P-ISSN: 2548-9704
E-ISSN: 2686-0880

PENGUJIAN TEKNIK ALGORITMA KLASIFIKASI TERHADAP TINGKAT KEMISKINAN PENDUDUK

Anna

Program Studi Sistem Informasi Kampus Kota Pontianak
Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika
Jl. Abdurrahman Saleh No.18A, Pontianak
E-mail: anna.nnz@bsi.ac.id

ABSTRACT

One of the problems is endless in Indonesia is poverty. The current government, while incessantly making efforts to reduce poverty through a variety of policies and speedy breakthrough. So that the required data mining to process data on poverty in order to provide an accurate infromasi about the level of poverty in every province in Indonesia. This research discusses the algorithm is fast and fit in the classification at the level of the poverty population in Indonesia in particular. Some algorithms are tested, namely K-Nearest Neighbors, Decision Tree (C4.5), Naive Bayes, Random Forest, and Decision Stump. The testing will be done in 35 provinces in Indonesia. The results obtained show that the algorithm C4.5 method is an appropriate method in the classification of the population poverty rate in Indonesia.

Keywords: Data Mining, Poverty, Classification Algorithms, Testing

ABSTRAK

Salah satu masalah yang tidak ada habisnya di Indonesia adalah kemiskinan. Pemerintah saat ini sedang gencar melakukan upaya penanggulangan kemiskinan melalui berbagai kebijakan dan terobosan yang cepat. Sehingga diperlukan data mining untuk mengolah data kemiskinan agar dapat memberikan infromasi yang akurat tentang tingkat kemiskinan di setiap provinsi di Indonesia. Penelitian ini membahas tentang algoritma fast and fit dalam klasifikasi pada tingkat penduduk miskin di Indonesia khususnya. Beberapa algoritma yang diuji yaitu K-Nearest Neighbors, Decision Tree (C4.5), Naive Bayes, Random Forest, dan Decision Stump. Pengujian akan dilakukan di 35 provinsi di Indonesia. Hasil yang diperoleh menunjukkan bahwa metode algoritma C4.5 merupakan metode yang tepat dalam klasifikasi angka kemiskinan penduduk di Indonesia.

Kata Kunci: Data Mining, Kemiskinan, Algoritma Klasifikasi, Pengujian

I. PENDAHULUAN

Kemiskinan merupakan salah satu permasalahan yang masih menjadi pokok utama pemerintah. Dalam menentukan suatu penduduk tergolong miskin atau tidak saat ini pemerintah masih menggunakan sampel rumah tangga miskin, dimana secara definisi rumah tangga miskin adalah suatu kondisi individu yang akan dinilai berdasarkan karaktersitik kemiskinan yang telah ditetapkan[1]. Pentingnya mengolah data kemiskinan untuk informasi daerah kemiskinan. Kemiskinan merupakan permasalahan bangsa yang memerlukan langkah-langkah penanganan dan pendekatan yang sistematik, terpadu, dan

P-ISSN: 2548-9704
E-ISSN: 2686-0880

menyeluruh[2]. Dalam rangka mengurangi beban dan memenuhi hak-hak dasar warga negara secara layak untuk mewujudkan kehidupan masyarakat Indonesia yang bermartabat. Menurut BPS dikatakan bahwa untuk menuju solusi kemiskinan penting bagi kita untuk menulusuri secara detail indikator-indikator kemiskinan tersebut. Adapun indikator-indikator kemiskinan yaitu:

1. Ketidakmampuan memenuhi kebutuhan konsumsi dasar (sandang, pangan dan papan).
2. Tidak adanya akses terhadap kebutuhan hidup dasar lainnya (kesehatan,pendidikan,sanitasi,air bersih dan transportasi).
3. Tidak adanya jaminan masa depan (karena tiadanya inventasi untuk pendidikan keluarga).
4. Kerentanan terhadap goncangan yang bersifat individual maupun massa.
5. Rendahnya kualitas sumber daya manusia dan terbtasnya sumber daya alam.
6. Kurangnya apresiasi dalam kegiatan sosial masyarakat.

