JUSTIN
Jurnal Sistem dan Teknologi Informasi
p-ISSN : 2460-3562 / e-ISSN : 2620-8989

DOI: 10.26418/justin.v9i2.44166
Vol. 09, No. 2, April 2021

Perbandingan Kinerja Metode Bagging dan Non-Ensemble Machine Learning pada Klasifikasi Wilayah di Indonesia menurut Indeks Pembangunan Manusia

Intan Kemala[a1], Arie Wahyu Wijayanto[a2]
[a]*Program Studi Komputasi Statistik, Politeknik Statistika STIS*
Jl. Otto Iskandardinata No. 64C, Jakarta, Indonesia
[1]221810343@stis.ac.id
[2]ariewahyu@stis.ac.id

Abstrak

Indeks Pembangunan Manusia (IPM) sendiri merupakan suatu ukuran yang digunakan untuk mengukur pencapaian pembangunan manusia di suatu wilayah. Capaian tinggi rendahnya nilai IPM di Indonesia tidak terlepas dari program pembangunan yang dilaksanakan pemerintah di tiap wilayah Indonesia baik itu tingkat pusat maupun daerah. Penentuan program pembangunan yang dilakasanakan harus tepat sasaran dan sesuai dengan prioritas daerah berdasarkan kategori IPM yang dimiliki. Untuk membantu efektifitas kinerja pemerintah dalam menganalisis kategori IPM masing-masing daerah di Indonesia, dilakukan penelitian terkait klasifikasi Indeks Pembangunan Manusia di Indonesia menggunakan 4 indikator berbeda yakni Indeks Pemberdayaan Gender, Indeks Keparahan Kemiskinan, Rata-Rata Lama Sekolah, dan Pengeluaran per Kapita dengan menerapkan beberapa algoritma klasifikasi sebagai perbandingan antara lain *Random Forest* untuk mewakili metode *Bagging*, serta *C4.5 Decision Tree, K Nearest Neighbors,* dan *Naive Bayes* yang mewakili metode *non-ensemble*. Proses klasifikasi diawali dengan pengumpulan data yang sudah tersedia di website BPS; *preprocessing* data berupa *feature selection*, *cleaning*, integrasi, dan transformasi; dilanjutkan proses pembuatan model pada data *training* dengan menerapkan *10-fold cross validation* serta metode *SMOTE* untuk menangani imbalance class; dan terakhir evaluasi hasil pemodelan pada data testing. Dari hasil pemodelan dan *testing* didapat hasil bahwa metode *Random Forest* dengan *mtry* = 2 dan *ntree* = 500 merupakan metode terbaik diantara metode lainnya dimana akurasi klasifikasi yang dihasilkan sebesar 95.41% dan nilai Kappa sebesar 83.46%.

Kata kunci: Pembangunan Manusia, Klasifikasi, Preprocessing, Random Forest, Akurasi

Comparison of the Performance of Bagging Methods and Non-Ensemble Machine Learning on Regional Classification in Indonesia according to the Human Development Index

Abstract

The Human Development Index (HDI) itself is a measure used to measure the achievement of human development in an area. The high and low HDI scores in Indonesia cannot be separated from the development programs implemented by the government in each region of Indonesia, both at the central and regional levels. Determination of the development program that is carried out must be right on target and in accordance with regional priorities based on the HDI category owned. To help the effectiveness of government performance in analyzing the HDI category of each region in Indonesia, research was carried out related to the classification of the Human Development Index in Indonesia using 4 different indicators, namely the Gender Empowerment Index, the Poverty Severity Index, the Average Length of Schooling, and the Per Capita Expenditure. Several classification algorithms for comparison include Random Forest to represent the Bagging method, as well as C4.5 Decision Tree, K Nearest Neighbors, and Naive Bayes which represent non-ensemble methods. The classification process begins with collecting data that is already available on the BPS website; preprocessing data in the form of feature selection, cleaning, integration, and transformation; continued with the process of modeling the training data by applying 10-fold cross validation and the SMOTE method to handle imbalance classes; and

finally evaluation of modeling results on testing data. From the modeling and testing results, the results show that the Random Forest method with mtry = 2 and ntree = 500 is the best method among other methods where the resulting classification accuracy is 95.41% and the Kappa value is 83.46%.

