ISSN:2302-8149
e-ISSN:2540-9719

Perbandingan Metode *Ensemble Learning: Random Forest, Support Vector Machine, AdaBoost* pada Klasifikasi Indeks Pembangunan Manusia (IPM)

Comparison of Ensemble Learning Method: Random Forest, Support Vector Machine, AdaBoost for Classification Human Development Index (HDI)

[1]Ressa Isnaini Arumnisaa*, [2]Arie Wahyu Wijayanto
[12]Program Studi DIV Statistika, Politeknik Statistika STIS
Jalan Otto Iskandardinata No. 64C, Jakarta, Indonesia
*e-mail: 211910805@stis.ac.id

(***received*:** 19 Desember 2022, ***revised*:** 20 Desember 2022, ***accepted*:** 24 Desember 2022)

Abstrak

Klasifikasi dalam *supervised learning* adalah cara untuk menemukan pola dari basis data yang telah diketahui labelnya. Dalam klasifikasi pada *machine learning*, dikenal dengan istilah *ensemble classifier*. Cara kerja *ensemble classifier* dinilai dapat meningkatkan akurasi model serta mengoptimalkan kinerja klasifikasi. Penelitian ini bertujuan untuk menganalisis perbandingan algoritma yang bekerja dengan *ensemble learning*, diantaranya *Random Forest, Support Vector Machine (SVM),* dan *AdaBoost.* Data yang digunakan penelitian ini merupakan data Indeks Pembangunan Manusia (IPM) kabupaten/kota di Indonesia dan variabel-variabel lain yang berkaitan kuat dengan pembangunan manusia seperti PDRB per kapita, APK, APM, TPAK, TPT, Tingkat Kemiskinan, Kedalaman Kemiskinan, Keparahan Kemiskinan, dan rata-rata konsumsi per kapita. Alasan digunakannya data IPM selain merupakan variable makro ekonomi yang penting dalam penggambaran keadaa SDM di Indonesia, juga karena IPM telah memiliki klasifikasi yang jelas menurut Badan Pusat Statistik (BPS) untuk dapat dilakukan *supervised learning.* Perbandingan evaluasi model menggunakan metrik akurasi, spesifisitas, sensitivitas, dan statistik *kappa.* Alur analisis dimulai dengan *data preprocessing*, dilanjutkan dengan *resampling* dan *cross-validation*, selanjutnya membangun model pengklasifikasi menggunakan algoritma *Random Forest; Support Vector Machine (SVM);* dan *AdaBoost.* Tahap terakhir adalah evaluasi model dengan membandingkan model terbaik dalam klasifikasi kabupaten/kota menurut IPM. Hasil penelitian menunjukkan bahwa model *Random Forest* memiliki perfoma terbaik dibandingkan model *Support Vector Machine (SVM)* dan *AdaBoost* dengan akurasi sebesar 85,23%, spesifisitas sebesar 71,63%, sensitivitas sebesar 95,05%, dan statistik *kappa* 0,7698. Dari penelitian ini, dapat dikembangkan sebuah *ensemble classifier* untuk membantu mengklasifikasikan Indeks Pembangunan Manusia di Indonesia.

Kata kunci: *AdaBoost*, *Random Forest*, *Support Vector Machine*, Pembelajaran Ansambel, Indeks Pembangunan Manusia

Abstract

Classification in supervised learning is a way to find patterns in database that the classes are already known. In the classification of machine learning, there is a term called ensemble classifier. The workings of the ensemble classifier aimed to improve model accuracy and optimize classification performance. This study aims to analyze the comparison of algorithms that work with ensemble learning, including Random Forest, Support Vector Machine (SVM), and AdaBoost. The data used is the Human Development Index (HDI) of districts/cities in Indonesia. Other variables that are strongly related to human development are GRDP per capita, gross enrollment rate, net enrollment rate, labor force participation rate, unemployment rate, poverty rate, poverty depth, poverty severity, and average consumption per capita. The reason for using HDI is that apart from being an important macroeconomic variable in describing the condition of human resources in Indonesia, HDI already has an obvious classification according to the Badan Pusat Statistik (BPS) so that supervised learning can be applied. Comparison of model evaluation using accuracy, specificity, sensitivity, and kappa

 ISSN:2302-8149
e-ISSN:2540-9719

statistics. The analysis flow starts with data preprocessing, resampling and cross-validation, then modeling using the Random Forest, Support Vector Machine (SVM), and AdaBoost algorithm. The final stage is the model evaluation by comparing the best models in the classifications of districts/ cities according to HDI. The results showed that the Random Forest model had the best performance compared to the Support Vector Machine (SVM) and AdaBoost models with an accuracy value of 85,23%, *specificity of 71,63%, sensitivity of 95,05%, and kappa coefficient of 0,7698. From this research, the an ensemble classifier can be developed to help classify scores on the Human Development Index in Indonesia.*

Keywords: *AdaBoost, Random Forest, Support Vector Machine, Ensemble Learning, Human Development Index*

1 Pendahuluan

Perkembangan dunia sains dan teknologi menghasilkan metode pembelajaran yang canggih, lebih efektif dan efisien dengan memanfaatkan *data mining.* Salah satu metode data mining yang popular digunakan ialah klasifikasi. Klasifikasi adalah cara mengelompokkan suatu objek ke dalam kelas atau kategori. Tujuan klasifikasi adalah menemukan fungsi keputusan yang secara akurat memprediksi kelas data yang berasal dari fungsi distribusi yang sama dengan data untuk pelatihan [1]. Untuk dapat melakukan klasifikasi, diperlukan teori atau rujukan untuk dapat mengetahui kelas/label dari suatu *dataset.* Dalam melakukan perbandingan algoritma klasifikasi, data Indeks Pembangunan Manusia (IPM) digunakan dalam penelitian ini. Alasan digunakannya data IPM karena selain merupakan variabel makro ekonomi yang penting dalam penggambaran keadaa SDM di Indonesia, juga karena IPM telah memiliki klasifikasi yang jelas untuk dapat dilakukan *supervised learning.* Penetapan klasifikasi IPM menurut Badan Pusat Statistik (BPS), yaitu dikatakan rendah jika IPM < 60, sedang jika 60 ≤ IPM < 70, tinggi jika 70 ≤ IPM < 80, dan sangat tinggi jika IPM ≥ 80 [2].

Indeks Pembangunan Manusia dibentuk berdasarkan tiga dimensi, yaitu dimensi Kesehatan yang digambarkan dengan Angka Harapan Hidup (AHH); dimensi Pendidikan yang digambarkan dengan Harapan Lama Sekolah (HLS) dan Rata-rata Lama Sekolah (RLS); serta dimensi pengeluaran yang digambarkan dengan Produk nasional Bruto (PNB). Indeks ini bermanfaat dalam memberikan gambaran menyeluruh terkait pencapaian pembangunan manusia dalam rangka meningkatkan kualitas hidup manusia. Tinginya nilai IPM pada suatu daerah menandakan pencapaian pembangunan manusia menjadi semakin baik. Untuk dapat melakukan klasifikasi dengan baik, maka dalam penelitian ini akan digunakan atribut-atribut lain diluar komponen penyusun IPM yang berhubungan dengan pembangunan manusia. Selanjutnya akan dibangun beberapa model *classifier* yang nantinya akan dipilih salah satu *classifier* terbaik yang dapat mengklasifikasikan IPM dengan tepat sesuai kelasnya.

