JURNAL SAINS DAN INFORMATIKA
RESEARCH OF SCIENCE AND INFORMATIC V8.I2

Vol.08No.02(2022) 77-85 **p-issn : 2459-9549**
http://publikasi.lldikti10.id/index.php/jsi **e-issn : 2502-096X**

Variable Importance Kesehatan dan Pendidikan dalam Pembentukan IPM dengan Algoritme *Machine Learning*

Cahya Alkahfi [a], Zein Rizky Santoso[b] , Anwar Fitrianto[c], Sachnaz Desta Oktarina[d]
Prodi Statistika dan Sains Data, IPB University
Email : [a]cahyaalkahfi@gmail.com, [b]rizkyzein@apps.ipb.ac.id, [c]anwarstat@gmail.com, [d]sachnazdes@gmail.com

Submitted: 31-10-2022, Reviewed: 27-11-2022, Accepted 29-11-2022
http://doi.org/10.22216/jsi.v8i2.1623

Abstract

HDI is an important indicator to measure the achievement of developing the quality of human life in a region. Java Island is the economic center of Indonesia as well as the most populous island in Indonesia, but it still has relatively large disparities in HDI among its regencies and cities. This research aims to determine the factors of health and education infrastructure that have a major influence on HDI in Java Island. The results can be used as input and consideration in policy-making, especially in Java. This research uses the "variable importance" based on 5 machine learning models, namely forward-selection and LASSO using the absolute value of the regression coefficient, as well as the random forest, extra tree, and gradient boosting models using MDI values. The bootstrap technique will be applied to all models to expand the sample space and produce more accurate indicators. The results of the five machine learning models show that the number of doctors and dentists per 1000 population is the factor that most influences HDI scores in Java. Also, the extra tree model provides the best performance based on the smallest RMSE value and shorter intervals than the other models.

*Keywords: M*achine learning, HDI, health, education, variable importance

Abstrak

IPM merupakan indikator penting untuk mengukur keberhasilan pembangunan kualitas hidup manusia pada suatu wilayah. Pulau Jawa merupakan pusat ekonomi serta penduduk terpadat di Indonesia namun antar kabupaten/kota masih menunjukkan ketimpangan nilai IPM yang relatif besar. Penelitian ini bertujuan untuk mengetahui faktor-faktor infrastruktur kesehatan dan pendidikan yang memiliki pengaruh besar terhadap IPM di Pulau Jawa. Hasil penelitian ini dapat digunakan sebagai bahan masukan dan pertimbangan dalam pembuatan kebijakan khususnya di Pulau Jawa. Metode yang digunakan adalah *variable importance* berdasarkan 5 model pembelajaran mesin yaitu *forward-selection* dan LASSO menggunakan nilai absolut koefisien regresi, serta model *random forest*, *extra trees* dan *gradient boosting* menggunakan nilai MDI. Teknik *bootstrap* akan diterapkan pada semua model dengan tujuan untuk memperluas ruang sampel dan menghasilkan indikator yang lebih akurat. Hasil penelitian dari lima model pembelajaran mesin menunjukkan faktor jumlah dokter dan dokter gigi per 1000 penduduk merupakan faktor yang paling mempengaruhi nilai IPM di Pulau Jawa. Sementara itu, model *extra trees* memberikan performa terbaik berdasarkan nilai RMSE yang terkecil serta interval yang lebih pendek dibandingkan model lainnya.

Kata kunci: Pembelajaran mesin, IPM, kesehatan, pendidikan, variabel penting

1. Pendahuluan

Teknologi pada beberapa tahun terakhir telah maju secara signifikan, khususnya dalam bidang *Machine Learning* (ML), yang efektif untuk meminimalkan tenaga kerja manusia. Pada bidang kecerdasan buatan, pembelajaran mesin menggabungkan statistik dan ilmu komputer untuk membuat algoritma yang menjadi lebih efisien ketika diberikan data yang bersangkutan. Pembelajaran mesin adalah studi tentang metode komputasi yang secara otomatis ditingkatkan oleh pengalaman sehingga dapat mengidentifikasi gambar, suara, dan teks.[1]

Pada dasarnya, pembelajaran mesin menggunakan algoritma yang terprogram untuk mempelajari dan mengoptimalkan operasinya dengan menganalisis data input untuk membuat prediksi dalam rentang yang dapat diterima. Apabila memasukkan data yang baru, algoritma ini cenderung membuat prediksi yang akurat. Terdapat tiga kategori utama pada algoritma pembelajaran mesin sesuai dengan tujuannya dan cara

mesin diajarkan. Ketiga kategori tersebut adalah *supervised*, *unsupervised*, dan *semi-supervised.*

Pada pembelajaran mesin *supervised*, solusi yang diinginkan dalam proses pembelajarannya disertakan. Data yang digunakan telah memiliki label dan algoritma akan mempelajari dari pola dari pasangan data dan label tersebut. Algoritma pembelajaran mesin *supervised* dapat diterapkan baik untuk pemodelan klasifikasi maupun regresi.. Pada konteks pemodelan klasifikasi, peubah target berupa kategorik ataupun diskret, sedangkan pada pemodelan regresi peubah target berupa nilai numerik dengan tipe data rasio.