Keywords: Human Development Index, Classification, Preprocessing, Random Forest, Accuracy

I. PENDAHULUAN

Menurut UNDP, pembangunan manusia dirumuskan sebagai upaya perluasan pilihan bagi penduduk (*enlarging the choices of people*) dan sekaligus sebagai taraf yang dicapai dari upaya tersebut. "Perluasan pilihan" hanya mungkin dapat direalisasikan jika penduduk paling tidak memiliki: peluang berumur panjang dan sehat, pengetahuan dan keterampilan yang memadai, serta peluang untuk merealisasikan pengetahuan yang dimiliki dalam kegiatan yang produktif. Indeks Pembangunan Manusia (IPM) sendiri merupakan suatu ukuran yang digunakan untuk mengukur pencapaian pembangunan manusia di suatu wilayah [2].

Pencapaian pembangunan manusia diukur dengan memperhatikan tiga aspek esensial, yaitu umur panjang dan sehat, pengetahuan, dan standar hidup layak. Indikator Angka Harapan Hidup (AHH) merepresentasikan aspek umur panjang dan sehat. Aspek pendidikan pada IPM dicerminkan oleh indikator Angka Melek Huruf (AMH) dan Rata-rata Lama Sekolah (MYS). Aspek terakhir standar hidup layak yang direpresentasikan melalui indikator pengeluaran per kapita per tahun yang disesuaikan [2]. Klasifikasi IPM sendiri dikategorikan menjadi kategori sangat tinggi (IPM ≥ 80), kategori tinggi (70 ≤ IPM < 80), kategori menengah bawah (60 ≤ IPM < 70), dan kategori rendah (IPM < 60) [3]. Namun, aspek sosial budaya terkait gender yang dapat direpresentasikan dengan Indeks Pemberdayaan Gender dan aspek kemiskinan yang dapat direpresentasikan dengan Indeks Keparahan Kemiskinan belum ikut menjadi indikator penentu IPM suatu negara. Padahal masalah kemiskinan dan kesetaraan gender sudah menjadi isu krusial dalam masyarakat.

Dilain hal, capaian tinggi rendahnya nilai IPM di Indonesia tidak terlepas dari program pembangunan yang dilaksanakan pemerintah di tiap wilayah Indonesia baik itu tingkat pusat maupun daerah. Dikarenakan pembangunan manusia tergolong perubahan yang dapat dilihat hasilnya pada jangka waktu panjang maka program pembangunan harus dilaksanakan secara berkesinambungan dan terpantau. Selain itu Penentuan program pembangunan yang dilakasanakan harus tepat sasaran dan sesuai dengan prioritas daerah berdasarkan kategori IPM yang dimiliki daerah tersebut. Untuk membantu efektifitas kinerja pemerintah dalam menganalisis kategori IPM masing-masing daerah di Indonesia perlu adanya suatu sistem keputusan yang dapat menentukan klasifikasi kategori IPM di masing-masing daerah secara cepat dan akurat. Dikarenakan hal inilah, peneliti ingin melakukan penelitian terkait klasifikasi Indeks Pembangunan Manusia menggunakan 4 indikator berbeda yakni Indeks Pemberdayaan Gender, Indeks Keparahan Kemiskinan, Rata-Rata Lama Sekolah, dan Pengeluaran per Kapita Disesuaikan dengan menerapkan beberapa algoritma klasifikasi sebagai perbandingan antara lain *Random Forest* untuk mewakili metode *Bagging*, serta *Decision Tree*, *K Nearest Neighbors*, dan *Naive Bayes* yang mewakili metode *non-ensemble*. Diharapkan penelitian ini dapat menghasilkan informasi terkait algoritma terbaik untuk mengklasifikasi Indeks Pembangunan Manusia di Indonesia.

II. TINJAUAN PUSTAKA

A. Penelitian Terdahulu

Beberapa penelitian terkait klasifikasi IPM di Indonesia yang pernah dilakukan yakni oleh Moh Yamin (2020) dengan judul *Klasifikasi Indeks Pembangunan Manusia (IPM) dengan Pendekatan K Nearest Neighbors (KNN)* dan menghasilkan akurasi klasifikasi sebesar 91.43% [4]. Fatkhurokhman Fauzi dkk (2017) dengan judul *Klasifikasi Indeks Pembangunan Manusia Kabupaten/Kota Se-Indonesia dengan Pendekatan Smooth Support Vector Machine (SSVM) Kernel Radial Basis Function (RBF)* dan didapat hasil akurasi sebesar 100% [5]. Fergie Joanda (2018) dengan judul *Penerapan Algoritma J48 Decision Tree untuk Analisis Tingkat Kemiskinan di Indonesia* dan menghasilkan akurasi sebesar 88.6% [7]. Erfan Karyadiputra (2018) dengan judul *Analisis Algoritma Naive Bayes untuk Klasifikasi Status Kesejahteraan Rumah Tangga Keluarga Binaan Sosial* dan didapat hasil akurasi sebesar 85.8% [8].

