dengan metode Support Vector Machine (SVM) dihasilkan rerata akurasi sebesar 76,07%. Berdasarkan *Confusion Matrix* yang tersaji pada tabel 4, dari 514 kabupaten/kota di Indonesia, rata-rata kabupaten/kota hasil prediksi secara tepat masuk pada kategori IPM rendah sekitar 41,6. Rata-rata kabupaten/kota hasil prediksi secara tepat masuk pada kategori IPM tinggi sekitar 34,4. Selain itu, terdapat 12,6 kabupaten/kota diprediksi IPM rendah namun pada kenyataannya masuk pada IPM berkategori tinggi, serta 11,3 kabupaten/kota diprediksi IPM tinggi namun kenyataannya justru berkategori IPM rendah. Algoritma SVM mampu mengklasifikasikan kab/kota di Indonsia dengan akurasi maksimal 80,39%, dihasilkan pada saat sigma = 0,4864648 dan C = 1.

3) K-Nearest Neighbor (K-NN)

Tabel 5. *Confusion Matrix* K-NN
(Sumber: diolah dari R Studio)

	True Rendah	True Tinggi
Pred. Rendah	43,2	13,8
Pred. Tinggi	9,7	33,3

Hasil analisis *training data* dengan metode K-Nearest Neighbor (K-NN) memiliki rata-rata akurasi sebesar 76,45%. Berdasarkan *Confusion Matrix* yang tertera pada tabel 5, dari 514 kabupaten/kota di Indonesia, rata-rata kabupaten/kota yang diprediksi dengan tepat masuk ke dalam kategori IPM rendah sekitar 43,2. Rata-rata kabupaten/kota yang diprediksi dengan tepat masuk ke dalam kategori IPM tinggi sekitar 33,3. Sementara itu, terdapat 13,8 kabupaten/kota diprediksi IPM rendah tetapi ternyata masuk ke dalam IPM tinggi, serta 9,7 kabupaten/kota diprediksi IPM tinggi tetapi ternyata masuk ke dalam IPM rendah. Algoritma K-NN mampu mengklasifikasikan kab/kota di Indonsia dengan akurasi maksimal 82,70%, dihasilkan pada k=11.

4) Random Forest

Tabel 6. *Confusion Matrix* Random Forest
(Sumber: diolah dari R Studio)

	True Rendah	*True* Tinggi
Pred. Rendah	41,6	14,0
Pred. Tinggi	11,3	33,1

Hasil analisis *training data* dengan metode Random Forest memiliki rata-rata akurasi sebesar 74,73%. Berdasarkan *Confusion Matrix* yang disajikan pada tabel 6, dari 514 kabupaten/kota di Indonesia, rata-rata kabupaten/kota yang diprediksi dengan tepat masuk ke dalam kategori IPM rendah sekitar 41,6. Rata-rata kabupaten/kota yang diprediksi dengan tepat masuk ke dalam kategori IPM tinggi sekitar 33,1. Sementara itu, terdapat 14,0 kabupaten/kota diprediksi IPM rendah tetapi ternyata masuk ke dalam IPM tinggi, serta 11,3 kabupaten/kota diprediksi IPM tinggi tetapi ternyata masuk ke dalam IPM rendah. Algoritma Random Forest mampu mengklasifikasikan kab/kota di Indonsia dengan akurasi maksimal 86,27%, dihasilkan pada saat *mtry* = 7 dan *ntree* = 300

5) Decision Tree

Tabel 7. *Confusion Matrix Decision Tree*
(Sumber: diolah dari R Studio)

	True Rendah	*True* Tinggi
Pred. Rendah	42,8	18,3
Pred. Tinggi	10,1	28,8

Hasil analisis *training data* dengan metode Decision Tree memiliki rata-rata akurasi sebesar 71,53%. Berdasarkan *Confusion Matrix* yang disajikan pada tabel 7, dari 514 kabupaten/kota di Indonesia, rata-rata kabupaten/kota yang diprediksi dengan tepat masuk ke dalam kategori IPM rendah sekitar 42,8. Rata-rata kabupaten/kota yang diprediksi dengan tepat masuk ke dalam kategori IPM tinggi sekitar 28,8. Sementara itu, terdapat 18,3 kabupaten/kota diprediksi IPM rendah namun pada kenyataannya masuk ke dalam IPM tinggi, serta 10,1 kabupaten/kota diprediksi IPM tinggi namun kenyataannya jsutru masuk ke dalam IPM rendah. Algoritma Decision Tree mampu mengklasifikasikan kab/kota di Indonsia dengan akurasi maksimal 84,91%, dihasilkan pada saat cp = 0,012.