Least Absolute Shrinkage and Selection Operator (LASSO) adalah salah satu metode dari pembelajaran mesin *supervised*. LASSO merupakan metode komputasi dengan menggunakan pemrograman kuadratik yang dapat memerankan prinsip regresi gulud serta melakukan seleksi peubah. Metode LASSO mulai dikenal setelah ditemukannya algoritma LAR pada tahun 2004.[2] Metode lainnya adalah *Forward Selection* yaitu salah satu teknik untuk mereduksi dimensi dataset dengan menghilangkan atribut-atribut yang kurang relevan atau redundan dan menyisakan peubah yang memiliki pengaruh signifikan saja. *Forward Selection* didasarkan pada model Regresi Linear. [3]

Tree-based juga merupakan salah satu dari pembelajaran mesin *supervised* yang melakukan tugas klasifikasi dan regresi dengan membangun struktur seperti pohon untuk menentukan kelas atau nilai variabel target sesuai dengan fitur. *Tree-based* adalah salah satu algoritma pembelajaran mesin yang populer digunakan untuk memprediksi kumpulan data tabular dan spasial. *Tree-based* yang populer digunakan antara lain *Random Forest*, *Gradient Boosting*, dan *Extra Trees*. *Random Forest* merupakan algoritma pembelajaran mesin yang fleksibel dan mudah digunakan yang dapat memperoleh hasil yang baik walaupun tanpa dilakukan *tuning* parameter. Secara singkat, *Random Forest* membangun beberapa pohon keputusan dan menggabungkannya untuk mendapatkan prediksi yang lebih akurat dan stabil. *Gradient Boosting* merupakan algoritma yang dimulai dari menghasilkan pohon klasifikasi awal dan terus menyesuaikan pohon baru melalui minimalisasi fungsi kerugian.[4] *Extra Trees* serupa dengan *Random Forest* yang membangun banyak pohon dan membagi simpul menggunakan subset atribut acak, namun pada *Extra Trees* keacakan tidak berasal dari bootstrap data, melainkan berasal dari pemisahan acak dari semua pengamatan.[5] Pada penelitian ini, akan membandingkan kelima metode *supervised* pada studi kasus mengetahui faktor - faktor yang dapat mempengaruhi nilai Indeks Pembangunan Manusia (IPM) di Pulau Jawa pada 2018.

Keberhasilan pembangunan suatu daerah dapat diukur dengan beberapa parameter, dan yang paling populer saat ini adalah Indeks Pembangunan Manusia (IPM).[6] UNDP menyusun indeks komposit bahwa IPM didasarkan pada tiga indikator : harapan hidup saat lahir, tingkat melek huruf penduduk dewasa dan rata-rata lama sekolah, dan daya beli.

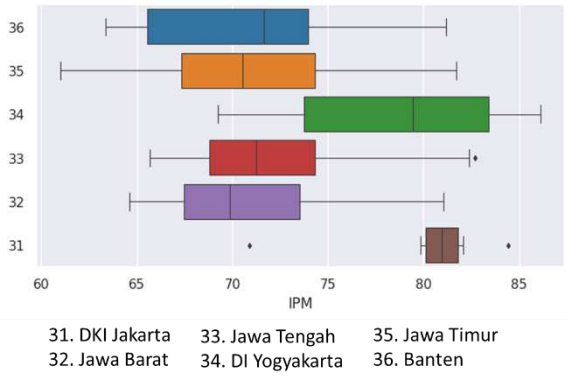
Perhitungan indeks pendidikan meliputi dua indikator yaitu melek huruf dan rata-rata sekolah. Penduduk yang masuk dalam pengukuran adalah penduduk usia 15 tahun ke atas karena pada kenyataannya penduduk usia sudah ada yang putus sekolah. Batasan tersebut diperlukan untuk mencerminkan kondisi yang sebenarnya mengingat penduduk yang berusi kurang dari 15 tahun masih dalam proses bersekolah atau bersekolah sehingga tidak sesuai dengan rata-rata lama sekolah. Kedua indikator pendidikan ini dimunculkan dengan harapan dapat mencerminkan tingkat pengetahuan.

Indeks harapan hidup menunjukkan jumlah tahun hidup yang diharapkan dapat dinikmati oleh penduduk suatu wilayah. Data dasar yang diperlukan dalam metode ini adalah rata-rata anak yang lahir hidup dan rata-rata anak yang masih hidup dari wanita pernah kawin.[7] Angka harapan hidup merupakan alat untuk mengevaluasi kinerja pemerintah dalam meningkatkan kesejahteraan penduduk pada umumnya, dan meningkatkan derajat kesehatan pada khususnya. Apabila ditemukan angka harapan hidup yang rendah di suatu daerah maka pemerintah harus mengadakan lebih banyak program pembangunan, kesehatan, dan program sosial lainnya. Sementara peningkatan angka harapan hidup menunjukkan bahwa bayi-bayi telah terjamin kesehatan dan kemiskinan sudah diatasi lebih baik. Pada indeks kesehatan ini pencapaiannya memerlukan upaya peningkatan terhadap status kesehatan masyarakat, akses dan tenaga kesehatan yang tersedia bagi masyarakat.[8]

Berdasarkan data IPM Indonesia pada tahun 2018 yang dilansir oleh Badan Pusat Statistik (BPS)[9], sebagian besar kota/kabupaten dari Pulau Jawa memiliki nilai IPM dengan status “tinggi” yaitu pada rentang nilai 70 - 80. Pada Gambar 1 dapat dilihat bahwa nilai IPM kabupaten/kota di Pulau Jawa masih cenderung bervariasi baik antar provinsi maupun antar kabupaten/kota pada masing-masing provinsi. Wilayah yang memiliki nilai IPM tertinggi pada tahun 2018 di Pulau Jawa adalah Kota Yogyakarta, DI Yogyakarta dengan nilai 86,11 dan masuk kategori “sangat tinggi”. Sementara itu wilayah dengan IPM terendah adalah Kabupaten Sampang, Jawa Timur yaitu 61,00 atau masuk kategori “sedang’. Nilai ini hanya sedikit saja berapa pada ambang batas IPM kategori “rendah” yaitu 60,00.

