e-ISSN :
p-ISSN :

JCI, Vol. 1 No. 1 (2021) 49 – 57 | https://doi.org/10.54066/jci.v1i1.212

Jurnal Cakrawala Informasi

Journal Homepage: http://www.itbsemarang.ac.id/sijies/index.php/jci
e-Mail: jci@itbsemarang.ac.id

Analisa dan Komparasi 5 Algoritma Klasifikasi untuk Penduduk Miskin berdasarkan Usia dan Jenis Kelamin

Munawir Hamzah

Akademi Komunitas BPPMNU Banat Kudus

INFO ARTIKEL

Histori artikel:

Diterima : 8 Juni 2021
Revisi : 17 Juni 2021
Disetujui : 29 Juni 2021
Publikasi : 30 Juni 2021

Kata kunci:
Komparasi
Algoritma
Decision Tree
Naive Bayes
K-Nearest Neighbor
Random Tree
Random Forest
T-Test

ABSTRACT

Poverty is diverse and reflects a spectrum of ideological orientations. The factor that influences the poverty rate is economic growth. So poverty is no longer just a matter of lack of food. Growth in all business sectors is urgently needed in an effort to reduce poverty levels. To handle and coordinate matters relating to poverty reduction, it is necessary to classify the age and sex of individuals with the lowest 30% welfare level. Along with the rapid development of technology, the development of computer algorithms is developing several algorithms to support the advancement of computerized systems. Well-known algorithms include the Decision Tree (C4.5) algorithm, Random Tree, Linear and Quadratic Discriminant Analysis, Neural Network, Least Square Support Vector Machines, K-NN, Random Forest, CART, and Naive Bayes.

ABSTRAK

Kemiskinan adalah berbeda-beda dan merefleksikan suatu spektrum orientasi ideologi. Faktor yang mempengaruhi tingkat kemiskinan adalah pertumbuhan ekonomi. Jadi kemiskinan tidak lagi sekedar masalah kekurangan makanan saja. Pertumbuhan keseluruh sektor usaha sangat dibutuhkan dalam upaya menurunkan tingkat kemiskinan. Untuk menangani dan berkoordinasi dalam hal-hal yang berkaitan dengan penanggulangan kemiskinan, maka perlu mengklasifikasikan usia dan jenis kelamin dari individu dengan tingkat kesejahteraan 30% terbawah. Seiring dengan perkembangan teknologi yang begitu pesat, perkembangan algoritma komputer sedang mengembangkan beberapa algoritma untuk mendukung kemajuan sistem komputerisasi. Algoritma yang terkenal diantaranya adalah algoritma *Decision Tree* (C4.5), *Random Tree*, *Linear and Quadratic Discriminant Analysis*, *Neural Network*, *Least*

* Korespondensi penulis: m.hamzah001@gmail.com

Square Support Vector Machines, K-NN, *Random Forest*, CART, dan *Naive Bayes*.

PENDAHULUAN

Kadar kemiskinan tidak lagi sekedar masalah kekurangan makanan, tetapi bagi warga masyarakat tertentu bahkan sudah mencapai tahap ekstrem sampai level kehabisan dan ketiadaan makanan. Tidak sedikit orang terkapar karena tidak tahan menderita kelaparan dan kekurangan gizi yang membuka jalan lebih cepat kearah kematian dini. Inilah proses kematian secara pelan-pelan tetapi kejam. Tidak sedikit orang gagal mengelola rasa lapar dan kemiskinan. Masih ada sebagian warga masyarakat untuk dapat makan sekali sehari saja sulit. Banyak definisi tentang kemiskinan telah diungkapkan dan menjadi bahan perdebatan. Kemiskinan telah didefinisikan berbeda-beda dan merefleksikan suatu spektrum orientasi ideologi. Faktor yang mempengaruhi tingkat kemiskinan adalah pertumbuhan ekonomi. Pertumbuhan ekonomi yang tinggi dan disertai pemerataan hasil pertumbuhan keseluruh sektor usaha sangat dibutuhkan dalam upaya menurunkan tingkat kemiskinan. Selain itu pengangguran juga berpengaruh terhadap tingkat kemiskinan.