Dalam klasifikasi menggunakan *machine learning,* dikenal dengan istilah *ensemble classifier*. *Ensemble classifier* adalah metode yang mengkombinasikan beberapa algoritma klasifikasi untuk meningkatkan kekuatan model serta untuk meningkatkan kinerja klasifikasi [3]. *Ensemble classifier* dinilai lebih powerful dibandingkan *single classifier* dikarenakan beberapa hal sebagai berikut: (1) *training set* tidak selalu memberikan informasi yang cukup untuk memilih satu hipotesis yang akurat, (2) proses pembelajaran dari sebuah *classifier* lemah dalam kondisi tertentu, (3) ruang hipotesis yang dicari mungkin tidak berisi fungsi target sebenarnya sementara *ensemble classifier* dapat memberikan perkiraan yang baik [4]. *Ensemble classifier* dinilai lebih resisten terhadap *noise,* mampu meminimalkan bias dan varians daripada *single learning*. Oleh karena itu, penelitian ini memiliki tujuan untuk melakukan analisis perbandingan algoritma-algoritma yang menggunakan konsep *ensemble learning*, seperti *Random Forest*, *Support Vector Machine (SVM)*, dan *AdaBoost* menggunakan data IPM. Hasil dari penelitian ini diharapkan mampu memberikan informasi mengenai algoritma ansambel terbaik untuk mengklasifikasikan Indeks Pembangunan Manusia di Indonesia.

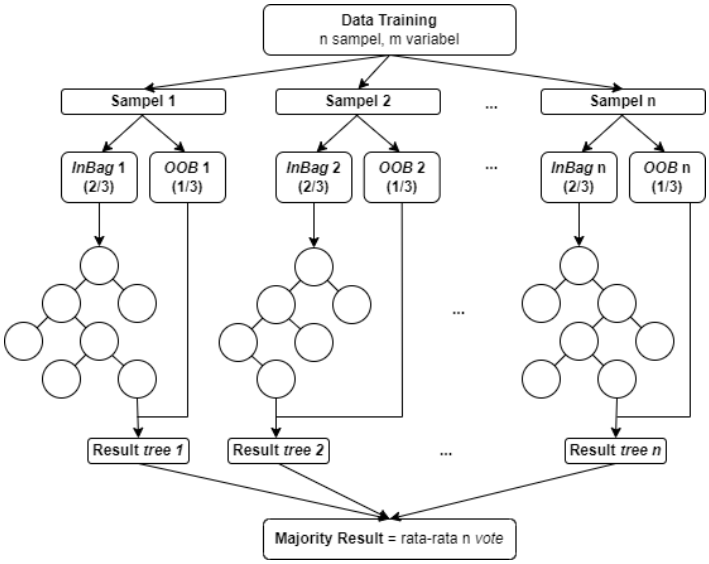
 ISSN:2302-8149
e-ISSN:2540-9719

2 Tinjauan Literatur

Random Forest

Random Forest atau *random decision forest* adalah metode klasifikasi berbasis pohon keputusan atau *decision tree*. Jika pada metode *decision tree* klasik seperti C4.5 dan ID3 hanya dihasilkan satu pohon untuk memodelkan seluruh *training set*, maka metode *Random Forest* akan dihasilkan banyak pohon dari sampel di *training set.* Hasil akhir dari klasifikasinya adalah *decision* mayoritas dari seluruh pohon yang terbentuk. Terdapat dua hal yang menjadi dasar pemikiran utama *Random Forest*: (1) Sebagian besar pohon dapat menghasilkan prediksi yang benar untuk Sebagian besar data, (2) untuk pohon dengan prediksi yang salah, prediksi jatuh pada kelas yang berbeda.

Algoritma pada *Random Forest* termasuk dalam metode *bagging ensemble learning.* Pada metode *Random Forest, training set* akan dilakukan pengambilan sampel sebanyak pohon yang diinginkan dengan metode *random sampling with replacement* (SRS WR). Proses ini disebut dengan istilah *bagging.* Kemudian dari setiap sampel *training set* akan dihasilkan satu *decision tree.* Satu objek pada *testing set* akan diproses oleh seluruh pohon yang terbentuk dan keputusan klasifikasi akhir adalah keputusan mayoritas atau *decision* yang paling banyak terpilih oleh pohon-pohon. Perhitungan *error* dilakukan untuk objek-objek yang tidak digunakan saat pembuatan pohon keputusan, disebut juga dengan *Out-of-bag* (OOB) *error estimate* [5]. Ilustrasi cara kerja *Random Forest* ditunjukkan oleh Gambar 1.



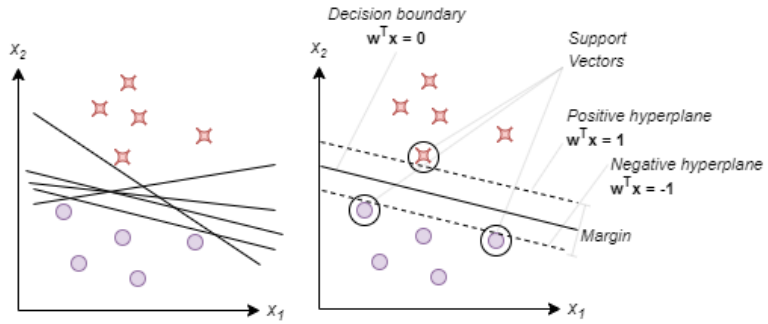
Gambar 1. Ilustrasi algoritma *Random Forest*

Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah metode pengelompokan diskriminatif dengan menggunakan sebuah *hyperplane* sebagai pemisah antar kelas dengan memaksimalkan margin diantara kelas-kelas tersebut. SVM merupakan suatu metode klasifikasi yang termasuk dalam kelas *Artificial Neural Network* yang mencapai solusi berupa global optimal [6]. Parameter pada SVM adalah C (*cost*) dan jenis kernel.

Algoritma pada SVM termasuk dalam metode *ensemble learning* karena sistem pembelajaran model ini menggunakan ruang hipotesis berupa fungsi-fungsi dalam sebuah ruang fitur berdimensi tinggi. Terdapat banyak batas yang dapat digunakan untuk memisahkan kelas-kelas tersebut, namun hanya terdapat satu batas saja yang dapat memaksimalkan margin atau jarak antar kelas. Ilustrasi SVM digambarkan pada Gambar 2 sebelah kiri, terdapat banyak *hyperplane* yang dapat memisahkan kelas lingkaran dan kelas bintang. Namun, *hyperplane* yang dapat memaksimalkan margin merupakan *classifier* terbaik (Gambar 2 kanan).