B. Kerangka Teori

1) Klasifikasi : suatu proses yang bertujuan untuk menentukan suatu obyek kedalam suatu kelas atau kategori yang sudah ditentukan sebelumnya. Menurut (Elly Susilowati, 2015) klasifikasi adalah proses dari pembangunan terhadap suatu model yang mengklasifikan suatu objek sesuai dengan atribut – atributnya. Metode klasifikasi ditujukan untuk pembelajaran fungsi-fungsi berbeda yang memetakan masing-masing data terpilih kedalam salah satu dari kelompok kelas yang telah ditetapkan sebelumnya [6].

2) Random Forest : algoritma Random Forest didesain oleh J. Ross Quinlan, dinamakan Random Forest karena merupakan keturunan dari pendekatan ID3 untuk membangun pohon keputusan. Random Forest merupakan metode pohon gabungan yang berasal dari metode Classification and Regression Tree (CART) dan didasarkan pada teknik pohon keputusan sehingga dapat diterapkan untuk data nonlinier. Proses klasifikasi Random Forest diawali dengan memecah data sampel yang ada kedalam decision tree secara acak, selanjutnya dilakukan voting pada setiap kelas dan terakhir mengkombinasikan vote dari setiap kelas kemudian diambil vote terbanyak. Pembangunan pohon pada Random Forest dilakukan dengan penerapan metode Random Feature Selection untuk meminimalisir kesalahan.

Tahapan algoritma *Random Forest*:

1. Pilih secara acak "k" fitur dari total "m" fitur. Dimana k <= m

2. Diantara "k" fitur, hitung node "d" menggunakan titik pisah terbaik.
3. Pisahkan node menjadi node anak menggunakan pemisahan terbaik.
4. Ulangi langkah 1 sampai 3 hingga jumlah "l" dari node telah tercapai.
5. Bangun hutan/*forest* dengan mengulangi langkah 1 sampai 4 untuk jumlah "n" kali untuk membuat "n" pohon.

3) Decision Tree : model prediksi menggunakan struktur pohon atau struktur hierarki. Manfaat utama Decision Tree adalah kemampuannya untuk mem-*break down* proses pengambilan keputusan yang kompleks menjadi lebih sederhana serta mampu mengeliminasi perhitungan atau data-data yang tidak diperlukan. Nama lain Decision Tree yakni *Classification and Regression Tree*, dimana metode ini merupakan gabungan dari dua jenis pohon yakni *classification tree* dan *regression tree*.

Salah satu algoritma *Decision Tree* yakni C4.5 yang bekerja dengan tahapan tahapan sebagai berikut [11]

1. Pilih nilai attribute dasar
2. Untuk setiap attribute A, cari gaint rasio informasi yang dinormalisasi dari pemisahan pada A.
3. Buat simpul keputusan
4. Ulangi kembali pada daftar yang diperoleh dengan memisahkan pada a_best, dan tambahkan simpul tersebut sebagai anak-anak simpul.

4) K Nearest Neighbors : suatu metode klasifikasi yang terdapat dalam data mining selain Support Vector Machine (SVM).kNN dilakukan dengan mencari dengan mencari k objek dalam data training yang paling dekat (mirip) dengan objek pada data testing (Wu, 2009)[yamin]. Algoritma KNN hanya bisa digunakan pada data bertipe numerik. Adapun cara untuk mengukur kedekatan antara data baru dengan data yang lama (data trining), diantaranya Euclidean distance, mahattan distance, Hamming Distance, dan Minkowski distance. Adapun tahapan algoritma KNN yakni:

1. Tentukan bilangan k bulat positif yang menunjukkan berapa jangkauan tetangga terdekat yang digunakan oleh objek, dapat genap atau ganjil namun disarankan ganjil
2. Pilih tetangga terdekat dari tetangga baru sebanyak k
3. Hitung kuadrat jarak misal euclidien objek terhadap data training yang diberikan
4. Mengurutkan hasil nomor 3 secara Ascending
5. Menggunakan kategori nearest neighbor yang paling mayoritas untuk memprediksi kategori objek.

5) Naive Bayes : pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Theorema Bayes. Teorema tersebut dikombinasikan dengan *naive* dimana diasumsikan konsdisi antar atribut saling bebas. Klasifikasi naive bayes mengasumsikan bahwa ada tidaknya ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya.