Variasi nilai IPM tersebut menunjukkan Pulau Jawa merupakan salah satu wilayah yang berada di Kawasan Barat Indonesia yang didominasi adanya pembangunan infrastruktur. Akan tetapi, antarwilayah di

Pulau Jawa masih memiliki kualitas sumber daya manusia yang bervariasi. Hal tersebut dikarenakan terdapat tidak meratanya pembangunan.[10]



Gambar 1. Sebaran Nilai IPM Kabupaten/Kota Menurut Provinsi di Pulau Jawa Tahun 2018

Berdasarkan latar belakang tersebut akan dibandingkan hasil dari lima metode pembelajaran mesin *supervised*, yaitu *Forward Selection,* LASSO, *Random Forest, Gradient Boosting,* dan *Extra Trees* pada studi kasus untuk mengetahui faktor-faktor infrastruktur kesehatan dan pendidikan di tingkat desa/kelurahan yang mempengaruhi skor IPM kabupaten/kota di Pulau Jawa.

2. Tinjauan Pustaka

Pada penelitian ini akan membandingkan lima metode pembelajaran mesin *supervised* yaitu *Forward Selection,* LASSO, *Random Forest, Gradient Boosting,* dan *Extra Trees.*

2.1 Forward Selection

Metode *Sequential Forward Selection* atau metode seleksi maju adalah algoritma pencarian paling sederhana. *Forward Selection* didasarkan pada model Regresi Linier. *Forward Selection* adalah salah satu teknik untuk mereduksi dimensi dataset dengan menghilangkan atribut-atribut yang tidak relevan atau redundan. Metode *Forward Selection* adalah pemodelan dimulai dari nol peubah (*empty model*), kemudian satu persatu peubah dimasukan sampai kriteria tertentu dipenuhi.[3]

2.2 LASSO

Metode *Least Absolute Shrinkage and Selection Operator* (LASSO) diperkenalkan pertama kali oleh Tibshirani pada tahun 1996. LASSO menyusutkan koefisien regresi dari variabel prediktor yang memiliki korelasi tinggi tinggi dengan galat, menjadi tepat pada nol atau mendekati nol.[11] Menurut Zhao dan Yu [12], persamaan secara umum LASSO dinyatakan sebagai berikut:

$$Y^{**} = X^{**}\beta + \varepsilon^{**} \quad (1)$$

keterangan :

Y^{**} = vektor variabel respon berukuran $(n \times 1)$
X^{**} = matriks variabel prediktor berukuran $(n \times p)$
β = vektor dari koefisien LASSO berukuran $(k+1) \times 1$
ε^{**} = vektor galat berukuran $(n \times 1)$

Menurut Tibshirani [11] estimasi koefisien LASSO menggunakan pemrograman kuadratik dengan kendala pertidaksamaan. Estimasi LASSO diperoleh dari persamaan berikut:

$$\hat{\beta}^{lasso} = argmin\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij} \quad (2)$$

dengan syarat $\sum_{j=1}^{k}|\beta_j| \leq t$. Nilai t merupakan parameter tuning yang mengontrol koefisien LASSO dengan $t \geq 0$. Menurut Tibshirani[11], jika $t < t_0$ dengan $t_0 = \sum_{j=1}^{p}|\hat{\beta}_j|$ maka akan menyebabkan koefisien menyusut mendekati nol atau tepat pada nol atau tepat pada nol, sehingga LASSO akan berperan sebagai seleksi variabel. Akan tetapi jika $t > t_0$ maka penduga koefisien LASSO memberikan hasil yang sama dengan penduga kuadrat terkecil. Koefisien regresi LASSO ditentukan berdasarkan parameter tuning yang sudah dibakukan $s = \frac{t}{\sum_{j=1}^{k}|\hat{\beta}_j^0|}$ dengan $t = \sum_{j=1}^{p}|\hat{\beta}_j|$, $\hat{\beta}_j^0$ adalah penduga kuadrat terkecil untuk model penuh, nilai s optimal diperoleh melalui validasi silang.[13]

2.3 Random Forest

Random Forest merupakan sebuah metode *ensemble*, yang mana metode *ensemble* merupakan cara untuk meningkatkan akurasi metode klasifikasi dengan mengkombinasikan metode klasifikasi dari sebuah pemilah tunggal yang tidak stabil melalui banyak kombinasi penyaringan dari suatu metode yang sama dengan proses keputusan (*voting*) untuk memperoleh prediksi klasifikasi akhir.[14]

Random Forest diawali dengan teknik dasar dari data mining yaitu *decision tree*. Pada proses *decision tree*, dimana *input* berupa data akan dimasukkan pada bagian atas proses *tree* berupa akar pohon (*root*) kemudian akan dibawa turun ke bagian bawah berupa daun pohon (*leaf*) pada proses, untuk menentukan data *input*-an tersebut termasuk ke dalam kelas apa pada proses. Dengan kata lain *Random Forest* terdiri dari sekumpulan *decision tree* (pohon keputusan), dimana kumpulan *decision tree* tersebut digunakan untuk mengklasifikasi data ke suatu kelas.

2.4 Gradient Boosting

Gradient boosting termasuk *supervised learning* berbasis *decision tree* yang dapat digunakan untuk klasifikasi. Algoritma *gradient boosting* bekerja secara sekuensial menambahkan prediktor sebelumnya yang kurang cocok dengan prediksi ke *ensemble*, memastikan kesalahan yang dibuat sebelumnya diperbaiki. Penggambaran sederhana konsep *ensemble* adalah keputusan-keputusan dari berbagai mesin pembelajaran digabungkan, kemudian untuk kelas yang menerima mayoritas 'suara' adalah kelas yang akan diprediksi oleh keseluruhan *ensemble*. *Gradient boosting* dimulai dengan menghasilkan pohon klasifikasi awal dan terus menyesuaikan pohon baru melalui minimalisasi fungsi kerugian :[4]

$$-logL_1 = -\sum_{i-1}^{N} \log(odds) + \log(1 + e^{\log(odds)} \quad (3)$$

2.5 Extra Trees

Extra Trees merupakan model yang diusulkan pertama kali oleh Geurts et al.[15] Model ini merupakan salah satu *tree-based* dari pembelajaran mesin *supervised* yang dapat digunakan untuk regresi dan klasifikasi. *Extra Trees Regression* membangun sebuah *ensemble* dari pohon regresi yang berdasarkan prosedur *top-down* yang sederhana. Perbedaan yang jelas dengan model *ensemble* lainnya adalah *Extra Trees* membangun pohon dengan semua sampel dan memilih titik potong acak untuk setiap fitur yang dipertimbangkan daripada menghitung yang optimal secara local.