Berdasarkan dasar permasalahan di atas maka persoalan penelitian yang ingin dipecahkan dalam penelitian ini adalah bagaimana mengklasifikasikan usia dan jenis kelamin dari individu dengan tingkat kesejahteraan 30% terbawah. Pada era globalisasi yang terjadi di seluruh dunia kebutuhan akan informasi sangat diperlukan. Perkembangan ilmu pengetahuan dan teknologi sudah semakin cepat, terlebih lagi informasi yang dihasilkan mengandung nilai yang benar, akurat, cepat, dan tepat, sehingga siapapun dan apapun yang menggunakan informasi tersebut dapat menangani berbagai masalah yang terjadi dengan cepat. Seiring dengan perkembangan teknologi yang begitu pesat, perkembangan algoritma komputer sedang mengembangkan beberapa algoritma untuk mendukung kemajuan sistem komputerisasi. Beberapa algoritma yang terkenal diantaranya adalah algoritma *Decision Tree* (C4.5), *Random Tree*, *Linear and Quadratic Discriminant Analysis*, *Neural Network*, *Least Square Support Vector Machines*, K-NN, *Random Forest*, CART, dan *Naive Bayes*. Dalam penelitian ini dilakukan pengujian terhadap lima algoritma klasifikasi yaitu *Decision Tree* (C4.5), *Naive Bayes*, K-*Nearest Neighbor*, *Random Tree*, dan *Random Forest*. Untuk metode validasi yang digunakan adalah 10-*fold cross validation* untuk *training* dan *testing dataset* serta menambahkan *t-test*.

TINJAUAN PUSTAKA

Analisa perbandingan eksperimen algoritma klasifikasi untuk keseimbangan telah dilakukan oleh seorang peneliti Brown dan Mues (2012) yang melakukan perbandingan beberapa teknik yang dapat digunakan dalam analisis keseimbang *dataset credit scoring* [1]. Dalam konteks *credit scoring*, *dataset* seimbang sering terjadi karena jumlah pinjaman *default* dalam portofolio biasanya jauh lebih rendah dari jumlah pengamatan yang tidak baku. Teknik-teknik yang akan diterapkan dalam penelitiannya adalah regresi logistik (LOG), *Linear dan Quadratic Discriminant Analysis* (LDA, QDA), *Least Square Support Vector Machines* (LS-SVM), *Decision Trees* (C4.5), *Neural Networks* (NN), *Nearest-Neighbour Classifiers* (k-NN10, k-NN100), meningkatkan gradien algoritma dan *Random Forests*. Jurnal tersebut sangat tertarik pada daya

dan kegunaan dari meningkatkan gradien dan *Random Forest Classifiers* yang belum ditelusuri secara menyeluruh dalam konteks penilaian kredit. Hasil dari jurnal tersebut menunjukkan bahwa *Random Forest* dan meningkatkan gradien acak tampil sangat baik dalam konteks penilaian kredit dan mampu mengatasi relatif baik dengan ketidak seimbangan kelas pada *dataset*. Jurnal tersebut juga menemukan bahwa, ketika dihadapkan dengan ketidakseimbangan kelas yang besar, algoritma C4.5 *Decision Tree*, analisis diskriminan kuadrat dan K-*Nearest Neighbours* melakukan secara signifikan lebih buruk daripada berkinerja pengklasifikasian terbaik.

Pada penelitian lain membahas teknik klasifikasi utama, yang meliputi model statistik tradisional (LDA, QDA, dan *Logistic Regression*), K-*Nearest Neighbors*, *Bayesian Networks* (*Naive Bayes* dan TAN), *Decision Trees* (C4.5), *Associative Classification* (CBA), *Neural Network*, dan *Support Vector Machines* (SVM), dan mereka berlaku untuk mengendalikan risiko kredit, Yu, *et al* (2007) [2]. Percobaan dilakukan pada 244 perusahaan dinilai terutama dari Industri dan Komersial Bank of China. Hasil menunjukkan bahwa sementara model statistik tradisional menghasilkan hasil yang paling rendah, C4.5 atau SVM tidak menunjukkan hasil yang memuaskan dan CBA tampaknya menjadi pilihan terbaik untuk peringkat kredit dalam hal prediktabilitas dan *interpretability*.