Persamaan 1. Teorema *Bayes* :

$$P(C|X) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

Keterangan :
x : Data dengan class yang belum diketahui
c : Hipotesis data merupakan suatu class spesifik
P(c|x) : Probabilitas hipotesis berdasar kondisi (posteriori probability)
P(c) : Probabilitas hipotesis (prior probability)
P(x|c) : Probabilitas berdasarkan kondisi pada hipotesis
P(x) : Probabilitas c

Penjabaran lebih lanjut rumus *Bayes* tersebut dilakukan dengan menjabarkan *(c|x1,…,xn)* menggunakan aturan perkalian pada persamaan 2 sebagai berikut :

$$\begin{aligned} P(C|X1,\dots,Xn) &= P(C)P(X1,\dots,Xn|C) \\ &= P(C)P(X1|c)(X2,\dots,Xn|C,X1) \\ &= P(C)P(X1|c)P(X2|C,X1)(X3,\dots,Xn|C,X1,X2) \\ &= P(C)P(X1|c)P(X2|C,X1)P(X3|C,X1,X2)\dots P(Xn|C,X1,X2,\dots,Xn-1) \end{aligned} \quad (2)$$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya faktor-faktor syarat yang mempengaruhi nilai probabilitas, yang hampir mustahil untuk dianalisa satu persatu. Akibatnya, perhitungan tersebut menjadi sulit untuk dilakukan. Disinilah digunakan asumsi independensi yang sangat tinggi (naif), bahwa masing masing petunjuk saling bebas (independen) satu sama lain. Dengan asumsi tersebut, maka berlaku suatu kesamaan sebagai berikut:

$$P(c|X1,\dots,Xn) = P(C)\prod\prod_{i=1}^{n} P(Xi|C)$$

$$P(c|X) = P(x_1|c)P(x_2|c)\dots P(x_n|c)P(c) \quad (3)$$

Persamaan 3 diatas merupakan model dari ***Teorema Naive Bayes*** yang selanjutnya akan digunakan dalam proses klasifikasi. Secara umum, tahapan algoritma *Naive Bayes* antara lain :[20]

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

C. Metode Penelitian

Metode penelitian yang digunakan terlihat pada Gambar 1. Adapun proses analisis dan klasifikasi data menggunakan bantuan software RStudio versi 4.0.2. Penjelasan dari setiap tahapan metode sebagai berikut :

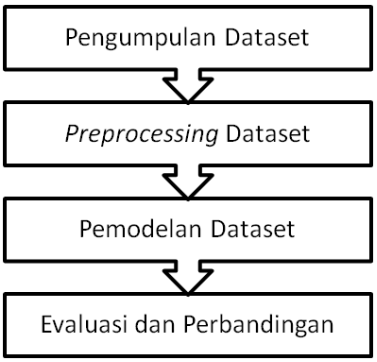
1) *Pengumpulan dataset* : proses pengambilan data dari sumber data untuk melanjutkan proses pengolahan data ke tingkat selanjutnya

2) *Preprocessing dataset* : proses yang harus dilakukan dalam data mining karna data yang didapat tidak

semuanya bisa langsung digunakan dalam proses data mining. proses penyiapan dataset dlakukan hingga data siap digunakan sesuai kebutuhan

3) *Pemodelan dataset* : proses penerapan algoritma data mining baik itu berupa klustering, klasifikasi, ataupun prediksi pada dataset yang telah melalui *preprocessing* data

4) *Evaluasi dan Perbandingan* : proses evaluasi dari model yang didapat pada pemodelan dataset dan melakukan perbandingan terhadap model-model tersebut



Gambar 1. Tahapan metode penelitian

D. Spesifikasi Hardware dan Software

Penelitian ini dilakukan dengan menggunakan satu set komputer/laptop dengan prosesor Intel Celeron N2840 berkecepatan 2.16 GHz dan RAM 2 GB. Dari sisi software, sistem operasi yang dipakai adalah Microsoft Windows 8.1 64 bit. Dan software yang digunakan untuk melakukan analisisi adalah Rstudio versi 4.0.2.

III. HASIL DAN PEMBAHASAN

A. Pengumpulan Dataset

Dataset yang digunakan pada penelitian ini antara lain data Indeks Pemberdayaan Gender 2017-2019, Indeks Keparahan Kemiskinan 2018-2020, Rata-Rata Lama Sekolah 2018-2020, Pengeluaran Perkapita Disesuaikan 2018-2020, dan Indeks Pembangunan Manusia Menurut Kabupaten/Kota 1996-2013. Semua data diunduh dari website BPS melalui link www.bps.go.id , dalam format xls dan xlsx. Masing-masing dataset terdiri dari 554 observasi dan 4 variabel. Tipe Data variabel pendukung dapat dilihat Tabel I.