Extra Trees memiliki tiga parameter utama, yaitu K, n_{min}, dan M. Ketiga parameter ini dapat ditentukan secara manual maupun otomatis, misalnya menggunakan validasi silang. Parameter K menunjukkan jumlah pemisahan acak dan rentang yang tersedia adalah 1 hingga n, di mana n menunjukkan jumlah atribut. Semakin kecil nilai K, semakin kuat pengacakan pohonnya. Eksperimen telah menunjukkan bahwa nilai optimal K adalah $K = \sqrt{n}$ untuk klasifikasi dan $K = n$ untuk regresi. Parameter n_{min} menunjukkan jumlah sampel untuk membagi sebuah node. n_{min} yang lebih besar mengarah ke bias yang lebih tinggi. Parameter M adalah jumlah pohon. Semakin banyak pohon, semakin baik akurasinya. Ukuran skor dalam *Extra Trees* adalah pengurangan varians relatif. Untuk sampel (S) dan split (s), skor didefinisikan sebagai berikut:

$$Score(s,S) = \frac{var\{y|S\} - \frac{|S_1|}{|S|} var\{y|S_1\} - \frac{|S_r|}{|S|} var\{y|S_r\}}{var\{y|S\}} \quad (4)$$

di mana S_1 dan S_r menyatakan dua himpunan bagian dari kasus dari S yang berkorespondensi dengan dua hasil dari split s.[16]

2.6 Bootstrap Method

Ide dasar dari *bootstrap* yaitu melibatkan pengambilan sampel acak berulang dengan penggantian dari data asli, untuk menghasilkan sampel acak dengan ukuran yang sama dari sampel asli, yang masing-masing dikenal sebagai sampel *bootstrap*, dan masing-masing memberikan parameter yang diinginkan, misalnya *mean* dan *standard deviation*. "Dengan pengembalian" berarti bahwa pengamatan apapun dapat diambil sampelnya lebih dari sekali dalam setiap sampel *bootstrap*. Hal ini penting karena pengambilan sampel tanpa pengembalian hanya akan memberikan permutasi acak dari data asli, dengan banyak statistik seperti rata-rata yang sama persis. Mengulangi proses lebih banyak memberikan informasi yang diperlukan tentang variabilitas estimator, karena kesalahan standar diperkirakan dari *standard deviation* yang berasal dari sampel *bootstrap*.[17]

3. Metodologi Penelitian

3.1 Data

Data yang digunakan adalah data Potensi Desa (PODES) yang dikumpulkan oleh Badan Pusat Statistik (BPS) pada tahun 2018. Pada penelitian ini menggunakan data PODES yang sudah diagregasi pada level kota/kabupaten untuk setiap Provinsi di Pulau Jawa. Provinsi tersebut yaitu DKI Jakarta, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur, dan Banten.

Pada penelitian ini akan dilakukan pencarian *variable importance* pada faktor-faktor infrastruktur kesehatan dan pendidikan yang dapat mempengaruhi skor IPM kota/kabupaten di Provinsi Pulau Jawa. Peubah yang digunakan sebanyak 16 peubah dengan rincian ditunjukkan pada Tabel 1. Seluruh peubah merupakan peubah numerik dengan rincian 15 peubah bebas dan 1 peubah respon yaitu nilai IPM. Di antara peubah bebas, terdapat perbedaan satuan serta interval nilai. Untuk itu akan dilakukan standardisasi agar setiap peubah memiliki interval nilai yang mirip. Adapun jumlah observasi adalah sebanyak 119 atau sebanyak jumlah kota/kabupaten di Pulau Jawa.

Tabel 1. Peubah – peubah yang digunakan pada penelitian ini

Peubah	Deskripsi
IPM	Nilai IPM
PR_NO_LIS	Proporsi keluarga yang tidak menggunakan listrik baik PLN maupun non-PLN
PR_SAMPAH	Proporsi desa/kelurahan pada kabupaten/kota dimana sebagian besar warganya membuang sampah di sungai/saluran irigasi/danau/laut serta got/selokan dan lainnya
PR_TINJA	Proporsi desa/kelurahan pada kabupaten/kota dimana tempat pembuangan akhir tinja sebagian besar warganya adalah sawah/kolam/sungai/danau/laut atau

	pantai/tanah lapang/kebun, lubang tanah dan lainnya
PR_MKM_SUNGAI	Proporsi desa/kelurahan di kabupaten/kota yang memiliki pemukiman di bantaran sungai
PR_SUNGAI_LMBH	Proporsi desa/kelurahan di kabupaten/kota yang memiliki sungai yang tercemar limbah
PR_KUMUH	Proporsi desa/kelurahan di kabupaten/kota yang memiliki pemukiman kumuh
PRA_1000	Jumlah PAUD dan TK per 1000 penduduk
SD_1000	Jumlah SD/MI per 1000 penduduk
SM_1000	Jumlah sekolah menengah (SMP/MTs, SMA/MA, SMK) per 1000 penduduk
RS_PKS_PDK_1000	Jumlah rumah sakit, puskesmas, puskesmas pembantu, poliklinik, praktek dokter per 1000 penduduk
LIN_BID_POS_1000	Jumlah rumah bersalin, praktek bidan, posyandu, polindes per 1000 penduduk
APT_OBT_1000	Jumlah apotek dan toko obat per 1000 penduduk
DOK_DRG_1000	Jumlah dokter dan dokter gigi per 1000 penduduk
BID_1000	Jumlah bidan per 1000 penduduk
GZ_BURUK_1000	Jumlah kejadian gizi buruk per 1000 penduduk