Romi Satria Wahono, et al (2014) dalam penelitiannya membandingkan model klasifikasi dalam bidang prediksi pada cacat *software* [3]. Berbagai jenis algoritma klasifikasi telah diterapkan untuk prediksi cacat *software*. Namun, tidak ada konsensus yang jelas dimana algoritma melakukan yang terbaik ketika studi individu yang melihat secara terpisah. Tujuan dari penelitian ini, menggunakan 10 pengklasifikasian yang dipilih dan diterapkan untuk membangun model klasifikasi dan menguji kinerja mereka di 9 NASA MDP *dataset*. Area di bawah kurva (AUC) digunakan sebagai indikator akurasi dalam *framework* untuk mengevaluasi kinerja pengklasifikasian. *Friedman* dan *Nemenyi Post Hoc Tests* digunakan untuk menguji signifikansi perbedaan AUC antara pengklasifikasian. Hasil penelitian menunjukkan bahwa *Logistic Regression* melakukan yang lebih baik pada *dataset* NASA MDP. *Naive Bayes*, *Neural Network*, *Support Vector Machine and Classifiers* juga terlihat baik. Berdasarkan klasifikasi *Decision Tree* cenderung di bawah, serta analisis linear diskriminan dan K-*Nearest Neighbor*.

Data Mining

Data mining adalah tentang pemecahan masalah dengan menganalisis data yang sudah ada dalam *database*. Misalkan, untuk mengambil contoh yang sudah usang, masalahnya adalah loyalitas pelanggan berubah-ubah dalam pasar yang sangat kompetitif. *Database* pilihan pelanggan, bersama dengan profil pelanggan, memegang kunci untuk masalah ini [4]. *Data mining* adalah studi tentang pengumpulan, pembersihan, pengolahan, analisis, dan mendapatkan wawasan yang berguna dari data. Sebuah variasi ada dalam hal masalah domain, aplikasi, representasi formulasi, dan data yang ditemui dalam aplikasi nyata. Oleh karena itu, "*data mining*" adalah istilah payung yang luas yang digunakan untuk menggambarkan aspek-aspek yang berbeda dari pengolahan data [5].

Klasifikasi

Klasifikasi adalah salah satu teknik *data mining* yang paling populer yang dapat digunakan untuk pengambilan keputusan yang cerdas [6]. Jadi proses untuk menyatakan suatu objek ke salah satu kategori yang sudah didefinisikan sebelumnya. Tujuannya adalah *record-record* yang sebelumnya tidak terlihat dinyatakan kelasnya seakurat mungkin. Model klasifikasi digunakan untuk:

1. Pemodelan deskriptif sebagai perangkat penggambaran untuk membedakan objek-objek dari kelas berbeda.
2. Pemodelan prediktif digunakan untuk memprediksi label kelas untuk *record* yang tidak diketahui atau tidak dikenal.

Algoritma *Decision Tree* (C4.5)

Decision Tree juga dikenal sebagai pohon klasifikasi, adalah struktur sederhana yang dapat digunakan sebagai penggolongan. Sebuah aspek penting dari pohon keputusan adalah ketidakstabilan yang melekat yang berarti bahwa mereka sensitif terhadap perubahan dalam contoh pelatihan dan karena itu hipotesis yang berbeda secara signifikan yang dihasilkan untuk *dataset* pelatihan yang berbeda. Pohon keputusan dibangun dari berbagai pelatihan subsampel dari domain masalah yang diberikan akan menghasilkan model yang sangat berbeda [7].

Algoritma *Naive Bayes* (NB)

Naive Bayes (NB) adalah penggolong probabilistik sederhana berdasarkan: (a) teori *Bayes*, (b) yang kuat (*naive*) asumsi kemandirian, dan (c) model fitur independen. *Naive Bayes* klasifikasi juga banyak digunakan untuk masalah klasifikasi dalam *data mining* dan *machine learning* bidang karena kesederhanaan dan akurasi klasifikasi mengesankan [6]. Metode *Naive Bayes*, yang merupakan bentuk sederhana dari *Bayesian Network* adalah metode *data mining* populer yang telah diterapkan untuk banyak domain, termasuk deteksi intrusi. Kesederhanaan metode bergantung pada asumsi bahwa semua fitur yang independen satu sama lain [8].

Algoritma K-*Nearest Neighbor* (K-NN)

Tujuan dari k terdekat *neighbour*s (K-NN) algoritma adalah dengan menggunakan *database* di mana titik data dipisahkan menjadi beberapa kelas terpisah untuk memprediksi klasifikasi titik sampel baru [9]. Cara dimana algoritma memutuskan mana dari *poin training set* cukup mirip yang harus dipertimbangkan ketika memilih kelas untuk memprediksi pengamatan baru yang memilih data poin yang paling dekat dengan pengamatan baru dan untuk mengambil yang paling umum diantara kelas. Inilah sebabnya mengapa hal itu disebut algoritma K-*Nearest Neighbour*s.