TABEL I
VARIABEL PENDUKUNG

No	Variabel	Tipe Data
1	Indeks Pemberdayaan Gender	Numerik
2	Indeks Keparahan Kemiskinan	Numerik
3	Rata-rata Lama Sekolah	Numerik
4	Pengeluaran per Kapita Disesuaikan	Numerik
5	Indeks Pembangunan Manusia	Numerik

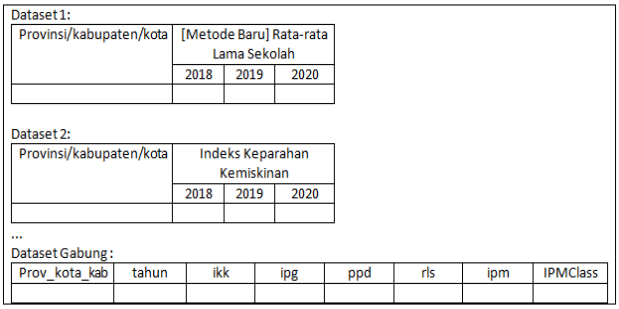
B. Preprocessing Dataset

Langkah-langkah yang dilakukan pada preprocessing dataset dalam penelitian ini antara lain :

1) Feature Selection / pemilihan fitur : proses pemilihan atribut data yang akan digunakan pada pemodelan nanti. Dalam penelitian ini, peneliti memutuskan untuk menggunakan data dari tahun 2018-2019 (2 tahun terakhir) pada masing-masing dataset, setelah mempertimbangkan ketersediaan data terbaru dan kesesuaian tahun pada kelima dataset. Selain itu dilakukan ekstraksi variabel IPM untuk membuat variabel kelas IPM sebagai target kelas klasifikasi pada dataset akhir/gabungan.

2) Cleaning / pembersihan : proses mengatasi noise dan data yang tidak relevan/inkonsisten. Dalam hal ini, terdapat penghilangan 3 observasi dimana 2 diantaranya memiliki missing value di semua kolom tahunnya dan 1 observasi dihilangkan (kecuali data Indeks Keparahan Kemiskinan) guna menyesuaikan record di masing-masing data.

3) Integration / integrasi : proses penggabungan data dari sumber dataset berbeda. Dalam penelitian ini, kelima dataset yang diambil digabung menjadi satu dataset utuh untuk selanjutnya dipakai dalam pemodelan. Hasil proses penggabungan ditunjukkan pada Gambar 2 dan penjelasan variabel dataset gabungan dapat dilihat pada Gambar 3.



Gambar 2. Proses integrasi dataset

Variabel	Tipe Data	Nilai	Deskripsi
Prov_kota_kab	Karakter	String	Nama provinsi, kabupaten, dan kota di Indonesia
tahun	Karakter	String	tahun
Ikk	Numerik	Real	Indeks Keparahan Kemiskinan
Ipg	Numerik	Real	Indeks Pemberdayaan Gender
ppd	Numerik	Real	Pengeluaran per Kapita Disesuaikan dalam ribu rupiah
rls	Numerik	Real	Rata-rata lama sekolah dalam tahun
ipm	Numerik	Real	Indeks Pembangunan Manusia
IPMClass	Kategorik	Rendah, Sedang, Tinggi, Sangat tinggi	Kategori nilai Indeks Pembangunan Manusia

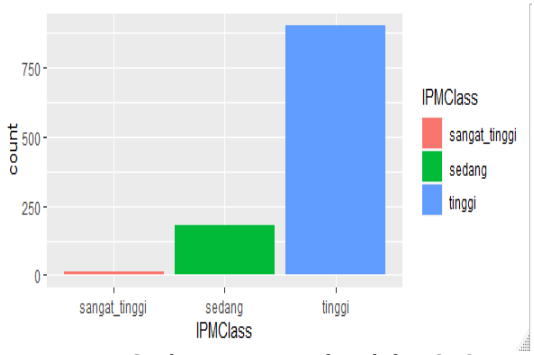
Gambar 3. Penjelasan variabel dataset gabungan

4) Transformation / transformasi : proses mengubah suatu nilai data agar didapat data dengan rentang nilai yang sama dalam satu dataset. Dalam hal ini, feature *scalling* yang dilakukan menggunakan metode *standardisation*.

C. Pemodelan Dataset

1) Data Training dan Testing: Dataset hasil integrasi terdiri dari 1096 observasi dengan 8 variabel termasuk variabel target kelas. Pembagian data Training dan data Testing menggunakan proporsi 7:3 dengan pendekatan *random split*. Metode 10-cross validation juga diterapkan pada data *Training* untuk membagi data menjadi *training* dan *validation*. Lima variabel utama untuk proses klasifikasi yakni ikk, ipg, rls, ppd, dan IPMClass. *Data Training* terdiri dari 767 observasi sedangkan *data Testing* sebanyak 329 observasi.