6) Naive Bayes

Tabel 8. *Confusion Matrix Naive Bayes*
(Sumber: diolah dari R Studio)

	True Rendah	*True* Tinggi
Pred. Rendah	39,3	11,9
Pred. Tinggi	13,6	35,2

Hasil analisis *training data* dengan metode Naïve Bayes memiliki rata-rata akurasi sebesar 74,51%. Berdasarkan *Confusion Matrix* yang disajikan pada tabel 8, dari 514 kabupaten/kota di Indonesia, rata-rata kabupaten/kota yang diprediksi dengan tepat masuk ke dalam kategori IPM rendah sekitar 39,3. Rata-rata kabupaten/kota yang diprediksi dengan tepat masuk ke dalam kategori IPM tinggi sekitar 35,2. Sementara itu, terdapat 11,9 kabupaten/kota diprediksi IPM rendah tetapi ternyata masuk ke dalam IPM tinggi, serta 13,6 kabupaten/kota diprediksi IPM tinggi tetapi ternyata masuk ke dalam IPM rendah. Algoritma Naive Bayes mampu mengklasifikasikan kab/kota di Indonesia dengan akurasi maksimal 82,35%, dihasilkan pada saat *laplace = 0, usekernel= TRUE* dan *adjust = 1*.

3.4 Perbandingan Metode

Berdasarkan hasil analisis dengan menggunakan enam macam metode klasifikasi dalam *machine learning* yaitu Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearset Neighbor (K-NN), Random Forest, Decision Tree, dan Naive Bayes, selanjutnya nilai akurasi dari setiap metode dibandingkan. Adapun hasil perbandingan akurasi masing-masing model disajikan pada Tabel 9.

Tabel 9. Performa Klasifikasi *Machine Learning*
(Sumber: diolah dari R Studio)

Metode	Rata-Rata Akurasi	Akurasi Maksimum
Artificial Neural Network (ANN)	53,31%	58,82%
Support Vector Machine (SVM)	76,08%	88,24%
K-Nearset Neighbor(K-NN)	76,45%	82,70%
Random Forest	74,73%	86,27%
Decision Tree	71,53%	84,91%
Naive Bayes	74,51%	82,35%

Tabel 9 menunjukan rata-rata akurasi dan akurasi optimal yang dihasilkan oleh beberapa algoritma *machine learning*. Berdasarkan tabel tersebut dapat dilihat bahwa algoritma dengan nilai akurasi paling baik adalah algoritma Support Vector Machine (SVM) yang memiliki akurasi maksimal sebesar 88,24%, diikuti oleh Random Forest dengan akurasi maksimal sebesar 86,27%, Decision Tree dengan akurasi maksimal sebesar 84,91%, K-Nearest Neighbor (K-NN) dengan akurasi maksimal 82,70% tetapi secara rata-rata memberikan akurasi yang paling baik, Naive Bayes dengan akurasi maksimal sebesar 82,35%, dan terakhir Artificial Neural Network (ANN) dengan akurasi maksimal sebesar 58,82%. Oleh karena itu, hasil klasifikasi dari algoritma Support Vector Machine (SVM) paling baik digunakan dalam mengklasifikasikan Indeks Pembangunan Manusia (IPM) di Indonesia Tahun 2021 dibandingkan lima algoritma lainnya yang digunakan dalam penelitian ini.

4. Kesimpulan

Berdasarkan hasil analisis pada penelitian ini, diperoleh kesimpulan bahwa algoritma klasifikasi terbaik adalah Support Vector Machine (SVM) dengan rata-rata akurasi sebesar 76,08% dan akurasi maksimal sebesar 88,24%. Model terbaik ini dihasilkan ketika menggunakan parameter kernel *Radial Basis Function* (RBF) dengan sigma = 0,4864648 and C = 1, diikuti oleh Random Forest dengan akurasi model sebesar 86,27%, Decision Tree dengan akurasi model sebesar 84,91%, K-Nearest Neighbor (K-NN dengan akurasi

82,70%, namun secara rata-rata memberikan akurasi yang paling baik, Naive Bayes dengan akurasi model sebesar 82,35%. Terakhir adalah algoritma Artificial Neural Network (ANN) dengan akurasi model sebesar 58,82%. Oleh karena itu, hasil klasifikasi dari algoritma Support Vector Machine (SVM) dapat diimplementasikan dalam pengklasifikasian IPM di Indonesia Tahun 2021.

Adapun manfaat yang diharapkan dari penelitian ini bagi pemerintah adalah untuk mengetahui apakah suatu wilayah memiliki taraf hidup yang rendah atau tinggi, sehingga pemerintah dapat mengambil kebijakan khususnya bagi wilayah yang memiliki IPM yang rendah serta sebagai alokator penentuan Dana Alokasi Umum (DAU). Selain itu, manfaat bagi penelitian lainnya dapat dijadikan sebegai referensi khususnya dalam kasus pengklasifikasian Indeks Pembangunan Manusia. Berdasarkan hasil penelitian ini disarankan untuk penelitian selanjutnya untuk menggunakan klasifikasi yang berbeda, misalnya dengan algoritma *boosting* seperti Adaboost, XGboost, dan lainnya. Selain itu, dapat pula melibatkan variabel lainnya, seperti angka kriminalitas, tingkat kemiskinan, kepadatan penduduk, atau pengeluaran per kapita.