Algoritma *Random Forest* (RF)

Saat ini, algoritma pembelajaran mesin yang disebut *Random Forest* (RF) secara luas dianggap sebagai salah satu pengklasifikasi paling akurat yang menarik perhatian banyak peneliti di daerah [10]. Algoritma ini pada dasarnya didasarkan pada *Decision Tree* bekerja di*ensemble*. Hal ini dikembangkan dari dua pendekatan yang sukses disarankan sebelumnya.

Algoritma *Random Tree* (RT)

Random tree adalah algoritma klasifikasi yang memecahkan masalah regresi dalam *Decision Tree*. Ini adalah sekumpulan *tree predictor*s disebut *forest*. Setiap pohon data di dalam *forest*. Dalam regresi, respon *classifier* adalah rata-rata

semua pohon di dalam *forest*. Algoritma pohon acak juga diterapkan pada data untuk mendapatkan informasi yang sama seperti Decision Tree [11].

Metode Validasi k-*Fold Cross Validation*

Dalam k-*fold cross* validasi, kadang-kadang disebut estimasi rotasi, *dataset* D secara acak dibagi menjadi k subset saling eksklusif (lipatan) D1, D2, ..., Dk [12]. *Cross validation* adalah teknik komputer yang intensif, menggunakan semua contoh tersedia sebagai pelatihan dan uji contoh [13].

T-Test

Pada dasarnya *t-test* tidak lain adalah *z-score*, jika *z-score* menunjukkan distribusi angka kasar maka *t-score* atau *t-test* adalah disribusi perbedaan *mean*. Fungsi *t-test* adalah sebagai uji komparasi antar 2 atau lebih sampel bebas (*independent*). Tes ini diterapkan jika analis data bertujuan untuk mengetahui apakah kelompok sampel berbeda dalam variabel tertentu. *T-test* diaplikasikan dengan beberapa kondisi antara lain:

1. Berhadapan dengan 2 atau lebih sampel bebas.
2. Tiap sampel diambil secara random.
3. Variabel yang dikomparasikan menghasilkan data paling rendah berskala interval.

METODE PENELITIAN

Dataset

Dalam penelitian ini *dataset* yang digunakan merupakan *dataset* yang diperoleh dari Tim Nasional Percepatan Penanggulangan Kemiskinan (TNP2K) adalah lembaga yang dibentuk untuk menangani dan berkoordinasi dalam hal-hal yang berkaitan dengan penanggulangan dan pengentasan kemiskinan di Republik Indonesia. Struktur datanya sebagai berikut:

Tabel 1. Struktur *Dataset*

No.	Attribut	Tipe Data	Keterangan
1	Kode Provinsi	Numeric	Attribut
2	Nama Provinsi	Polynominal	Attribut
3	Kelompok Usia	Polynominal	Attribut
4	Jenis Kelamin	Binominal	Label
5	Jumlah Individu	Integer	Attribut

Penelitian ini mengambil sampel sebanyak 330 *record* data dengan jumlah atribut empat dan satu label.

Algoritma Klasifikasi

Penelitian ini membandingkan hasil akurasi lima algoritma klasifikasi untuk mengklasifikasi usia dan jenis kelamin dari individu dengan tingkat kesejahteraan. Pada penelitian ini digunakan lima algoritma klasifikasi, yaitu *Decision Tree* (C4.5), *Naive Bayes*, K-*Nearest Neighbor, Logistic Regression*, dan *Neural Network*. Kelima algoritma ini dipilih karena algoritma ini populer digunakan dalam analisis komparasi.