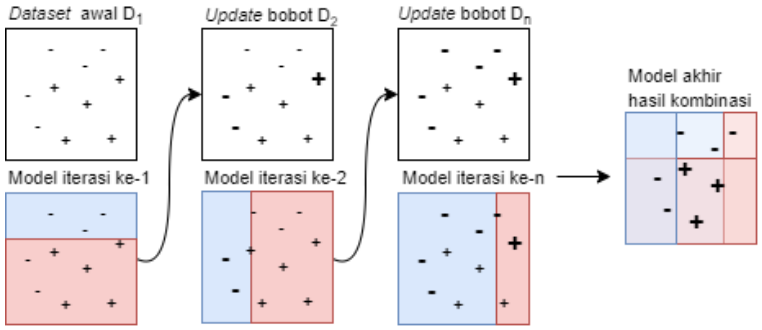
 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 2. Ilustrasi *Support Vector Machine*

AdaBoost

Metode *Boosting* (*AdaBoost*) adalah metode *ensemble* karena dalam proses klasifikasi dan prediksi, konsep kerja *AdaBoost* adalah dengan dengan cara membangkitkan kombinasi dari suatu model. Metode ini dapat meningkatkan ketelitian untuk masalah kelas yang tidak seimbang dan meningkatkan identifikasi dari kelas minoritas yang sulit serta menjaga kemampuan klasifikasi dari kelas mayoritas. Pada dasarnya, metode boosting Setiap model yang dibangkitkan melalui sejumlah *bagging* dan *boosting* yang diinginkan akan memiliki atribut berupa nilai bobot. Hasil klasifikasi akhir atau prediksi yang dipilih adalah model yang memiliki nilai bobot paling besar [7]. Berikut ilustrasi cara kerja *AdaBoost* ditunjukkan pada Gambar 3.



Gambar 3. Ilustrasi *Bagging AdaBoost*

Penelitian Terdahulu

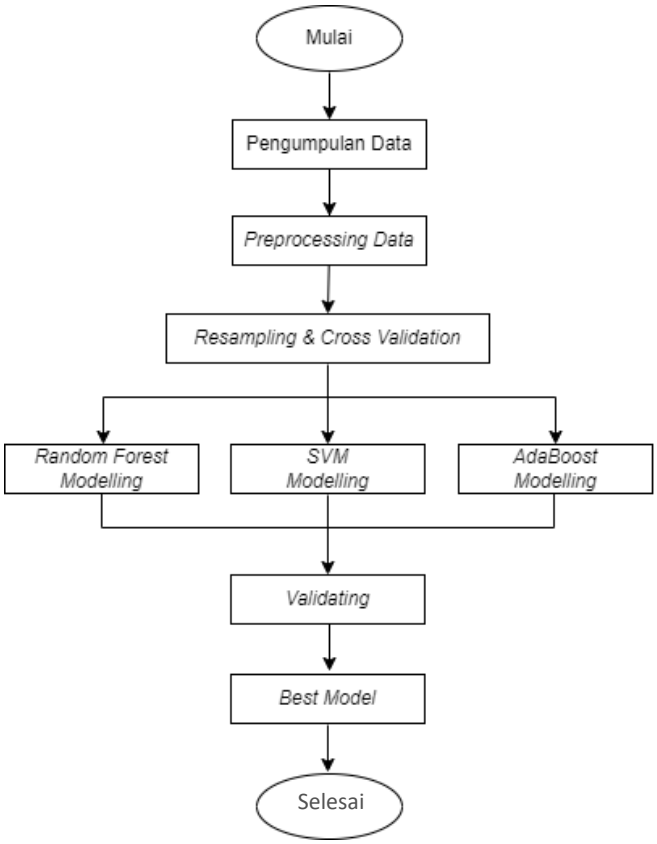
Sejumlah penelitian tentang perbandingan metode klasifikasi dikembangkan dalam berbagai persoalan. Penelitian [8] melakukan perbandingan algoritma *Random Forest* dengan *Support Vector Machine* (SVM) dalam mengklasifikasikan rumah tangga penerima Raskin. Hasil penelitian menunjukkan metode SVM dengan akurasi 72,33 persen; sensitivitas 71,04 persen; spesifisitas 73,26 persen mempunyai performa lebih baik dibandingkan dengan *Random Forest*. Penelitian lain, [9] melakukan perbandingan *Random Forest* dengan *Support Vector Machine* dalam klasifikasi status Indeks Mitigasi dan Kesiapsiagaan Bencana (IMKB) satuan kerja BPS. Hasil penelitian menunjukkan bahwa metode klasifikasi *Random Forest* menghasilkan akurasi, presisi, dan recall yang lebih tinggi dibandingkan SVM. Selanjutnya, penelitian mengenai algoritma *Support Vector Machine*, *conditional inference trees*, dan *Random Forest* pada klasifikasi capaian belajar siswa SMP di Indonesia tahun 2019. Penelitian ini menunjukan bahwa algoritma SVM mengasilkan model terbaik dengan akurasi 80 persen, recall 97 persen, *kappa statistics* 0,38, dan F1-score 87 persen yang tertinggi dibandingkan algoritma *conditional tree* dan *Random Forest* [10]. Penelitian [11] melakukan analisis komparasi algoritma klasifikasi *data mining* dalam pengelompokan *website phising* menggunakan *Naïve Bayes*, *Random Forest, Decision Tree,* dan *Support Vector Machine*. Hasil analisis didapatkan bahwa algortma *Random Forest* memiliki kinerja terbaik dengan metrik akurasi 90,77%; presisi80,90%; dan sensitivitas 95,61%.

ISSN:2302-8149
e-ISSN:2540-9719

Sementara itu, beberapa penelitian terdahulu mengenai komparasi algoritma *machine learning* pada klasifikasi Indeks Pembangunan Manusia telah dilakukan. Penelitian mengenai klasifikasi IPM dengan lokus Provinsi Jawa Tengah menggunakan metode *K-Nearest Neighbor* dan *Support Vector Machine*. Hasil penelitian ini didapatkan bahwa metode *SVM* memiliki akurasi lebih besar dibandingkan *KNN* dengan nilai akurasi sebesar 95,36% [12]. Penelitian ini menunjukkan bahwa metode *ensemble SVM* memiliki kinerja lebih baik daripada metode *single KNN*. Penelitian selanjutnya klasifikasi IPM dengan lokus Pulau Jawa dilakukan menggunakan metode *K-Nearest Neighbor* dan *Support Vector Machine* [13]. Kesimpulan penelitian ini menunjukkan hasil yang sama bahwa metode yang lebih baik adalah *Support Vector Machine* dengan parameter terbaik yang digunakan yaitu kernel-linier, *gamma* sebesar 1, dan *cost* sebesar 5, serta akurasi yang dihasilkan adalah sebesar 88,89 persen. Penelitian lain mengenai klasifikasi IPM dengan lokus Pulau Sumatera menggunakan algoritma *Support Vector Machine* dan *Artificial Neural Network* [14]. Hasil penelitian menunjukkan nilai presisi metode *Artificial Neural Network (ANN)* lebih tinggi daripada nilai presisi metode *Support Vector Machine (SVM)*.