2) Kelas Tidak Seimbang (Imbalance Class): Berdasarkan Gambar 4 terlihat bahwa proporsi masing-masing kelas target pada data Training tidak seimbang, terutama antara kelas IPM tinggi dan sangat tinggi. Hal ini dapat berpengaruh pada keakuratan dan reliabilitas model klasifikasi yang dibangun nantinya. Metode yang peneliti gunakan untuk menangani ketidakseimbangan kelas yakni metode SMOTE, dengan memanfaatkan package DMwR pada RStudio. Data hasil SMOTE terdiri dari 2440 observasi dengan 5 variabel (lihat Gambar 5).

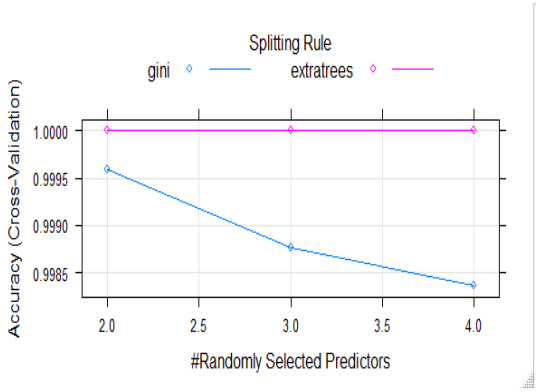


Gambar 4. Proporsi Kelas sebelum *SMOTE*

Gambar 5. Proporsi Kelas setelah *SMOTE*

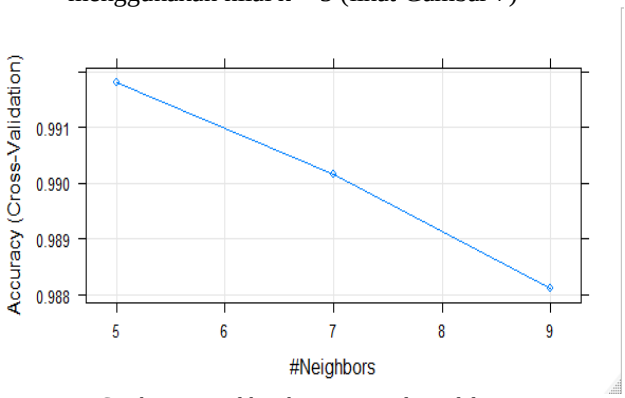
3) Pembuatan Model: Proses klasifikasi yang dilakukan dalam penelitian ini menggunakan 4 algoritma klasifikasi yakni Random Forest, K Nearest Neighbors, Decision Tree C4.5, dan Naive Bayes. Pengujian klasifikasi juga menggunakan konsep k-fold cross validation dengan k=10. Pemilihan model terbaik masing-masing algoritma memperhatikan nilai akurasi yang dihasilkan. Berikut model terbaik dari masing-masing algoritma :

1. Algoritma *Random Forest*, model terbaik dihasilkan pada saat *mtry = 2* , *splitrule = extratrees*, dan *ntree = 500* (lihat Gambar 6)



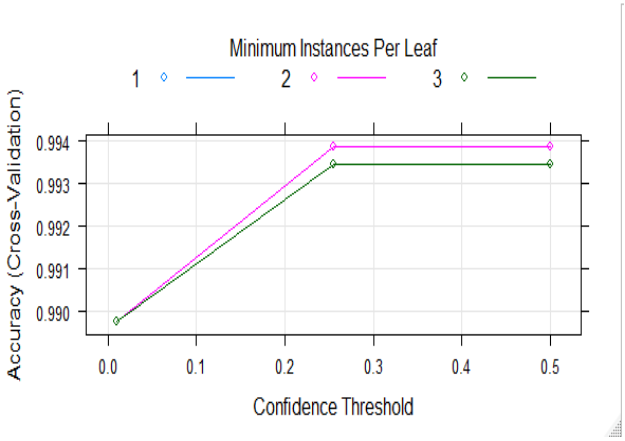
Gambar 6. Pemilihan nilai *mtry* dan *splitting rule* pada model *random forest*