Metode Perbandingan

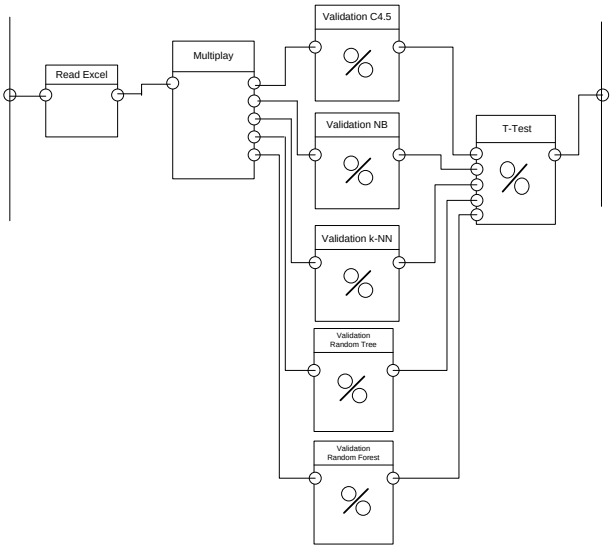
Ada tiga golongan dari uji statistik yang dapat digunakan untuk membandingkan dua atau lebih pengklasifikasi selama beberapa *dataset* tes parametrik (pasangan *t-test* dan ANOVA), tes non-parametrik (*Wilcoxon* dan uji *Friedman*) dan non-parametrik tes yang tidak bertanggung *commensurability* hasil (uji tanda). Demsar merekomendasikan tes *Friedman* untuk perbandingan *classifier*, yang bergantung pada asumsi kurang membatasi. Berdasarkan rekomendasi ini, dalam rangka kita uji *Friedman*

digunakan untuk membandingkan AUC dari pengklasifikasi yang berbeda. Tes *Friedman* didasarkan pada rata-rata peringkat (R) kinerja algoritma klasifikasi pada setiap *dataset* [14].

PEMBAHASAN DAN HASIL

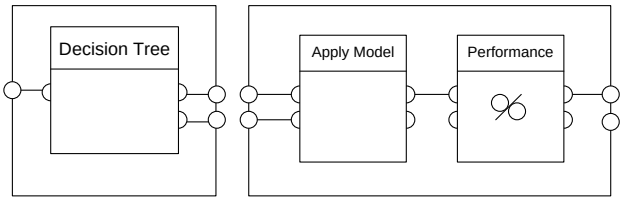
Training dan *Testing Dataset* Menggunakan Algoritma Klasifikasi

Penelitian ini digunakan *tool RapidMiner Studio* 6.3.000 untuk mengukur performansi dengan *confusion matrix* dan ROC *curve*. Berikut skema pengukuran performansi lima algoritma yaitu (*Decision Tree* (C4.5), *Naive Bayes*, K-*Nearest Neighbor* (K-NN), *Random Forest*, dan *Random Tree*).

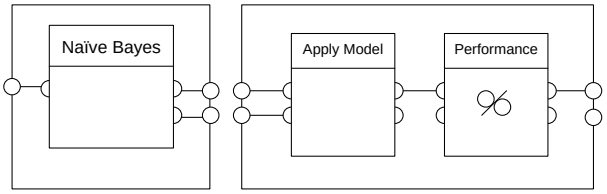


Gambar 1. Skema Performansi Algoritma Klasifikasi

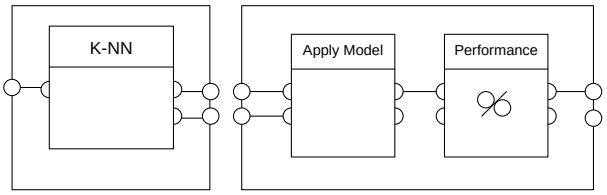
Pada proses masing-masing validasi algoritma klasifikasi terdapat subproses. Berikut skema masing-masing proses validasi pada algoritma klasifikasi:



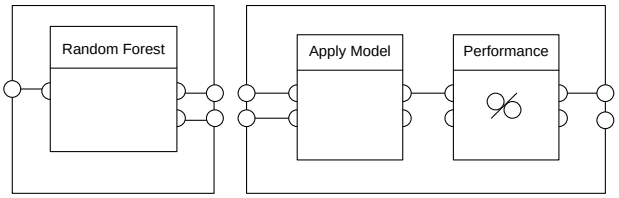
Gambar 2. Skema Subproses Validasi Algoritma *Decision Tree*



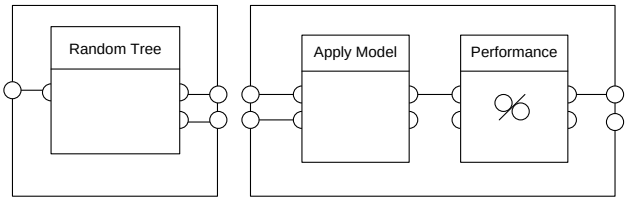
Gambar 3. Skema Subproses Validasi Algoritma *Naive Bayes*