3 Metode Penelitian

Implementasi metode yang digunakan adalah metode data mining menggunakan algoritma *Random Forest, Support Vector Machine*, dan *AdaBoost* untuk memprediksi Indeks Pembangunan Manusia. Secara keseluruhan, tahapan-tahapan dalam penelitian seperti pada Gambar 4 berikut.



Gambar 4. Diagram Alur Penelitian

3.1 Pengumpulan Data

Data yang digunakan merupakan data sekunder yang bersumber dari *website* resmi www.bps.go.id dan apkapm.data.kemdikbud.go.id. Data tersebut terdiri dari 505 objek dengan unit analisis berupa kabupaten/kota di Indonesia pada tahun 2021. Daftar atribut yang digunakan ini tertera pada Tabel 1. Dataset tersebut diolah menggunakan program R.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 1. Daftar Atribut yang Digunakan

Deskripsi	Variabel	Satuan	Tipe Data
Indeks Pembangunan Manusia (IPM)	Dependen	Indeks	Kategorik
Produk Domestik Regional Bruto (PDRB) per kapita	Independen	Ribu Rupiah	Numerik
Angka Partisipasi Kasar (APK)	Independen	Persen	Numerik
Angka Partisipasi Murni (APM)	Independen	Persen	Numerik
Tingkat Partisipasi Angkatan Kerja (TPAK)	Independen	Persen	Numerik
Tingkat Pengangguran Terbuka (TPT)	Independen	Persen	Numerik
Tingkat Kemiskinan (P0)	Independen	Persen	Numerik
Kedalaman Kemiskinan (P1)	Independen	Indeks	Numerik
Keparahan Kemiskinan (P2)	Independen	Indeks	Numerik
Rata-rata konsumsi per kapita	Independen	Satuan Komoditas	Numerik

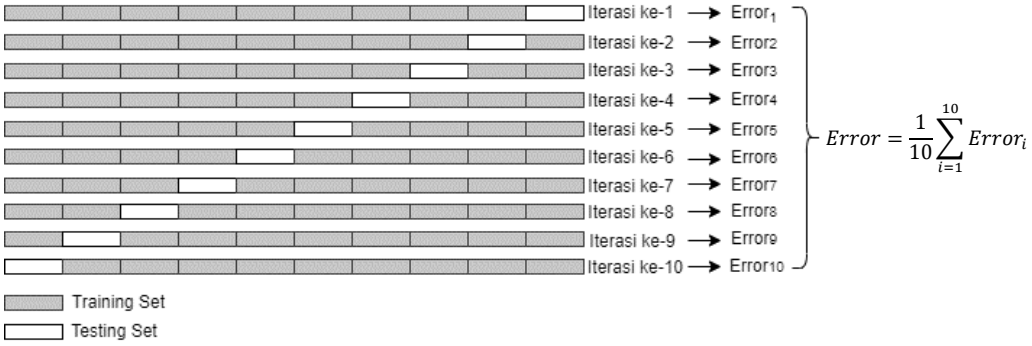
3.2 Preprocessing Data

Data *preprocessing* merupakan langkah awal dalam menyiapkan data dan melakukan transformasi terhadap data mentah sesuai dengan format untuk dilakukan analisis selanjutnya [1]**.** Tahapan-tahapan dalam *preprocessing* meliputi pembersihan data, integrasi data, reduksi data, dan transformasi data. *Data cleaning* meliputi pembersihan *missing value* dan penanganan pencilan. *Data integration* meliputi penyatuan arsip data yang berbeda-beda menjadi satu kesatuan *dataset*. *Data reduction* meliputi pemilihan atribut atau *features* yang berkorelasi kuat dengan label dan penanganan redundansi data. *Data transformation* meliputi standarisasi atau normalisasi data yang memiliki satuan berbeda.

3.3 Resampling & Cross-Validation

Untuk dapat melakukan evaluasi antar model, maka dilakukan pembagian *dataset* menjadi dua yakni data pelatihan dan data uji. Data pelatihan atau *training set* digunakan untuk membangun model dan data uji atau *testing set* digunakan untuk mengetahui seberapa baik klasifikasi yang dihasilkan. Sebelum dilakukan pembagian *dataset*, terlebih dahulu dilakukan pengecekan distribusi data untuk mengetahui keseimbangan kelas. Ketidakseimbangan kelas (*class imbalance*) dapat menyebabkan *misclassification* pada kelas minoritas. Dalam penelitian ini akan dilakukan resampling *Synthetic Minority Oversampling Technique* (SMOTE) jika terdapat indikasi ketidakseimbangan data. Kemudian, untuk membagi *data set* menjadi *training set* dan *testing set*, maka akan dilakukan *K-fold cross-validation.*

SMOTE merupakan salah satu metode *resampling* untuk mengatasi data *imbalance*. Konsep kerja *SMOTE* adalah dengan menambahkan jumlah sampel pada kelas minoritas agar setara dengan kelas mayoritas. Penambahan sampel ini dilakukan dengan cara membangkitkan data sintetik berdasarkan tetangga terdekat *k-nearest neighbor* yang dipilih berdasarkan jarak *Euclidean* antara kedua data [15].



Gambar 5. Ilustrasi 10-*fold cross-validation*

 ISSN:2302-8149
e-ISSN:2540-9719

K-fold cross-validation merupakan metode untuk membagi data secara acak menjadi beberapa *k* bagian yang kemudian beberapa bagian digunakan untuk training dan beberapa bagian yang lain untuk testing. Pada penelitian ini akan digunakan sejumlah 10 bagian. Pada 10-*fold cross-validation, dataset* dibagi menjadi 10 bagian (S_1, S_2, …, S_{10}) yang kemudian pada iterasi pertama dilakukan *training* dengan menggunakan S_1 sampai dengan S_9, selanjutnya model yang dihasilkan dari *training* akan diuji dengan menggunakan S_{10}. Ilustrasi 10-*fold cross-validation* ditunjukkan pada Gambar 5.

3.4 Modeling

Setelah membagi *dataset* menjadi *training set* dan *testing set*, langkah selanjutnya adalah pemodelan menggunakan beberapa algoritma *ensemble learning.* Cara kerja klasifikasi dengan metode ansambel adalah menggabungkan kumpulan pengklasifikasi tunggal untuk membangun model komposit yang memberikan performa klasifikasi yang lebih baik. Sejumlah studi *machine learning* membuktikan bahwa kesalahan generalisasi dapat dikurangu dengan menggabungkan output dari beberapa pengklasifikasian [16]**.** Metode pembelejaran ansambel dapat meminimalkan varians, hal ini terjadi karena efek bias rata-rata ansambel untuk mengurangi varian dari satu set pengkalsifikasian [7]. Dalam penelitian ini akan digunakan algoritma *Random Forest, Support Vector Machine* (SVM), dan *AdaBoost.*

3.4 Validating Model

Setelah dilakukan pemodelan dengan beberapa algoritma, maka untuk melakukan perbandingan antar *classifier* perlu dilakukan evaluasi seberapa baik *classifier* melakukan prediksi pada label *testing set.* Jika hasil klasifikasi oleh model yang dihasilkan sama dengan kelas/label suatu objek dalam *testing set*, maka disebut *correct classification.* Sebaliknya jika hasil klasifikasi oleh model yang dihasilkan berbeda dengan kelas/ label suatu objek dalam *testing set*, maka disebut dengan *misclassification.* Model dikatakan semakin baik kinerjanya apabila model tersebut menghasilkan semakin banyak jumlah *correct classification*.