3.2 Metode Penelitian

Pada tahap ini akan melihat hasil dari lima metode pembelajaran mesin *supervised* dalam memperoleh *variable importance* dari faktor – faktor infrastruktur kesehatan dan pendidikan yang mempengaruhi skor IPM pada kota/kabupaten di Provinsi Pulau Jawa. Metode *bootstrap* akan diterapkan pada semua metode pembelajaran mesin, yaitu melakukan penarikan contoh acak dengan pemulihan berukuran n dari gugus data latih. Pembelajaran mesin *supervised* yang digunakan adalah :

a. *Forward Selection*
b. LASSO
c. *Random Forest*
d. *Gradient Boosting*
e. *Extra Trees*

3.3 Penentuan Variable Importance

Pendekatan naif untuk mengukur tingkat *variable importance* (tingkat kepentingan) adalah menghitung berapa kali variabel tersebut muncul dalam kelompok pohon keputusan. Semakin besar dampaknya, semakin penting variabelnya.[16] Pada tahap ini, untuk metode *forward-selection* dan LASSO, penentuan *variable importance* menggunakan nilai absolut koefisien regresi, dimana semakin besar nilai koefisien menunjukkan semakin besar kontribusi peubah yang bersangkutan terhadap nilai IPM untuk setiap perubahan satu satuan peubah tersebut. Adapun untuk *random forest*, *extra trees* dan *gradient boosting* akan menggunakan nilai *Mean Decrease in Impurity* (MDI), yang menunjukkan rata-rata besarnya reduksi keragaman data sebelum dan sesudah dilakukan pembagian kelompok berdasarkan nilai suatu peubah.

Pada penelitian ini, penentuan *variabel importance* tidak hanya menggunakan satu model menggunakan set data tertentu pada setiap motodenya. Penentuannya akan menggunakan metode *bootstrap* dengan jumlah iterasi sebanyak 1000. Adapun tahapan yang dilakukan adalah sebagai berikut:

- Membangkitkan sebanyak 1000 bilangan acak,
- Setiap bilangan acak akan menjadi *seed* pada penarikan sampel dengan teknik *bootstrap* pada data PODES untuk memperoleh 1000 *set data* yang berbeda,
- Membangun model untuk setiap *set data*, sehingga akan menghasilkan 1000 model untuk setiap metode,
- Menghitung nilai koefisien atau MDI setiap model untuk masing-masing metode, dan
- Menentukan *variable importance* pada setiap metode dengan mempertimbangkan nilai rata-rata, median serta interval 95% dari nilai koefisien atau MDI.

4. Hasil dan Pembahasan

Pada bagian ini disajikan hasil dari pembentukan model untuk setiap metode yang digunakan. Hasil yang diperoleh meliputi nilai rataan koefisien atau MDI, median dan selang 95 persen.

4.1 Forward Selection

Koefisien regresi pada model-model *forward-selection* menunjukkan bahwa peubah DOK_DRG_1000 memiliki rataan maupun median yang paling tinggi dibandingkan peubah lainnya. Berdasarkan Tabel 2, dari 5000 model diperoleh nilai rataan koefisien sebesar 2,566 serta median yang relatif serupa yaitu 2,547. Adapun 95 persen nilai koefisien DOK_DRG ini berada pada kisaran nilai 1,465 hingga 3,705. Peubah berikutnya dengan nilai koefisien (absolut) tertinggi yaitu SD_1000 dengan rataan -1,740 dan median -1,747 diikuti PR_SAMPAH dengan rataan -1,020 dan median -1,012. Adapun peubah lainnya walaupun sebagian memiliki nilai yang relatif besar, namun pada interval 95 persen seluruhnya dapat dikatakan tidak begitu konsisten karena memiliki rentang nilai yang berbeda tanda. Sehingga pengaruh peubah tersebut dapat dikatakan tidak signifikan secara statistik.

Tabel 2. Sebaran Nilai Koefisien Peubah pada Model *Forward-Selection*

Peubah	Mean	Median	Q2.5	Q97.5
DOK_DRG_1000	2,566	2,547	1,465	3,705
SD_1000	-1,740	-1,747	-2,731	-0,654
PR_SAMPAH	-1,020	-1,012	-1,656	-0,450

APT_OBT_1000	-0,817	-0,900	-1,674	0,317
PR_TINJA	-0,861	-0,861	-1,816	0,101
LIN_BID_POS_1000	0,597	0,599	-0,368	1,716
RS_PKS_PDK_1000	0,577	0,575	-0,567	1,814
SM_1000	-0,444	-0,437	-1,271	0,255
BID_1000	-0,471	-0,419	-1,691	0,394
PR_KUMUH	0,193	0,185	-0,793	1,075
PR_NO_LIS	-0,214	-0,123	-1,107	0,478
GZ_BURUK_1000	0,079	0,084	-0,388	0,499
PR_MKM_SUNGAI	-0,110	0,000	-1,121	1,031
PR_SUNGAI_LMBH	0,057	0,000	-0,713	0,988

4.2 LASSO

Model LASSO memberikan hasil yang tidak jauh berbeda seperti sebelumnya. Pada Tabel 3, menunjukkan bahwa peubah DOK_DRG_1000 memiliki nilai rataan dan median yang tertinggi yaitu 2,119 dan 2,135. Adapun pada selang 95 persen koefisien peubah ini memiliki nilai berkisar antara 1,038 sampai dengan 3,099. Untuk urutan kedua dan ketiga berturut-turut yaitu peubah SD_1000 dengan rataan -1,466 dan median -1,488 serta peubah PR_SAMPAH dengan rataan -0,986 dan median -0,988. Pada urutan keempat yaitu PR_TINJA dengan rataan -0,817 dan median 0,808. Nilai koefisien 0,000 atau -0,000 menunjukkan nilai koefisien peubah tersebut disusutkan menjadi 0 dalam proses LASSO dan menandakan bahwa pada model bersangkutan peubah tersebut dianggap tidak diperlukan dalam pembentukan model.