Gambar 4. Skema Subproses Validasi Algoritma K-NN



Gambar 5. Skema Subproses Validasi Algoritma *Random Forest*



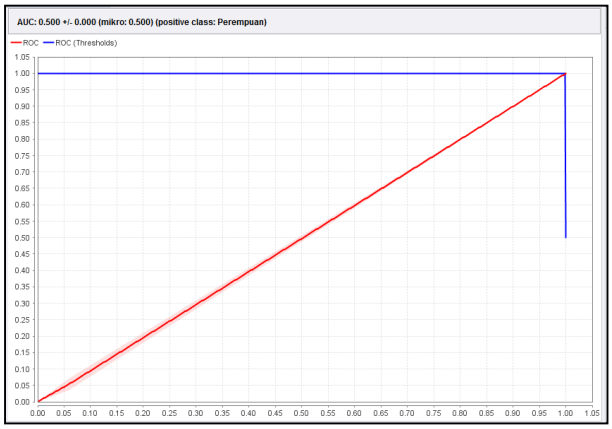
Gambar 6. Skema Subproses Validasi Algoritma *Random Tree*

Nilai Akurasi Algoritma Klasifikasi dengan *Confusion Matrix* dan ROC *Curve*

Dari proses validasi yang menggunakan metode k-*fold cross validation*, diperoleh nilai akurasi masing-masing algoritma klasifikasi:

Tabel 2. *Confusion Matrix* Algoritma C4.5

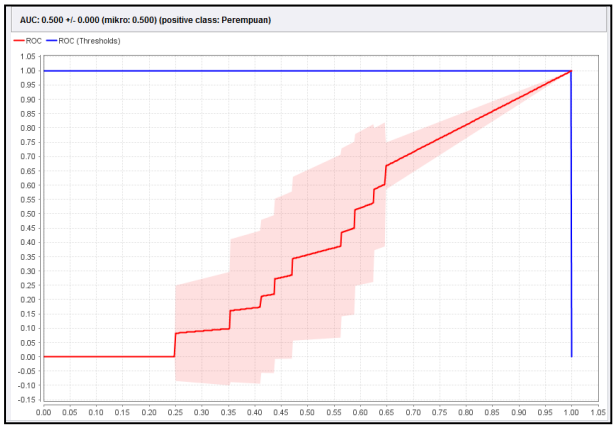
Accuracy: 48.48% +/- 0.00% (mikro: 48.48%)			
	true Laki-Laki	true Perempuan	class precision
pred. Laki-Laki	96	101	48.73%
pred. Perempuan	69	64	48.12%
class recall	58.18%	38.79%	



Gambar 7. ROC *Curve* Algoritma C4.5

Tabel 3. *Confusion Matrix* Algoritma K-NN

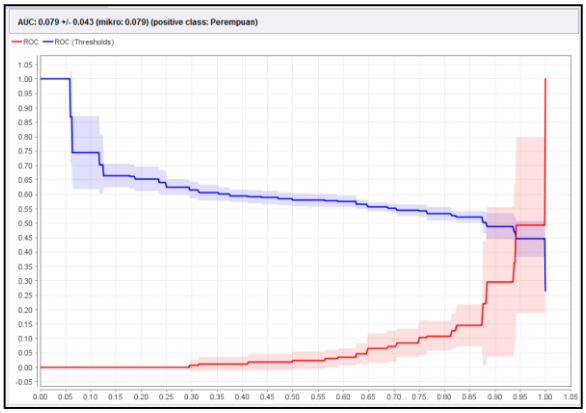
Accuracy: 52.42% +/- 6.22% (mikro: 52.42%)			
	true Laki-Laki	true Perempuan	class precision
pred. Laki-Laki	89	81	52.35%
pred. Perempuan	76	84	52.50%
class recall	53.94%	50.91%	