Pemerintah saat ini sedang gencar-gencarnya melakukan upaya penanggulangan kemiskinan dengan melakukan berbagai kebijakan dan gebrakan cepat. Program-program bantuan dari pemerintah agar tersalur dengan baik jika sudah diketahui secara akurat tingkat kemiskinan suatu provinsi. Permasalahan yang muncul adalah belum diketahuinya dareah yang tingkat kemiskinan penduduknya rendah, sedang dan tinggi dilihat berdasarkan perbandingan setiap tahunnya [1]. Untuk itu penelitian ini dilakukan bertujuan untuk menentukan algoritma klasifikasi yang tepat dan akurasi yang cepat dalam menentukan pengelompokkan kategori tingkat kemiskinan penduduk di setiap provinsi di seluruh Indonesia. Upaya pengklasifikasian kemiskinan ini dilakukan karena adanya fakta bahwa terjadinya pertumbuhan ekonomi yang tidak merata di wilayah Indonesia [2]. Sehingga penelitian ini dapat bermanfaat bagi pemerintah dalam menyalurkan bantuan-bantuan atapun sebuah informasi untuk menentukan program yang tepat diberikan kepada daerah yang tergolong tinggi tingkat kemiskinan penduduk.

Data mining merupakan metode yang digunakan untuk menggali ilmu pengetahuan dari kumpulan data yang berjumlah benar atau kompleks dengan tujuan untuk menemukan pola atau kecenderungan yang penting yang biasanya tidak disadari keberadaannya [3]. Pada *data mining* dikenal metode klasifikasi, secara sederhana algoritma klasifikasi merupakan sebuah catatan *record* data hendak diklasifikasikan ke dalam salah satu dari sekian klasifikasi data yang tersedia pada variabel tujuan berdasarkan nilai-nilai "variabel prediktor"[4]. Penerapan algoritma klasifikasi dapat menghasilkan tingkat kemiskinan penduduk yang dapat mengidentifikasikan sasaran penerima manfaat Program Penanggulangan Kemiskinan.

2. METODOLOGI

Model klasifikasi yang baik dapat dilihat dari tingkat akurasinya dalam memprediksi berdasarkan kategori respon. Hal yang dapat memengaruhi akurasi dari model klasifikasi salah satunya adalah masalah ketidakseimbangan data[5]. Klasifikasi dalam data mining bekerja pada data historis atau data sejarah. Data historis disebut data latihan atau training data. histori data digunakan sebagai cara mendapatkan pengetahuan dan disebut data pengalaman.

P-ISSN: 2548-9704
E-ISSN: 2686-0880

Dalam penelitian ini menggunakan *data training* tingkat kemiskinan penduduk di 35 provinsi yang ada di Indonesia dengan label kategori rendah, sedang, dan tinggi. Dengan data tersebut pula dapat digunakan sebagai data testing. Untuk mendapatkan ketiga kategori tersebut maka akan diuji dengan lima algoritma klasifikasi yaitu C4.5, k-Nearest Neighbour (k-NN), Random Forest, Naïve Bayes, dan Decision Stump dengan tingkat akurasi yang berbeda-beda.

Algoritma C4.5 merupakan salah satu algoritma decision tree yang paling efektif untuk melakukan klasifikasi. *Decision Trees* atau dengan nama lain pohon keputusan yang merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan[6]. Pohon keputusan ini dibangun dengan cara membagi data secara rekursif hingga tiap bagian terdiri dari data yang berasal dari kelas yang sama.

Bentuk pemecahan (*split)* yang digunakan untuk membagi data tergantung dari jenis atribut yang digunakan dalam split. Tahapan dalam membuat sebuah pohon keputusan dengan algoritma C4.5 yaitu:

a) Mempersiapkan *data training*, dapat diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan dalam kelas-kelas tertentu.
b) Menentukan akar dari pohon dengan menghitung nilai gain yang tertinggi dari masing-masing atribut atau berdasarkan nilai *index entropy* terendah. Sebelumnya dihitung terlebih dahulu nilai *index entropy*, dengan rumus:

$$Entropy\,(i) = -\sum_{j=1}^{m} f(i,j).\log 2\, f[(i,j)]$$

Keterangan:
I = himpunan kasus
m = jumlah partisi i

f(i,j) = proporsi j terhadap i

c) Hitung nilai gain dengan rumus :

$$Entropy\ split = -\sum_{i=1}^{p} \frac{n1}{n} .\ IE\,(i)$$

Keterangan:
p = jumlah partisi atribut
ni= proporsi ni terhadap i

n = jumlah kasus dalam n

d) Ulangi langkah ke-2 hingga semua *record* terpartisi.
e) Tidak ada *record* di dalam cabang yang kosong.