Tabel 3. Sebaran Nilai Koefisien Peubah pada Model *LASSO*

Peubah	Rataan	Median	Q2.5	Q97.5
DOK_DRG_1000	2,119	2,135	1,038	3,099
SD_1000	-1,466	-1,488	-2,222	-0,638
PR_SAMPAH	-0,986	-0,988	-1,502	-0,460
PR_TINJA	-0,817	-0,808	-1,650	-0,012
SM_1000	-0,487	-0,493	-1,170	0,000
RS_PKS_PDK_1000	0,289	0,141	0,000	1,179
PR_NO_LIS	-0,188	-0,090	-0,799	0,033
APT_OBT_1000	-0,234	-0,090	-0,851	0,120
PR_KUMUH	0,185	0,071	-0,068	0,785
BID_1000	-0,151	-0,046	-0,669	0,000
PRA_1000	-0,119	-0,000	-0,696	0,135
LIN_BID_POS_1000	0,015	-0,000	-0,330	0,407
PR_SUNGAI_LMBH	0,050	0,000	-0,403	0,596
GZ_BURUK_1000	0,016	0,000	-0,208	0,240
PR_MKM_SUNGAI	0,005	0,000	-0,616	0,732

4.3 Random Forest

Pada penelitian ini, model-model berbasis pohon, termasuk *random forest* diukur berdasarkan nilai MDI, semakin besar nilainya maka semakin tinggi tingkat kepentingan peubah tersebut. Dari hasil pada tabel 4, diperoleh nilai MDI tertinggi ada pada peubah DOK_DRG_1000 dengan rata-rata 0,502 dan median 0,523. Adapun 95 persen nilai MDI untuk peubah ini berkisar antara 0,139 hingga 0,75. Untuk urutan kedua adalah peubah PR_TINJA dengan rata-rata 0,136 dan median 0,091 dan diikuti oleh LIN_BID_POS_1000 serta SD_1000.

Tabel 4. Sebaran Nilai MDI pada Model *Random Forest*

Peubah	Rataan	Median	Q2.5	Q97.5
DOK_DRG_1000	0,502	0,523	0,139	0,750
PR_TINJA	0,136	0,091	0,012	0,500
LIN_BID_POS_1000	0,096	0,055	0,006	0,415
SD_1000	0,065	0,053	0,012	0,186
RS_PKS_PDK_1000	0,050	0,041	0,009	0,127
SM_1000	0,039	0,032	0,010	0,096
BID_1000	0,027	0,019	0,004	0,095
APT_OBT_1000	0,013	0,012	0,004	0,031
PRA_1000	0,012	0,010	0,004	0,035
PR_SAMPAH	0,014	0,009	0,002	0,056
PR_KUMUH	0,010	0,009	0,004	0,026
PR_SUNGAI_LMBH	0,009	0,008	0,003	0,022
GZ_BURUK_1000	0,010	0,008	0,003	0,028
PR_MKM_SUNGAI	0,009	0,008	0,003	0,018
PR_NO_LIS	0,009	0,006	0,002	0,031

4.4 Extra Trees

Berdasarkan Tabel 5, peubah DOK_DRG_1000 memiliki nilai rata-rata dan median MDI tertinggi untuk model *extra trees* yaitu sebesar 0,214 dan 0,207 dengan 95 persen nilainya berada di antara 0,125 hingga 0,339. Pada urutan kedua, peubah PR_TINJA memiliki nilai MDI dengan rata-rata 0,175 dan median 0,169 diikuti dengan LIN_BID_POS_1000 pada urutan ketiga dengan rata-rata 0,129 dan median 0,127. Adapun peubah SD_1000 dan RS_PKS_PDK_1000 memiliki nilai rata-rata dan median tertinggi pada urutan empat dan lima.

Tabel 5. Sebaran Nilai MDI pada Model *Extra Trees*

Peubah	Rataan	Median	Q2.5	Q97.5
DOK_DRG_1000	0,214	0,207	0,125	0,339
PR_TINJA	0,175	0,169	0,074	0,302
LIN_BID_POS_1000	0,129	0,127	0,052	0,221
SD_1000	0,112	0,108	0,048	0,198
RS_PKS_PDK_1000	0,083	0,078	0,039	0,154
PR_SAMPAH	0,075	0,070	0,029	0,145
BID_1000	0,050	0,043	0,016	0,123
SM_1000	0,031	0,029	0,014	0,058
PR_NO_LIS	0,022	0,021	0,009	0,044
PRA_1000	0,022	0,021	0,010	0,044
APT_OBT_1000	0,022	0,020	0,010	0,045
PR_KUMUH	0,020	0,016	0,008	0,057
PR_SUNGAI_LMBH	0,017	0,015	0,008	0,038
PR_MKM_SUNGAI	0,014	0,013	0,007	0,027
GZ_BURUK_1000	0,013	0,011	0,006	0,028

4.5 Gradient Boosting

Tabel 6 menunjukkan sebaran nilai MDI pada model *gradient boosting*. Berdasarkan tabel tersebut peubah DOK_DRG_1000 merupakan peubah dengan nilai rata-rata dan median MDI tertinggi yaitu 0,538 dan 0,605. Peubah tersebut memiliki 95 persen nilai MDI pada rentang nilai antara 0,097 hingga 0,757. Adapun peubah pada urutan kedua jika dilihat berdasarkan nilai rataan maka PR_TINJA memiliki nilai 0,108 yang lebih tinggi dari peubah SD_1000 sebesar 0,083. Namun jika dilihat dari nilai median maka sebaliknya SD_1000 memiliki nilai lebih tinggi dibandingkan PR_TINJA.

Jika dilihat berdasarkan selang nilai 95 persen, maka peubah SD_1000 memiliki selang yang lebih pendek dibandingkan PR_TINJA yaitu antara 0,010 hingga 0,198. Hal ini menunjukkan nilai MDI pada SD_1000 lebih stabil dibandingkan PR_TINJA.