Untuk melakukan validasi model klasifikasi yang dibangun, maka akan digunakan *confusion matrix.* Dari *confusion matrix* dapat diturunkan beberapa metrik evaluasi *classifier*. Dalam penelitian ini akan digunakan metrik akurasi, spesifisitas, sensitivitas dan stiatistik *kappa*. Berikut adalah format umum dari *confusion matrix.*

Tabel 2. Confusion Matrix Klasifikasi Dua Kelas

		Kelas Prediksi		Total
		Ya	Tidak	
Kelas Sebenarnya	Ya	TP	FN	P’
	Tidak	FP	TN	N’
Total		P	N	All

1) Akurasi

Akurasi adalah persentase objek yang diklasifikasikan dengan benar. *Accuracy* = (TP+TN) / All.

2) Kappa

Kappa statistic merupakan ukuran yang penting dalam evaluasi model klasifikasi khususnya pada data *imbalance*. *Kappa* = (*observed accuracy* - *expected accuracy*) / (1 - *expected accuracy*).

3) Sensitivitas

Sensitivitas adalah rasio dari kelas ‘Ya’ yang diprediksi benar terhadap seluruh kelas ‘Ya’ pada keadaan yang sebenarnya. Ukuran ini disebut dengan *recall* atau *true positive rate* (TPR). *Sensitivity* = TP / (TP+FN)

4) Spesifisitas

Spesifisitas adalah rasio dari kelas ‘Tidak’ yang diprediksi benar terhadap seluruh kelas ‘Tidak’ pada keadaan sebenarnya. *Spesificity* = TN / (TN+FP).

ISSN:2302-8149
e-ISSN:2540-9719

4 Hasil dan Pembahasan

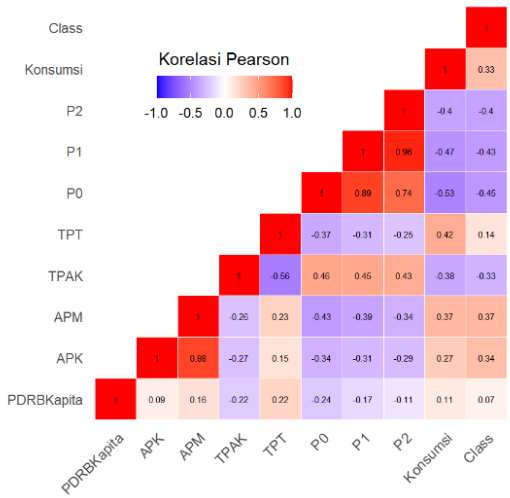
4.1 Preprocessing Dataset

Preprocessing data pada penelitian ini terdiri atas *data cleaning*, *data reduction*, dan *data transformation*. Langkah pertama adalah melakukan *data cleaning*. Pemeriksaan missing data menunjukkan terdapat 10 *missing value* pada atribut Tingkat Pengangguran Terbuka (TPT). Untuk mengatasi *missing value* tersebut, digunakan *k-nearest neighbor* dengan jumlah tetangga terdekat sebanyak 5. Selanjutnya, dilakukan deteksi *outlier* atau pencilan pada data yang sudah tidak mengandung *missing value*. Hasil deteksi menunjukkan terdapat pencilan di semua atribut yang digunakan dengan jumlah pencilan tergambar pada Tabel 3. Untuk mengatasi hal itu, maka dilakukan transformasi data menggunakan transformasi min-max. Transformasi ini juga diperlukan karena range dan satuan antar atribut yang digunakan berbeda-beda. Atribut dengan skala besar akan mendominasi pemodelan.

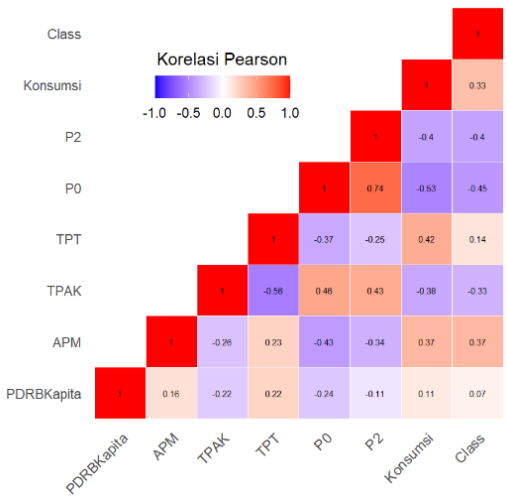
Tabel 3. Jumlah Pencilan pada Masing-masing Atribut

Atribut	Jumlah *Outlier*
PDRB per kapita	47
APK	17
APM	19
TPAK	20
TPT	8
P0	35
P1	36
P2	40
Konsumsi	5

Langkah selanjutnya adalah *data reduction*. *Data reduction* dilakukan untuk menghindari multikolinearitas yang menyebabkan redundansi data. Pada Gambar 6 terlihat bahwa plot korelasi menunjukkan adanya korelasi yang tinggi antara atribut APK dengan APM, P0 dengan P1, dan atribut P1 dengan P2. Oleh karena itu, atribut APK dan P1 dihilangkan dari *dataset* karena dinilai memiliki korelasi yang kecil terhadap label atau kelas. Antar atribut tidak terdapat multikolinieritas jika koefisien korelasi tidak lebih dari 0,80 [17]. Plot korelasi *dataset* yang telah dilakukan *data reduction* (Gambar 7) menunjukkan korelasi antar antribut 0.74 yang berarti tidak terdapat multikolinearitas yang menyebabkan redundansi data.