Nearest Neighbors adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama, yaitu berdasarkan pada pencocokkan bobot dari sejumlah fitur yang ada. Algoritma k-nearest neighbors (k-NN atau K-NN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut[3]. Tujuan dari algoritma ini adalah untuk mengklasifikasikan obyek baru berdasarkan atribut dan sampel-sampel dari *data training*. K-nearest neighbor bekerja berdasarkan jarak minimum dari data baru ke *data training samples* untuk menentukan K-tetangga terdekat. Klasifikasi k-NN menghitung Jarak Kedekatan antara data baru dengan *data training* menggunakan *Euclidean Distance*, yaitu dengan formula sebagai berikut:

$$d(p,q) = \sqrt{\sum_{i} (p_i - q_i)^2}$$

P-ISSN: 2548-9704
E-ISSN: 2686-0880

Dimana:

- p dan q = titik pada ruang vektor *n* dimensi
- pi dan qi adalah besaran skalar untuk dimensi ke-*i* dalam ruang vektor *n* dimensi

Tahapan algoritma k-NN sebagai berikut:

a) Tentukan parameter K = jumlah banyaknya tetangga terdekat
b) Hitung jarak antara data baru dan semua data yang ada di *data training*
c) Urutkan jarak tersebut dan tentukan tetangga mana yang terdekat berdasarkan jarak minimum ke-K.
d) Tentukan kategori dari tetangga terdekat.
e) Gunakan kategori mayoritas yang sederhana dari tetangga yang terdekat tersebut sebagai nilai prediksi dari data yang baru.

Random Forest (RF) merupakan salah satu algoritma klasifikasi dengan tingkat akurasi yang baik. Random Forest merupakan sebuah metode ensemble yang terdiri dari beberapa pohon keputusan sebagai classifier. Kelas yang dihasilkan dari proses klasifikasi ini diambil dari kelas terbanyak yang dihasilkan oleh pohon-pohon keputusan yang ada pada Random Forest. Dengan melakukan voting pada pohon-pohon keputusan yang tersedia membuat akurasi dari Random Forest meningkat. Random Forest tidak membutuhkan preprocessing khusus dalam pengimpelementasiannya untuk mencapai akurasi yang bagus. Diuji juga performansi Random Forest menggunakan fitur ciri Histogram of Gradient yang sering dipakai untuk mengenali objek dalam citra. Random Forest yang dihasilkan memiliki banyak tree, dan setiap tree ditanam dengan cara yang sama. Tree dengan variabel x akan ditanam sejauh mungkin dengan tree dengan variabel y. Dan dalam perkembangannya, sejalan dengan bertambahnya data set, maka tree pun ikut berkembang. Penempatan tree yang saling berjauhan membuat apabila terdapat tree disekitar treex berarti pohon tersebut merupakan perkembangan dari tree x. Beberapa fungsi learning yang dihasilkan random forest digunakan strategi ensemble ”bagging” untuk mengatasi masalah overfitting apabila dihadapkan data set yang kecil.