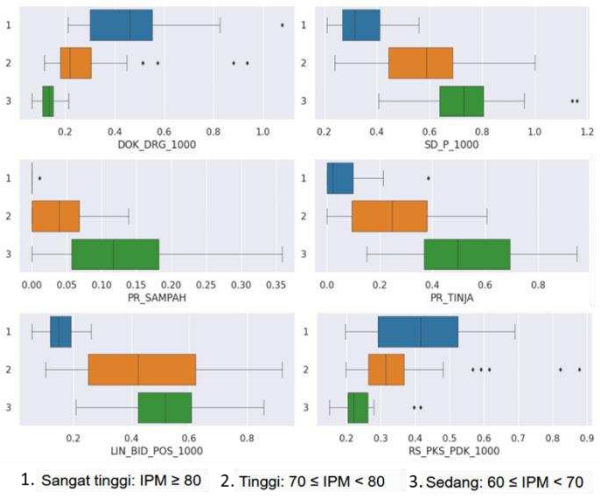
Tabel 6. Sebaran Nilai MDI pada Model *Gradient Boosting*

Peubah	Rataan	Median	Q2.5	Q97.5
DOK_DRG_1000	0,538	0,605	0,097	0,757
SD_1000	0,083	0,076	0,010	0,198
PR_TINJA	0,108	0,051	0,002	0,538
RS_PKS_PDK_1000	0,052	0,044	0,009	0,132
LIN_BID_POS_1000	0,091	0,042	0,001	0,528
SM_1000	0,033	0,027	0,006	0,090
APT_OBT_1000	0,016	0,014	0,002	0,043
BID_1000	0,022	0,013	0,001	0,091
PRA_1000	0,009	0,007	0,001	0,038
PR_SUNGAI_LMBH	0,009	0,007	0,001	0,027
PR_SAMPAH	0,010	0,006	0,000	0,034
GZ_BURUK_1000	0,008	0,006	0,001	0,025
PR_KUMUH	0,007	0,005	0,001	0,025
PR_MKM_SUNGAI	0,007	0,005	0,001	0,023
PR_NO_LIS	0,007	0,004	0,000	0,038

4. 6 Pembahasan

Menurut hasil yang sudah diperoleh pada Tabel 2 hingga Tabel 6, dapat dilihat bahwa peubah DOK_DRG_1000 selalu menempati urutan pertama sebagai peubah dengan koefisien tertinggi atau nilai MDI terbesar. Hal ini menunjukkan peubah tersebut, yaitu jumlah dokter dan dokter gigi per 1000 penduduk merupakan peubah paling penting diantara peubah lainnya sebagai faktor yang paling mempengaruhi besaran nilai IPM di Pulau Jawa.

Selanjutnya, 3 dari 5 model yaitu *forward-selection*, LASSO, dan *gradient boosting* menempatkan SD_1000 sebagai peubah terpenting berikutnya. Adapun pada *random forest* dan *extra tress* peubah SD_1000 juga masih memiliki nilai yang cukup tinggi yaitu masing-masing pada urutan ke-4. Namun yang perlu menjadi catatan bahwa nilai koefisien peubah ini pada model *forward-selection* dan LASSO adalah negatif, yang mana dapat dimaknai bahwa semakin sedikit jumlah SD per 1000 penduduk maka semakin tinggi nilai IPM dan sebaliknya. Pada Gambar 2 dapat dilihat pula bahwa kabupaten/kota dengan jumlah SD per 1000 penduduk yang lebih kecil memiliki kecenderungan nilai IPM yang lebih tinggi dibandingkan kabupaten/kota dengan jumlah SD per 1000 penduduk yang lebih besar.

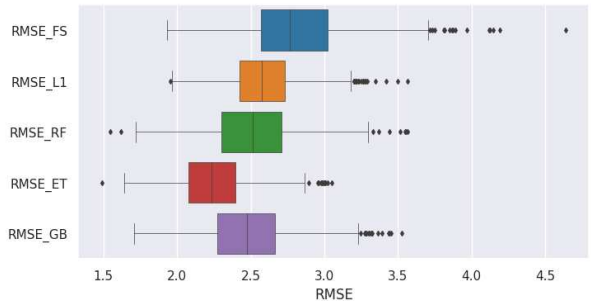


Gambar 2. Boxplot hubungan beberapa peubah penting menurut kategori IPM

Pada urutan ke-3 dan seterusnya, masing-masing metode menetapkan peubah penting yang cukup beragam. Namun, jika dilihat secara keseluruhan, maka peubah PR_TINJA, PR_SAMPAH, LIN_BID_POS_1000 dan RS_PKS_PDK_1000 merupakan peubah-peubah yang cukup penting pada setiap model.

Gambar 2 menunjukkan bagaimana hubungan antara enam peubah penting terhadap nilai IPM. Pada peubah DOK_DRG_1000 memperlihatkan pola dimana semakin besar nilai DOK_DRG_1000 maka kecenderungan nilai IPM juga semakin tinggi. Sebaliknya pada PR_SAMPAH dan PR_TINJA dimana kabupaten/kota dengan nilai yang lebih tinggi cenderung memiliki IPM yang lebih rendah. Adapun anomali yang layak dikaji lebih jauh adalah pada peubah SD_P_1000 dan LIN_BID_POS_1000. Berdasarkan data PODES 2018 memiliki kecenderungan yang kontradiktif. Wilayah dengan jumlah SD per 1000 penduduk yang lebih tinggi atau jumlah rumah bersalin, praktek bidan dan Posyandu yang lebih tinggi memiliki nilai IPM yang cenderung lebih rendah.

Untuk melihat model terbaik pada penelitian ini menggunakan nilai RMSE. Secara keseluruhan pada Gambar 3 menyajikan perbandingan masing-masing metode dimana model *extra trees* memberikan nilai rata-rata RMSE yang terkecil serta interval nilai terpendek. Hasil ini menunjukkan bahwa metode *extra trees* memberikan hasil yang lebih konsisten dibandingkan motode lainnya. Sementara itu model berbasis *forward-selection* memiliki nilai RMSE yang cenderung lebih besar serta interval yang lebar, menunjukkan hasil model yang dapat dikatakan tidak begitu baik dibandingkan model lainnya.