Gambar 6. Plot Korelasi Sebelum *Data Reduction*

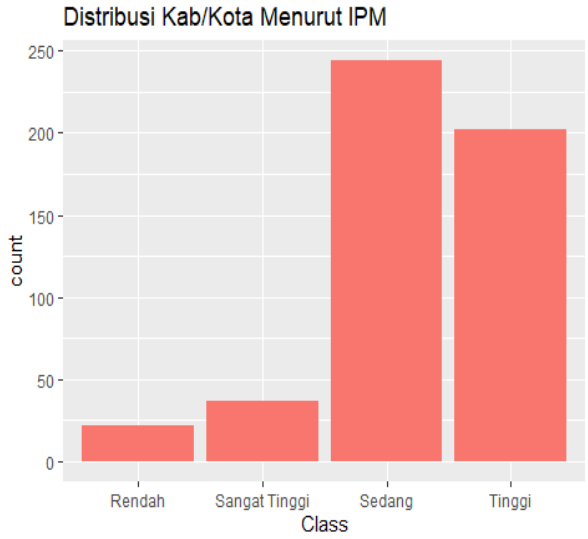


Gambar 7. Plot Korelasi Sesudah *Data Reduction*

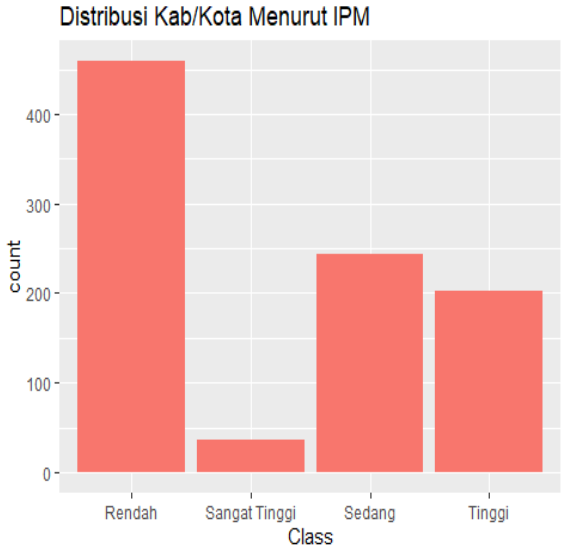
 ISSN:2302-8149
e-ISSN:2540-9719

4.2 *Resampling dan Cross-Validation*

Terlebih dahulu melihat distribusi atau sebaran data yang telah melalui *data preprocessing*. Pada Gambar 8 terlihat bahwa sebelum dilakukan *resampling*, sebaran data menurut kelas IPM kab/kota terdistribusi *imbalance*. Kelas 'sedang' dan 'tinggi' mendominasi lebih besar dibandingkan kelas rendah dan sangat tinggi. Ketidakseimbangan kelas (*class imbalance*) dapat memengaruhi kinerja klasifikasi karena akan banyak *misclassification*. Oleh karena itu perlu dilakukan penyeimbangan data dengan cara *resampling* menggunakan metode *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE dianggap metode resampling yang baik karena karena meminimalisir redundansi data dan tidak menghilangkan informasi penting. Setelah dilakukan *resampling*, maka distribusi data baru seperti pada Gambar 9. Distribusi pada gambar tersebut dikatakan sudah seimbang karena jumlah sampel pada masing-masing kelas lebih merata khususnya kelas 'rendah', 'sedang', dan 'tinggi', selain itu juga terdapat kelas yang dapat dikatakan kelas mayor yaitu 'rendah' dan kelas minor yaitu 'sangat tinggi'.



Gambar 8. Distribusi data sebelum dilakukan *resampling*



Gambar 9. Distribusi data setelah dilakukan *resampling*

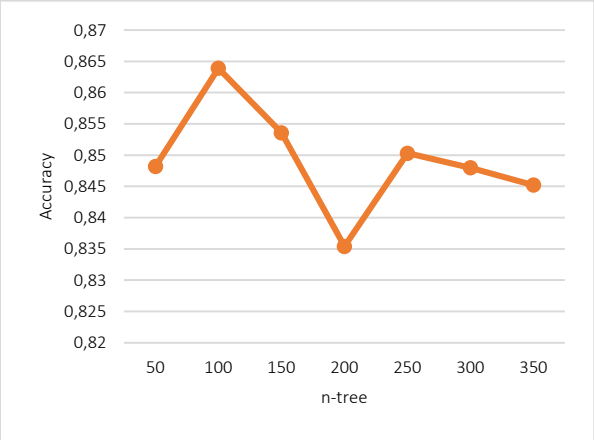
Setelah dilakukan *resampling*, maka proses selanjutnya adalah pembagian *dataset* menjadi *data train* dan *data test*. Pada penelitian ini diterapkan *k-fold cross-validation* dengan 10 bagian dan satu iterasi. Penelitian [18] menganjurkan untuk menggunakan 10-*fold cross-validation* (jumlah *k=10*) karena dinilai memberikan hasil terbaik untuk uji validitas. Selain itu juga dikatakan bahwa semakin kecil iterasi, akan semakin reliabel pula hasil klasifikasi [19]. Dengan demikian, penelitian ini akan menggunakan 10-*folds cross-validation* dengan satu kali iterasi.

4.3 *Modeling dan Validating*

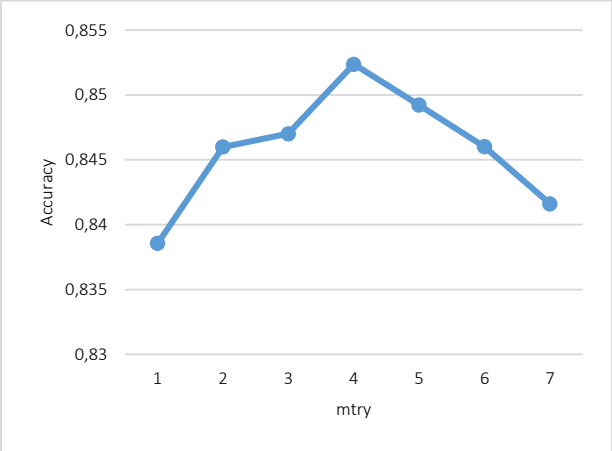
1) Random Forest

Pada model *Random Forest*, dilakukan beberapa kali percobaan dengan jumlah pohon (*ntree*) dan jumlah atribut untuk setiap pohon (*mtry*) yang berbeda-beda. Hasil percobaan menunjukkan bahwa model terbaik dengan akurasi tertinggi diperoleh saat jumlah pohon sebanyak 100 dan jumlah atribut pada setiap pohonnya sebanyak 4. Hasil percobaan untuk menentukan jumlah pohon dan jumlah atribut yang menghasilkan akurasi terbaik ditunjukkan pada Gambar 9 dan Gambar 10.

ISSN:2302-8149
e-ISSN:2540-9719



Gambar 10. Jumlah Pohon dan Akurasi yang Dihasilkan



Gambar 11. Jumlah Mtry dan Akurasi yang Dihasilkan

2) Support Vector Machine

Pada model *Support Vector Machine* (*SVM*), dilakukan percobaan menggunakan kernel linier, radial, dan polynomial. Hasil percobaan menunjukkan bahwa model dengan akurasi terbaik adalah model dengan kernel polinomial, *cost* = 0.5, *scale* = 0.1, dan *degree* = 3.

Tabel 4. Hasil Percobaan Model *SVM*

Kernel	**Akurasi**
Radial	0.8385
Linier	0.8281
Polinomial	0.8461

3) AdaBoost

Pada model *AdaBoost*, dilakukan percobaan menggunakan model *bagging AdaBoost*. Hasil akhir pemodelan dengan akurasi tertinggi menggunakan banyaknya iterasi pada proses *boosting* didapatkan saat parameter *mfinal* = 50 dan *maxdepth* = 3.