Naive Bayes adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class[2]. Naive Bayes didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan decision tree dan neural network. Naive Bayes terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar[4].

$$P(H|X) = \frac{P(X|H)xP(H)}{P(X)}$$

Dasar dari Naïve Bayes yang dipakai dalam pemrograman adalah rumus Bayes:

P (A|B) = (P(B|A) * P(A))/P(B)

Peluang kejadian A sebagai B ditentukan dari peluang B saat A, peluang A, dan peluang B. Pada pengaplikasiannya nanti rumus ini berubah menjadi:

P(Ci|D) = (P(D|Ci)*P(Ci)) / P(D)

Naïve Bayes Classifier atau bisa disebut sebagai Multinomial Naïve Bayes merupakan model penyederhanaan dari Metoda Bayes yang cocok dalam pengklasifikasian teks atau dokumen. Persamaannya adalah:

$$V_{MAP} = \arg\max P(V_j | a_1, a_2, ...a_n)$$

Algoritma Decision Stump(DS) adalah model mesin pembelajaran yang terdiri dari satu tingkat pohon keputusan. Artinya, itu adalah pohon keputusan dengan satu simpul internal (akar) yang langsung tersambung

P-ISSN: 2548-9704
E-ISSN: 2686-0880

dengan node terminal (daunnya). Sebuah tunggul keputusan membuat prediksi berdasarkan pada nilai hanya fitur input tunggal. Kadang-kadang mereka juga disebut aturan.

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, terdapat lima algoritma klasifikasi yaitu Decision Tree (C4.5), k-Nearest Neighbour(k-NN), Naive Bayes (NB), Random Forest (RF), dan Decision Stump (DS).

Tabel 3.1 hasil akurasi

	C4.5	k-NN	RF	NB	DS
Accuracy	97.50%	100%	90.83%	95.00%	85,83%

Berikut hasil performance dari tiap algoritma terdiri dari class precision dan class recall.

1. Decision Tree (C4.5)

Tabel 3.2 hasil C45

accuracy: 97.50% +/- 7.50% (mikro: 97.06%)

	true Sedang	true Rendah	true Tinggi	class precision
pred. Sedang	8	0	0	100.00%
pred. Rendah	1	22	0	95.65%
pred. Tinggi	0	0	3	100.00%
class recall	88.89%	100.00%	100.00%	

2. K-Nearest Neighbor (k-NN)

Tabel 3.3 hasil KNN

accuracy: 100.00% +/- 0.00% (mikro: 100.00%)

	true Sedang	true Rendah	true Tinggi	class precision
pred. Sedang	9	0	0	100.00%
pred. Rendah	0	22	0	100.00%
pred. Tinggi	0	0	3	100.00%
class recall	100.00%	100.00%	100.00%	

3. Random Forest (RF)

Tabel 3.4 hasil RF

accuracy: 90.83% +/- 14.17% (mikro: 91.18%)

	true Sedang	true Rendah	true Tinggi	class precision
pred. Sedang	7	1	0	87.50%
pred. Rendah	2	21	0	91.30%
pred. Tinggi	0	0	3	100.00%
class recall	77.78%	95.45%	100.00%	

4. Naïve Bayes (NB)

Tabel 3.5 hasil NB

accuracy: 95.00% +/- 10.00% (mikro: 94.12%)

	true Sedang	true Rendah	true Tinggi	class precision
pred. Sedang	9	2	0	81.82%
pred. Rendah	0	20	0	100.00%
pred. Tinggi	0	0	3	100.00%
class recall	100.00%	90.91%	100.00%	

5. Decision Stump (DS)

Tabel 3.6 hasil DS

accuracy: 85.83% +/- 18.28% (mikro: 85.29%)

	true Sedang	true Rendah	true Tinggi	class precision
pred. Sedang	4	0	0	100.00%
pred. Rendah	5	22	0	81.48%
pred. Tinggi	0	0	3	100.00%
class recall	44.44%	100.00%	100.00%	

Berikut hasil uji t-Test.