Gambar 3. Perbandingan Sebaran Nilai RMSE Berdasarkan 5 Model

5. Kesimpulan

Berdasarkan hasil yang diperoleh, dapat ditarik kesimpulan bahwa banyaknya jumlah dokter dan dokter gigi per 1000 penduduk merupakan peubah paling penting dalam model untuk menentukan nilai IPM. Peubah lainnya yang menjadi peubah penting adalah jumlah jumlah SD per 1000 penduduk serta proporsi desa yang menjadikan sungai/saluran irigasi/danau/laut serta got/selokan dan lainnya sebagai tempat pembuangan sampah. Adapun model yang dapat dianggap sebagai model terbaik dalam penelitian ini adalah model dengan metode *extra trees*, dimana memiliki nilai RMSE yang cenderung lebih kecil serta interval yang juga lebih pendek dibandingkan model lainnya.

Saran untuk penelitian selanjutnya dapat menggunakan model dengan mempertimbangkan pengaruh geografis serta deret waktu (*spatio-temporal*) pada data IPM. Serta memperluas observasi dengan menambahkan data-data untuk wilayah lainnya.

6. Daftar Rujukan

[1] B. Taha Jijo dan A. Mohsin Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, hlm. 20–28, Mar 2021, doi: 10.38094/jastt20165.

[2] A. Pusporini, "Penerapan Regresi Gulud dan Least Absolute Shrinkage and Selection Operator (LASSO) dalam Penyusutan Koefisien Regresi," 2012.

[3] B. Walczak dan D. L. Massart, "Chapter 15 Calibration in wavelet domain," *Data Handling in Science and Technology*, vol. 22, no. C, hlm. 323–349, 2000, doi: 10.1016/S0922-3487(00)80040-4.

[4] A. Natekin dan A. Knoll, "Gradient boosting machines, a tutorial," *Front Neurorobot*, vol. 7, no. DEC, hlm. 21, 2013, doi: 10.3389/FNBOT.2013.00021/BIBTEX.

[5] I. Tamara, "Kajian Kinerja Algoritme Klasifikasi Extra-Trees pada Permasalahan Data Kelas Tak Seimbang," 2022.

[6] P. Ipm, "Analisis arah kebijakan ekonomi terhadap sektor pendidikan dalam peningkatan ipm," vol. 5, no. 62, hlm. 271–279, 2012, doi: 10.15294/JEJAK.V7I1.3596.

[7] W. W. Lestari dan V. E. Sanar, "Analysis Indicator of Factors Affecting Human Development Index (IPM)," *Geosfera Indonesia*, vol. 2, no. 1, hlm. 11, Apr 2018, doi: 10.19184/GEOSI.V2I1.7333.

[8] S. Sularno, D. Prima Mulya, Z. Zulfahmi, F. Faradika, dan M. Razi A, "Sistem Penunjang Keputusan Pelayanan Kesehatan (Padang Health) dengan Metode AHP (Studi Kasus : Pelayanan Kesehatan untuk Dosen dan Karyawan Universitas Dharma Andalas)," *Jurnal Sains dan Informatika*, vol. 7, no. 2, hlm. 63–72, Nov 2021, doi: 10.22216/JSI.V7I2.724.

[9] "Badan Pusat Statistik." https://www.bps.go.id/pressrelease/2019/04/15/1557/pada-tahun-2018--indeks-pembangunan-manusia--ipm--indonesia-mencapai-71-39.html (diakses Okt 25, 2022).

[10] I. A. A. S. Pratiwi dan A. W. Wijayanto, "Perbandingan Klasifikasi Indeks Pembangunan Manusia (IPM) dengan Metode K-Nearest Neighbour (K-NN) dan Support Vector Machine (SVM) Kabupaten/Kota di Pulau Jawa Tahun 2019," *Jurnal Ilmu Komputer*, vol. 15, no. 1, hlm. 8–21, 2022.

[11] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, hlm. 267–288, Jan 1996, doi: 10.1111/J.2517-6161.1996.TB02080.X.

[12] P. Zhao dan B. B. Edu, "On Model Selection Consistency of Lasso Bin Yu," *Journal of Machine Learning Research*, vol. 7, hlm. 2541–2563, 2006, doi: 10.5555/1248547.1248637.

[13] Y. S. Dewi, "OLS, LASSO dan PLS Pada data Mengandung Multikolinearitas," *Jurnal Ilmu Dasar*, vol. 11, no. 1, hlm. 83–91, Jan 2010.

[14] M. van Wezel dan R. Potharst, "Improved customer choice predictions using ensemble methods," *Eur J Oper Res*, vol. 181, no. 1, hlm. 436–452, Agu 2007, doi: 10.1016/J.EJOR.2006.05.029.

[15] P. Geurts, D. Ernst, dan L. Wehenkel, "Extremely randomized trees," *Mach Learn*, vol. 63, no. 1, hlm. 3–42, Apr 2006, doi: 10.1007/S10994-006-6226-1.

[16] K. Shang *dkk.*, "Fusion of Five Satellite-Derived Products Using Extremely Randomized Trees to Estimate Terrestrial

Latent Heat Flux over Europe," *Remote Sensing 2020, Vol. 12, Page 687*, vol. 12, no. 4, hlm. 687, Feb 2020, doi: 10.3390/RS12040687.

[17] E. Christy, K. Suryowati, J. Statistika, F. Sains Terapan, dan I. AKPRIND Yogyakarta, "Analisis Klasifikasi Status Bekerja Penduduk Daerah Istimewa Yogyakarta Mengunakan Metode Random Forest," *Jurnal Statistika Industri dan Komputasi*, vol. 6, no. 01, hlm. 69–76, Jan 2021.