Tabel 5. Hasil Percobaan Model *AdaBoost*

mfinal	***maxdepth***	**Akurasi**
1	50	0,7229
1	100	0,7229
1	150	0,7229
2	50	0,7909
2	100	0,7898
2	150	0,7940
3	50	0,8036
3	100	0,8015
3	150	0,7983

4) Perbandingan Random Forest, SVM, dan AdaBoost

Setelah dibangun model dan dihasilkan *classifier* dengan tiga algoritma berbeda, maka langkah selanjutnya adalah membandingkan ketiga model tersebut dengan hasil evaluasi yang dihasilkan.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 6. Hasil Evaluasi Model

Variabel	Akurasi	Sensitivitas	Spesifisitas	*Kappa*
Random Forest	0,8523	0,7163	0,9505	0,7698
SVM	0,8461	0,7090	0,9481	0,7596
AdaBoost	0,8036	0,5820	0,9324	0,6909

Hasil pemodelan menggunakan *Random Forest*, nilai akurasi sebesar 0,8523 menunjukkan bahwa sebesar 85,23 persen kabupaten/kota diklasifikasikan dengan benar sesuai dengan kelas yang sebenarnya. Nilai sensitivitas sebesar 0,7163 menunjukkan bahwa *Random Forest* berhasil mengklasifikasikan 71,63 persen kabupaten/kota dengan IPM yang tergolong sangat tinggi (kelas minor) sebagai kelompok IPM yang tergolong sangat tinggi. Nilai spesifisitas sebesar 0,9505 menunjukkan bahwa *Random Forest* berhasil mengklasifikasikan 95,05 persen kabupaten/kota dengan IPM yang tergolong rendah (kelas mayor) sebagai kelompok IPM yang tergolong rendah. Nilai koefisien *kappa* sebesar 0,7698 berarti hasil proses klasifikasi *Random Forest* sudah cukup baik dan dapat diterima.

Hasil pemodelan menggunakan *SVM*, nilai akurasi sebesar 0,8461 menunjukkan bahwa sebesar 84,61 persen kabupaten/kota diklasifikasikan dengan benar sesuai dengan kelas yang sebenarnya. Nilai sensitivitas sebesar 0,7090 menunjukkan bahwa *SVM* berhasil mengklasifikasikan 70,90 persen kabupaten/kota dengan IPM yang tergolong sangat tinggi (kelas minor) sebagai kelompok IPM yang tergolong sangat tinggi. Nilai spesifisitas sebesar 0,9481 menunjukkan bahwa *SVM* berhasil mengklasifikasikan 94,81 persen kabupaten/kota dengan IPM yang tergolong rendah (kelas mayor) sebagai kelompok IPM yang tergolong rendah. Nilai koefisien *kappa* sebesar 0,7596 berarti hasil proses klasifikasi *SVM* sudah cukup baik dan dapat diterima.

Hasil pemodelan menggunakan *AdaBoost*, nilai akurasi sebesar 0,8036 menunjukkan bahwa sebesar 80,36 persen kabupaten/kota diklasifikasikan dengan benar sesuai dengan kelas yang sebenarnya. Nilai sensitivitas sebesar 0,5820 menunjukkan bahwa *AdaBoost* berhasil mengklasifikasikan 58,20 persen kabupaten/kota dengan IPM yang tergolong sangat tinggi (kelas minor) sebagai kelompok IPM yang tergolong sangat tinggi. Nilai spesifisitas sebesar 0,9324 menunjukkan bahwa *AdaBoost* berhasil mengklasifikasikan 93,24 persen kabupaten/kota dengan IPM yang tergolong rendah (kelas mayor) sebagai kelompok IPM yang tergolong rendah. Nilai koefisien *kappa* sebesar 0,6909 berarti hasil proses klasifikasi *AdaBoost* sudah cukup baik dan dapat diterima.

Dengan menggunakan *10-fold cross-validation*, hasil menunjukkan bahwa *ensemble classifier* yang menunjukkan performa terbaik adalah metode *Random Forest* ditunjukkan dengan nilai akurasi, koefisien *kappa*, sensitivitas, dan spesifisitas yang lebih besar dibandingkan *Support Vector Machine* (*SVM*) dan *AdaBoost*. Oleh karena itu, klasifikasi Indeks Pembangunan Manusia kabupaten/kota di Indonesia dengan metode *Random Forest* memiliki kinerja yang lebih baik daripada *SVM* atau *AdaBoost.* Selain itu, *running time* model *Random Forest* memiliki waktu relatif lebih singkat daripada *SVM* atau *AdaBoost.*

5 Kesimpulan

Berdasarkan pembahasan yang telah dipaparkan, maka beberapa kesimpulan yang dapat diambil adalah sebagai berikut. Rata-rata akurasi dan kappa yang dihasilkan dari ketiga metode memiliki selisih yang kecil. Hal ini menunjukkan bahwa metode klasifikasi radom forest, SVM dan *AdaBoost* sudah cukup baik dan memiliki kinerja yang hampir setara dalam melakukan klasifikasi IPM kabupaten/kota di Indonesia. Metode *Random Forest* mempunyai nilai metrik terbesar diantara dua metode lainnya, hal ini membuktikan bahwa metode ini merupakan model terbaik dalam mengklasifikasikan IPM kabupaten/kota di Indonesia. Selain itu, tahapan data *preprocessing* pada data *imbalance* sangat penting dilakukan terutama jika satuan variabel berbeda-beda. Penerapan *resampling* dan *ensemble method* dapat menangani data *imbalance* terutama pada klasifikasi dengan kelas lebih dari dua (*multiclass*). Secara keseluruhan hasil penelitian ini bahwa dapat dikembangkan

 ISSN:2302-8149
e-ISSN:2540-9719

sebuah *ensemble classifier* untuk membantu memudahkan klasifikasi Indeks Pembangunan Manusia di Indonesia menggunakan variabel-variabel ekonomi yang tersedia.

Saran yang dapat diberikan dari penelitian ini antara lain, untuk pemerintah tingkat di bawah kabupaten/kota, dapat dilakukan klasifikasi indeks pembangunan manusia dengan memanfaatkan salah satu *metode ensemble learning* dan variabel ekonomi yang tersedia untuk menganalisis tingkat pembangunan manusia. Sehingga, dapat diambil kajian untuk meningkatkan pembangunan manusia pada daerah dengan IPM rendah. Untuk penelitian lain, dapat dilakukan eksplorasi variabel ekonomi lainnya yang berhubungan kuat dengan pembangunan manusia agar dapat meningkatkan akurasi dan kinerja klasifikasi.

Referensi

[1] S. S. Pangastusi, "Perbandingan Metode Ensemble Random Forest Dengan Smote-Boosting dan Smoote-Bagging Pada Klasifikasi Data Mining Untuk Kelas Imbalance," Tesis. Departemen Statistika, Insitut Teknologi Sepuluh Nopember, Surabaya, 2018.