2. Algoritma *K Nearest Neighbors*, model terbaik menggunakan nilai k = 5 (lihat Gambar 7)

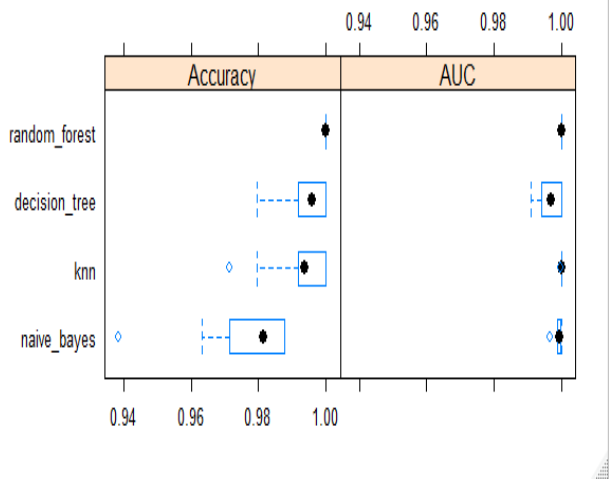


Gambar 7. Pemilihan *k* tetangga pada model KNN

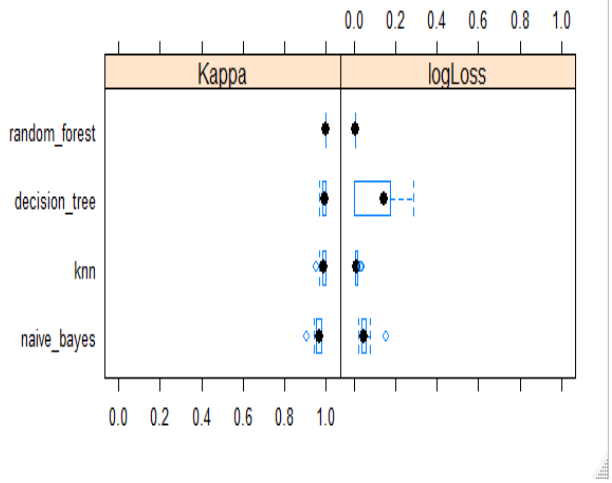
3. Algoritma *Naive Bayes*, model terbaik disaat *usekernel TRUE, laplace = 0,* dan *adjust = 1.*
4. Algoritma *Decision Tree C4.5*, model terbaik dihasilkan saat C = 0.255 dan M = 1. (lihat Gambar 8)



Gambar 8. Pemilihan nilai C (*Confidence Threshold*) dan M (*Minimum Instances per Leaf*) pada model *Decision Tree C4.5*



Gambar 9. Perbandingan nilai Akurasi dan AUC Hasil tiap Pemodelan



Gambar 10. Perbandingan nilai *Kappa* dan *logLoss* Hasil tiap Pemodelan

Berdasarkan gambar 9 dan 10 dapat dilihat bahwa model dengan nilai akurasi, *kappa*, dan *AUC* terbesar, serta nilai *logLoss* terkecil dimiliki oleh model klasifikasi dengan algoritma *Random Forest* dengan akurasi model sebesar 0.1 atau 100 %, diikuti oleh *KNN* dengan akurasi model sebesar 0.9918 atau 99.18%, *Decision Tree C4.5* dengan akurasi model sebesar 0.9938 atau 99.38 %, dan *Naive Bayes* dengan akurasi model sebesar 0.9758 atau 97.58%.

4) Evaluasi dan Perbandinga: Setelah melakukan pemodelan data Training, proses selanjutnya yakni prediksi data Testing menggunakan model yang dihasilkan pada masing-masing algoritma. Perbandingan Hasil prediksi tiap model dapat dilihat pada Tabel II.

Berdasarkan tabel II terlihat bahwa hasil prediksi dengan model *Random Forest* merupakan yang terbaik dengan nilai akurasi sebesar 95.14%, diikuti prediksi dengan model *KNN* sebesar 94.22%, prediksi dengan model *Decision tree C4.5* sebesar 93.01%, dan terakhir prediksi dengan model *Naive Bayes* sebesar 86.02%.

TABEL II
PERBANDINGAN HASIL PREDIKSI TIAP MODEL

Algoritma	Akurasi	Kappa	95%CI
Random Forest	0.9514	0.8346	0.9222,0.972
KNN	0.9422	0.7967	0.9113, 0.9649
Decision Tree C4.5	0.9301	0.7679	0.8969, 0.9552
Naive Bayes	0.8602	0.5953	0.8179, 0.8958

IV. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan dapat ditarik kesimpulan, bahwa :

1) Proses *Preprocessing* pada dataset penting untuk dilakukan terutama yang berkaitan dengan proses *scalling*, dan penanganan *imbalance class* pada kasus *multiclass.*

2) Berdasarkan perbandingan akurasi pada masing-masing model, proses klasifikasi Indeks Pembangunan Manusia di Indonesia dengan algoritma *Random Forest* merupakan yang terbaik dengan akurasi sebesar 95.14% dan nilai Kappa sebesar 0.8346.