A	B	C	D	E	F
	0.975 +/- 0.075	1.000 +/- 0.000	0.908 +/- 0.142	0.950 +/- 0.100	0.858 +/- 0.183
0.975 +/- 0.075		0.306	0.205	0.535	0.078
1.000 +/- 0.000			0.056	0.131	0.025
0.908 +/- 0.142				0.457	0.503
0.950 +/- 0.100					0.181
0.858 +/- 0.183					

Keterangan:

B = Decision Tree (C4.5)

C = k-Nearest Neighbors (k-NN)

D = Random Forest (RF)

E = Naive Bayes (NB)

F = Decision Stump(DS)

Berdasarkan hasil uji t-Test di atas, maka dapat diterapkan interpretasi H0 (tidak ada perbedaan signifikan) dan Ha(ada perbedaan signifikan). Dimana jika nilai lebih kecil dari alpha (0,05) maka mengindikasikan H0 ditolak berarti ada perbedaan signifikan. Algoritma C4.5 dibandingkan dengan k-NN, RF, NB, DS menghasilkan H0 diterima berarti tidak ada perbedaan signifikan. Algoritma k-NN dibandingkan dengan RF dan NB menunjukkan hasil H0 diterima, dan jika dibandingkan dengan algoritma DS maka H0 ditolak. Kemudian algoritma RF dibandingkan dengan NB maka H0 ditolak, dan dengan DS maka H0 diterima karena lebih dari nilai alpha. Dan yang terakhir algoritma NB.

4. KESIMPULAN

Dari pembahasan dan penjelasan di atas, dapat ditarik kesimpulan bahwa penelitian yang dilakukan yaitu pengujian teknik algoritma klasifikasi terhadap tingkat kemiskinan penduduk ini dengan menggunakan 35 provinsi *data training*

P-ISSN: 2548-9704
E-ISSN: 2686-0880

tingkat kemiskinan penduduk di setiap provinsi yang ada di Indonesia. Hasil akurasi yang tertinggi terdapat pada Algoritma k-Nearest Neighbors, hal ini berarti algoritma ini tepat dan memiliki tingkat akurasi 100% dalam pengelompokkan kategori tingkat kemiskinan penduduk dengan tiga kategori yakni tingkat rendah, sedang, dan tinggi.

Tentunya informasi ini sangat bermanfaat bagi pemerintah dalam mengidentifikasikan sasaran penerima dari banyak program bantuan yang dicanangkan dari pemerintah, salah satunya adalah Program Penanggulangan Kemiskinan.

Untuk kedepannya dapat dikembangkan penelitian lebih lanjut misalnya tentang Beras Miskin (Raskin) atau program lainnya dengan menggunakan algoritma klasifikasi yang tertinggi nilainya yakni algoritma k-Nearest Neighbor(k-NN).

DAFTAR PUSTAKA

[1] H. Harliana and F. N. Putra, "Klasifikasi Tingkat Rumah Tangga Miskin Saat Pandemi Dengan Naïve Bayes Classifier," *J. Sains dan Inform.*, vol. 7, no. 2, pp. 165–173, 2021, doi: 10.34128/jsi.v7i2.339.

[2] M. Rasyida, "Naïve Bayes Classification untuk Penentuan Status Penduduk Miskin," *J. Inform. Kaputama(JIK)*, vol. 4, no. 2, pp. 175–180, 2020.

[3] F. Kurnia, J. Kurniawan, I. Fahmi, and S. Monalisa, "Klasifikasi Keluarga Miskin Menggunakan Metode K- Nearest Neighbor Berbasis Euclidean Distance," no. November, pp. 230–239, 2019.

[4] W. P. Nurmayanti, "Penerapan Naive Bayes dalam Mengklasifikasikan Masyarakat Miskin di Desa Lepak," *Geodika J. Kaji. Ilmu dan Pendidik. Geogr.*, vol. 5, no. 1, pp. 123–132, 2021, doi: 10.29408/geodika.v5i1.3430.

[5] F. N. Umma, B. Warsito, and D. A. I. Maruddani, "Klasifikasi Status Kemiskinan Rumah Tangga Dengan Algoritma C5.0 Di Kabupaten Pemalang," *J. Gaussian*, vol. 10, no. 2, pp. 221–229, 2021, doi: 10.14710/j.gauss.v10i2.29934.

[6] M. Mansyur, "Penerapan Algoritma C4.5 Untuk Klasifikasi Status Kesejahteraan Rumah Tangga Keluarga Binaan Sosial Di Kabupaten Bulukumba," *Inspir. J. Teknol. Inf. dan Komun.*, vol. 9, no. 2, p. 147, 2019, doi: 10.35585/inspir.v9i2.2514.