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

[3] I. Syarif, E. Zaluska, A. Bennett dan G. Wills, "Application of Bagging, Boosting, and Stacking to Intrusion Detection," *Springer-Verlag Berlin Heidelberg,* vol. 7376, pp. 593-602 , 2012.

[4] T. G. Dietterich, "Ensemble Methods in Machine Learning," *Springer-Verlag Berlin Heidelberg,* vol. 1857, pp. 1-15, 2000.

[5] S. Pramana, B. Yuniarto, S. Mariyah, I. Santoso dan R. Nooraeni, Data Mining Dengan R: Konsep Serta Implementasi, Jakarta: IN MEDIA, 2018.

[6] P. R. Sihombing dan O. P. Hendarsin, "Perbandingan Metode Artificial Neural Network (ANN) dan Support Vector Machine (SVM) untuk Klasifikasi Kinerja Perusahaan Daerah Air Minum (PDAM) di Indonesia," *Jurnal Ilmu Komputer,* vol. XIII, no. 1, pp. 9-20, 2022.

[7] A. Bisri dan R. S. Wahono, "Penerapan Adaboost untuk Penyelesaian Ketidakseimbangan Kelas pada Penentuan Kelulusan Mahasiswa dengan Metode Decision Tree," *Journal of Intelligent Systems,* vol. 1, no. 1, pp. 27-32, 2015.

[8] Q. Iman dan A. W. Wijayanto, "Klasifikasi Rumah Tangga Penerima Beras Miskin (Raskin)/Beras Sejahtera di Provinsi Jawa Barat Tahun 2017 dengan Metode Random Forest dan Support Vector Machine," *Jurnal Sistem dan Teknologi Informasi,* vol. 9, no. 2, pp. 178-184, 2021.

[9] A. A. Nurkhaliza, " Perbandingan Algoritma Klasifikasi Support Vector Machine dan Random Forest pada Prediksi Status Indeks Mitigasi dan Kesiapsiagaan Bencana (IMKB) Satuan Kerja BPS di Indonesia Tahun 2020," *Jurnal Informatika Universitas Pamulang,* vol. 7, no. 1, pp. 54-59, 2022.

[10] A. Nurpiana dan A. W. Wijayanto , "Comparison of Models for Classification of Learning Achievement of Middle School Students in Indonesia in 2019 using the Support Vector Machine Algorithm, Conditional Inference Trees, and Random Forest," *Jurnal Matematika, Statisika & Komputasi,* vol. 18, no. 3, pp. 447-455, 2022.

[11] N. B. Putri dan A. W. Wijayanto, "Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phising," *Jurnal Sistem Komputer,* vol. 11, no. 1, pp. 59-66, 2022.

[12] F. Fauzi, "K-Nearest Neighbor (KNN) dan Support Vector Machine (SVM) untuk Klasifikasi Indeks Pembangunan Manusia Provinsi Jawa Tengah," *Jurnal MIPA,* vol. 40, no. 2, pp. 118-124, 2017.

[13] I. A. A. S. Pratiwi dan A. W. Wijayanto, "Klasifikasi Indeks Pembangunan Manusia dengan Metode K-Nearest Neighbor dan Support Vector Machine di Pulau Jawa," *Jurnal Ilmu Komputer,* vol. 15, no. 1, pp. 8-21, 2022.

[14] M. Fathurrahman dan N. Qisthi, "Klasifikasi Indeks Pembangunan Manusia (IPM) di Pulau

 ISSN:2302-8149
e-ISSN:2540-9719

Sumatera pada Dataset Multiclass Dengan Metode Artificial Neural Network," *Prosiding Seminar Nasional Fisika 7.0*, pp. 377-384, 2021.

[15] N. V. Chawla, A. Lazarevic, L. O. Hall dan Bowyer, "SMOTEBoost: Improving Prediction of The Minority Class in Boosting.," dalam *European Conference on Principles and Practice of Knowledge Discovery*, Dubrovnik, 2003.

[16] J. R. Quinlan, C4.5 : Programs For Machine Learning., San Mateo, California: Morgan Kaufman, 1993.

[17] D. Gujarati, Ekonometrika Dasar, Jakarta: Erlangga, 2006.

[18] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, vol. 2, pp. 1137-1145, 1995.

[19] J. G. Moreno-Torres, J. A. Saez dan F. Herrera, "A Study on the Impact of Partition-Induced Dataset Shift on k-fold Cross-Validation," *IEEE Trans. Neural Network Learn.Syst*, vol. 23, no. 8, pp. 1304-1312, 2012.

[20] O. Steinki dan Z. Mohammad, Introduction to Ensemble Learning, Schwyz: Evolutiq, 2015.

[21] I. Kemala dan A. W. Wijayanto, "Perbandingan Kinerja Metode Bagging dan Non-Bagging Machine Learning pada Klasifikasi Wilayah di Indonesia Menurut Indeks Pembangunan Manusia," *Jurnal Sistem dan Teknologi Informasi*, pp. 269-127, 2021.

[22] M. Y. Darsyah, "Klasifikasi Indeks Pembangunan Manusia (IPM) Dengan Pendekatan K-Nearest Neighbor (K-NN)," *Seminar Nasional Pendidikan, Sains dan Teknologi Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Muhammadiyah Semarang*, pp. 29-35, 2017.

[23] E. Alpaydin, Introduction to Machine Learning Fourth Edition, Cambridge: MIT Press, 2020.

[24] R. A. Wijayanti, M. T. Furqon dan S. Adinugroho, "Penerapan Algoritma Support Vector Machine Terhadap Klasifikasi Tingkat Risiko Pasien Gagal Ginjal," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 10, pp. 3500-3507, 2018.

[25] R. Kusumodestoni dan Sarwido, "Komparasi Support Vector Machines (SVM) dan Neural Network Untuk Mengetahui Tingkat Akurasi Prediksi Tertinggi," *Jurnal Informatika UPGRIS*, vol. 3, no. 1, pp. 1-9, 2017.

[26] Y. M. Hutahaean dan A. W. Wijayanto, "Klasifikasi Rumah Tangga Penerima Subsidi Listrik di Provinsi," *Gorontalo Tahun 2019 dengan Metode K-Nearest Neighbor dan Support Vector Machine*, vol. 10, no. 1, pp. 64-68, 2022.

[27] F. Fauzi, M. Darsyah dan W. Utami, "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*, pp. 88-97, 2017.

[28] Z.-H. Zhou, "Ensemble Learning," *Encyclopedia of Biometrics. Springer*, vol. 7, p. 270–273, 2009.

[29] M. I. Fachruddin, "Perbandingan Random Forest Classification Untuk Deteksi Epilepsi Menggunakan Data Rekaman Electroencephalograph (EEG)," dalam *Skripsi Program Studi S1 Statistika, Institut Teknologi Sepuluh Nopember*, Surabaya, 2015.