Reference

[1] United Nations Development Programme, "Human development report 2019: beyond income, beyond averages, beyond today", In United Nations Development Program, 2019.

[2] Fauzi, Fatkhurokhman, "K-nearset neighbor (k-nn) dan support vector machine (svm) untuk klasifikasi indeks pembangunan manusia provinsi jawa tengah", Jurnal Mipa, 40(2), 118–124, 2019.

[3] Gunadi, G., & Sensuse, D. I, "Penerapan metode data mining market basket analysis terhadap data penjualan produk buku dengan menggunakan algoritma apriori dan frequent pattern growth (fp-growth): studi kasus percetakan pt. Gramedia", Telematika MKOM, 4(1), 118-132, 2016.

[4] Irmawati, Irmawati, Zainudin, Zahir, & Yuyun, Yuyun, "Data mining untuk penentuan model kelulusan murid sma pada perguruan tinggi negeri; studi kasus di iain bone", JIKO (Jurnal Informatika Dan Komputer), 3(2), 113– 118. https://doi.org/10.33387/jiko.v3i2.1800, 2020.

[5] Kranjčić, Nikola, Medak, Damir, Župan, Robert, & Rezo, Milan, "Support vector machine accuracy assessment for extracting green urban areas in towns", Remote Sensing, 11(6), https://doi.org/10.3390/rs11060655, 2019.

[6] Fauzi, Fatkhurokhman, Yamin, Moh, & Wahyu, Tiani, "Klasifikasi indeks pembangunan manusia kabupaten / kota se-indonesia dengan pendekatan smooth support vector machine (ssvm) kernel radial basis function (rbf)", Seminar Nasional Pendidikan, Sains Dan Teknologi Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Muhammadiyah Semarang, 88–97, Retrieved from https://jurnal.unimus.ac.id/index.php/psn 12012010/article/view/2986, 2017.

[7] Fathurrahman, M., & Qisthi, N, " Klasifikasi indeks pembangunan manusia (ipm) di pulau sumatera pada dataset multi-class dengan metode artificial neural network (ann)", In *Seminar Nasional Fisika* (Vol. 1, No. 1, pp. 377-384, 2021.

[8] Darsyah, M. Y, "Klasifikasi indeks pembangunan manusia (ipm) dengan pendekatan k-nearset neighbor (k-nn)". In *Prosiding Seminar Nasional & Internasional*, 2017.

[9] Kemala, I., & Wijayanto, A. W, "Perbandingan kinerja metode bagging dan non-ensemble machine learning pada klasifikasi wilayah di indonesia menurut indeks pembangunan manusia", *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, *9*(2), 269-275, 2021.

[10] Kusumodestoni, R. Hadapiningradja, & Sarwido, Sarwido, "Komparasi model support vector machines (svm) dan neural network untuk mengetahui tingkat akurasi prediksi tertinggi harga saham", Jurnal Informatika Upgris, 3(1). https://doi.org/10.26877/jiu.v3i1.1536, 2017.

[11] Tomasevic, Nikola, Gvozdenovic, Nikola, & Vranes, Sanja, "An overview and comparison of supervised data mining techniques for student exam performance prediction", Computers and Education, 143, 103676. https://doi.org/10.1016/j.compedu.2019. 103676, 2020.

[12] Wibisono, A. B., & Fahrurozi, A, "Perbandingan algoritma klasifikasi dalam pengklasifikasian data penyakit jantung coroner", Jurnal Ilmiah Teknologi Dan Rekayasa, 24(3), 161–170, https://doi.org/10.35760/tr.2019.v24i3.239, 2019.

[13] Sihombing, Pardomuan Robinson, "Perbandingan metode artificial neural network (ann) dan support vector machine (svm) untuk klasifikasi kinerja perusahaan daerah air minum (pdam) di indonesia", Jurnal Ilmu Komputer, 13(1), 9. https://doi.org/10.24843/jik.2020.v13.i01. p02, 2020.

[14] Wu, X., & Kumar, V, "*The top ten algorithms in data mining*", CRC press, 2009.

[15] Bramer, M, "Introduction to classification: na¨ive bayes and nearest neighbour", Principles of Data Mining, 23-39, 2007.

[16] Rahmad, F., Suryanto, Y., & Ramli, K, "Performance comparison of anti-spam technology using confusion matrix classification". IOP Conference Series: Materials Science And Engineering, 879(1), https://doi.org/10.1088/1757-899X/879/1/012076, 2020.

[17] Purwaningsih, E, "Seleksi mobil berdasarkan fitur dengan komparasi metode klasifikasi neural network, support vector machine, dan algoritma c4.5", *Jurnal Pilar Nusa Mandiri*, *12*(2), 153-160, 2016.