3) Hasil penelitian ini dapat digunakan sebagai salah satu pertimbangan dalam menganalisis data Indeks Pembangunan Manusia di Indonesia, sehingga dapat mempermudah pemerintah dalam merumuskan berbagai kebijakan terkait pembangunan ekonomi dan sosial budaya.

DAFTAR PUSTAKA

[1] Annur, Haditsah. 2018. *Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes*. ILKOM Jurnal Ilmiah, vol. 10, No. 2, hlm 160-165.

[2] BPS. 2013. *Indeks Pembangunan Manusia 2013*. Jakarta: BPS.

[3] ___. 2014. Indeks Pembangunan Manusia 2014 Metode Baru. Jakarta: BPS.

[4] Darsyah, Moh. Yamin. 2020. *Klasifikasi Indeks Pembangunan Manusia (IPM) Dengan Pendekatan K-Nearset Neighbor (K-NN).* Seminar Nasional Pendidikan, Sains dan Teknologi Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Muhammadiyah Semarang, hlm 29-35.

[5] Fauzi, Fathurokhman dkk. 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, hlm 88-97.

[6] Hanum, Nugraha Listiana dan Achmad Udin Zaelani. 2020. Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra Sejahtera. Journal Of Technology Information, vol. 6 No. 1, hlm 7-14.

[7] Kaunang, Fergie Juanda. 2018. Penerapan Algoritma J48 Decision Tree Untuk Analisis Tingkat Kemiskinan di Indonesia. Cogito Smart Journal. Vol. 4 No. 2, hlm 348-357.

[8] Karlik, Bekir dkk. 2016. Comprising Feature Selection and Classifier Methods with SMOTE for Prediction of Male Infertility. International Journal Of Fuzzy Systems And Advanced Applications, vol. 3, hlm 1-6.

[9] Karyadiputra, Erfan. 2016. Analisis Algoritma Naive Bayes untuk Klasifikasi Status Kesejahteraan Rumah Tangga Keluarga Binaan Sosial. Technologia, vol. 7, No. 4, hlm 199-208.

[10] Longadge, Mr. Rushi dkk. 2013. Class Imbalance Problem in Data Mining: Review. International Journal of Computer Science and Network (IJCSN) , Volume 2, Issue 1.

[11] Putra, Ridwan Miftahul dkk. 2018. Prediksi Indeks Pembangunan Manusia Menggunakan Algoritma C4.5 Di Kabupaten Kampar. Jurnal Teknologi Informasi & Komunikasi Digital Zone, Volume 9, Nomor 2, hlm 204-214.

[12] Sari, Betha Nurina dan Priati. 2016. Identifikasi Keterkaitan Variabel dan Prediksi Indeks Pembangunan Manusia (IPM) Provinsi Jawa Barat Menggunakan Dynamic Bayesian Networks. Jurnal Infotel, Vol.8 No.2 , hlm 150-155.
[13] Bahera, Santi Kumari dkk. 2020. Maturity Status Classification of Papaya Fruits based on Machine Learningand Transfer Learning Approach. Science Direct, Elsevier, hlm 1-7.
[14] You, Jihao dkk. 2020. Application Random Forest Classification to Predict Daily Oviposition events in Broiler Breeders Fed by Precision Feeding System. Science Direct, Elsevier, hlm 1-8.
[15] E.Ramos, Minerva dkk. 2020. Gender Inequality and Gender-based Poverty. Science Direct, Elsevier, vol 6, Issue 1, hlm 1-10.
[16] Erlando, Angga dkk. 2020. Financial Inclusion, Economic Growth, and Poverty Alleviation : Evidence from Eastern Indonesia. Science Direct, Elsevier, vol 6, Issue 10, hlm 1-13.
[17] Surjono, dkk. 2015. Gender Inequality and Social Capital as Rural Development Indicators in Indonesia (Case : Malang Regency, Indonesia). Science Direct, Elsevier, vol 211, hlm 370-374.
[18] Kasinathan, Thenmozhi dkk. 2020. Insect Classification and Detection in Field Crops using Modern Machine Learning Techniques. Science Direct, Elsevier, hlm 1-12.
[19] Nnamoko, Nonso & Ioannis Korkontzelos. 2020. Efficient Treatment of Outliers and Class Imbalance for Diabetes Prediction. ScienceDirect, Elsevier, hlm 1-12.
[20] Informatikalogi. *Algoritma Naive Bayes*. [online]. Tersedia : https://informatikalogi.com/algoritma-naive-bayes/ [Diakses pada 3 Desember 2020].