Gambar 8. ROC *Curve* Algoritma K-NN

Tabel 4. *Confusion Matrix* Algoritma *Naive Bayes*

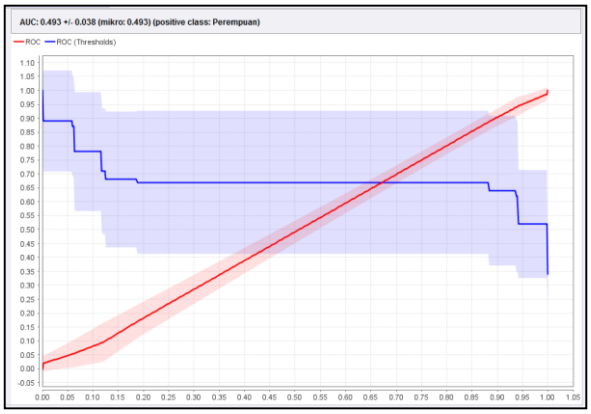
Accuracy: 14.55% +/- 5.72% (mikro: 14.55%)			
	true Laki-Laki	true Perempuan	class precision
pred. Laki-Laki	18	135	11.76%
pred. Perempuan	147	30	16.95%
class recall	10.91%	18.18%	



Gambar 9. ROC *Curve* Algoritma *Naive Bayes*

Tabel 5. *Confusion Matrix* Algoritma *Random Forest*

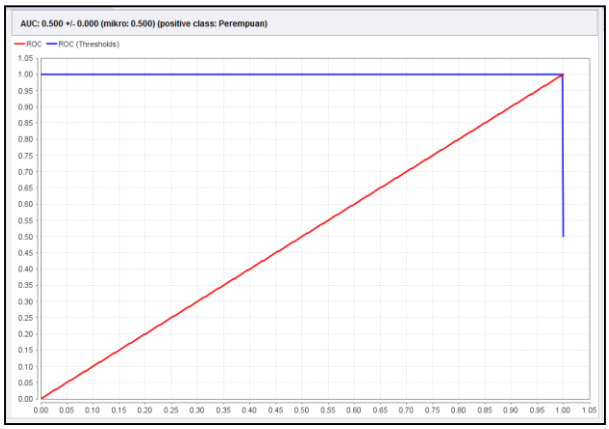
Accuracy: 50.00% +/- 7.70% (mikro: 50.00%)			
	true Laki-Laki	true Perempuan	class precision
pred. Laki-Laki	124	124	50.00%
pred. Perempuan	41	41	50.00%
class recall	75.15%	24.85%	



Gambar 10. ROC *Curve* Algoritma *Random Forest*

Tabel 6. *Confusion Matrix* Algoritma *Random Tree*

Accuracy: 48.48% +/- 0.00% (mikro: 48.48%)			
	true Laki-Laki	true Perempuan	class precision
pred. Laki-Laki	180	85	48.48%
pred. Perempuan	85	80	48.48%
class recall	48.48%	48.48%	



Gambar 11. ROC *Curve* Algoritma *Random Tree*

Dari proses pengukuran akurasi masing-masing algoritma dengan *confusion matrix* dan ROC *curve*, diperoleh nilai *accuracy* dan AUC:

Tabel 7. Perbandingan *Accuracy* dan AUC

	C4.5	**NB**	**K-NN**	**RF**	**RT**
Accuracy	0.4848	0.1455	0.5242	0.5000	0.4848
AUC	0.5000	0.0790	0.5000	0.4930	0.5000

Akurasi dari masing-masing algoritma klasifikasi diuji dengan uji beda parametrik *t-test*, diperoleh tabel uji *t-test*:

Tabel 8. *T-Test Significance*

	0.482 +/- 0.009 C4.5	0.142 +/- 0.058 NB	0.527 +/- 0.068 K-NN	0.503 +/- 0.015 RF	0.485 +/- 0.000 RT
0.482 +/- 0.009 C4.5		**0.000**	0.058	**0.001**	0.443
0.142 +/- 0.058 NB			**0.000**	**0.000**	**0.000**
0.527 +/- 0.068 K-NN				0.406	0.075
0.503 +/- 0.015 RF					**0.001**
0.485 +/- 0.000 RT					

Nilai yang dicetak tebal berarti lebih kecil dari alpha = 0.050 yang mengindikasikan adanya perbedaan signifikan diantara nilai rata-rata aktual. Dari tabel 13 dapat ditarik kesimpulan, algoritma *Decision Tree* (C4.5) memiliki akurasi paling bagus (dominan) terhadap algoritma yang lain. Berikutnya ada algoritma *Naive Bayes*, *Random Forest*, dan *Random Tree*, tidak ada perbedaan signifikan diantara algoritma tersebut. Namun demikian, algoritma *Logistic Regression* dan *Neural Network* tidak lebih baik dari algoritma K-*Nearest Neighbor* (K-NN).

KESIMPULAN

Penelitian dengan membandingkan lima algoritma klasifikasi *Decision Tree* (C4.5), *Naive Bayes*, K-*Nearest Neighbor*, *Random Tree*, dan *Random Forest* untuk memprediksi usia dan jenis kelamin dari individu dengan tingkat kesejahteraan, menggunakan metode validasi k-*fold cross validation*, serta dilakukan uji beda terhadap akurasi masing-masing algoritma dengan uji beda parametrik *t-test* menunjukkan bahwa algoritma *Decision Tree* (C4.5) memiliki akurasi paling bagus (dominan) terhadap algoritma yang lain. Berikutnya ada algoritma *Naive Bayes*, *Random Tree*, dan *Random Forest*, tidak ada perbedaan signifikan diantara algoritma tersebut. Namun demikian, algoritma *Random Tree* dan *Random Forest* tidak lebih baik dari algoritma K-*Nearest Neighbor* (K-NN). Setelah diperoleh algoritma yang paling bagus tingkat akurasinya yakni algoritma *Decision Tree* (C4.5), maka algoritma tersebut dapat digunakan untuk memprediksi usia dan jenis kelamin dari individu dengan tingkat kesejahteraan 30% terbawah sehingga dapat penanggulangan dan pengentasan kemiskinan di Republik Indonesia.

DAFTAR PUSTAKA

[1] I. Brown and C. Mues, “An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets,” *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3446–3453, 2012, doi: 10.1016/j.eswa.2011.09.033.

[2] L. Yu, G. Chen, A. Koronios, S. Zhu, and X. Guo, “Application and Comparison of Classification Techniques in Controlling Credit Risk,” pp. 111–145.

[3] R. S. Wahono, N. S. Herman, and S. Ahmad, “A Comparison Framework of Classification Models for Software Defect Prediction,” vol. 20, no. 10, pp. 1945–1950, 2014, doi: 10.1166/asl.2014.5640.

[4] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining “Practical Machine Learning Tools and Techniques,”* Third Edit. Burlington: Morgan Kaufmann Publishers, 2011.

[5] C. C. Aggarwal, *Data Mining: The Textbook 2015th Edition*. New York: Springer, 2015.

[6] D. Farid, L. Zhang, C. Mofizur, M. A. Hossain, and R. Strachan, “Expert Systems with Applications Hybrid Decision Tree and Naive Bayes Classifiers for Multiclass Classification Tasks,” *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1937–1946, 2014, doi: 10.1016/j.eswa.2013.08.089.

[7] C. J. Mantas and J. Abellán, “Expert Systems with Applications Credal-C4 . 5 : Decision Tree based on Imprecise Probabilities to Classify Noisy Data,” *Expert Syst. Appl.*, vol. 41, no. 10, pp. 4625–4637, 2014, doi: 10.1016/j.eswa.2014.01.017.

[8] L. Koc, T. A. Mazzuchi, and S. Sarkani, “Expert Systems with Applications a Network Intrusion Detection System based on a Hidden Naive Bayes Multiclass Classifier,” *Expert Syst. Appl.*, vol. 39, no. 18, pp. 13492–13500, 2012, doi: 10.1016/j.eswa.2012.07.009.

[9] C. Sutton, “Nearest Neighbor Methods,” *Semant. Sch.*, vol. 4, no. 3, pp. 307–309, 2012.

[10] V. Sazonau, “Implementation and Evaluation of a Random Forest Machine Learning Algorithm,” *Comput. Sci.*, 2012.

[11] A. Jameela and P. Revathy, “Comparison of Decision and Random Tree Algorithms on a Web Log Data for Finding Frequent Patterns,” *Int. J. Res. Eng. Technol.*, vol. 3, no. 7, 2014.

[12] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” *IJCAI Proc.*, vol. 95, no. 2, pp. 1137–1143, 1995.

[13] Y. Bengio and Y. Grandvalet, “No Unbiased Estimator of the Variance of K-Fold Cross Validation,” *J. Mach. Learn. Res.*, vol. 5, pp. 1089–1105, 2004.

[14] R. S. Wahono, “Introduction Data Mining,” in *Data Mining*, Jakarta, 2011, pp. 1–53.