

---

## KOMPARASI ALGORITMA C4.5 DENGAN NAÏVE BAYES DALAM PENGKLASIFIKASIAN TINGKAT PENDIDIKAN ANAK MISKIN

Andi Nurhayati<sup>1</sup>, Andi Baso Kaswar<sup>2</sup>

<sup>1),2)</sup>*Teknik Informatika Universitas Cokroaminoto Palopo  
Jl Latamcelling No 19 Kota Palopo, 91913*

Email : [andinurhayati991@gmail.com](mailto:andinurhayati991@gmail.com)<sup>1)</sup>, [a.baso.kaswar@gmail.com](mailto:a.baso.kaswar@gmail.com)<sup>2)</sup>

### Abstrak

*Pendidikan merupakan hal sangat penting bagi setiap orang untuk meningkatkan kualitas Sumber Daya Manusia (SDM). Pendidikan merupakan suatu proses untuk mendapatkan pengajaran dan pelatihan, baik itu melalui pendidikan formal, non formal, maupun informal. Oleh karena itu, perludilakukan klasifikasi terhadap tingkat pendidikan anak miskin di Indonesia dengan menggunakan teknik data mining klasifikasi. Penelitian ini bertujuan untuk membandingkan tingkat pendidikan anak miskin dengan tingkat kesejahteraan 30% dengan menggunakan algoritma C 4.5 dan Naïve Bayes. Perbandingan dilakukan untuk mendapatkan algoritma klasifikasi dengan tingkat akurasi, presisi, dan recall yang tinggi. Hasil penelitian menunjukkan bahwa algoritma C4.5 memiliki persentase tingkat akurasi sebesar 90,48% dibandingkan dengan Naïve Bayes yang persentase tingkat akurasinya hanya berkisar 25,32%. Selain itu, nilai presisi dan recall C4.5 yang berkisar antara 0,8-0,9 juga menunjukkan bahwa algoritma C 4.5 lebih baik dibandingkan Naïve Bayes. Sehingga, dalam pengklasifikasian tingkat pendidikan anak miskin, algoritma C4.5 lebih baik dibandingkan Naïve Bayes.s*

**Kata kunci:** C4.5, Naïve Bayes, Pendidikan

### 1. Pendahuluan

Pendidikan merupakan hal yang sangat penting bagi setiap orang untuk meningkatkan kualitas Sumber Daya Manusia (SDM). Pendidikan merupakan suatu proses untuk mendapatkan pengajaran dan pelatihan, baik itu melalui pendidikan formal, non formal, maupun informal. Hasil survey data statistik dari Badan Pusat Statistik (BPS) menunjukkan bahwa dari 261890,90 ribu jiwa penduduk Indonesia pada tahun 2016, ada sekitar 3,9% yang tidak/belum sekolah, 12,27 % yang tidak tamat SD, 33,08% yang pendidikannya hanya sampai Sekolah Dasar, 16,49% yang pendidikannya hanya sampai Sekolah Menengah, dan 34,27% yang pendidikannya sampai Sekolah Tinggi [1]. Melihat data tersebut, dapat dilihat bahwa tingkat pendidikan di Indonesia masih rendah jika dibandingkan dengan negara-negara lainnya. Hal ini didukung dengan fakta yang didapatkan dari OECD (*Organization for Economic Co-operation and Development*), yang menyatakan bahwa pada tahun 2017, Indonesia menempati peringkat ke 57 dari total 65 negara [2]. Sementara di ASEAN, Indonesia menempati peringkat 5 dari 10 negara [3].

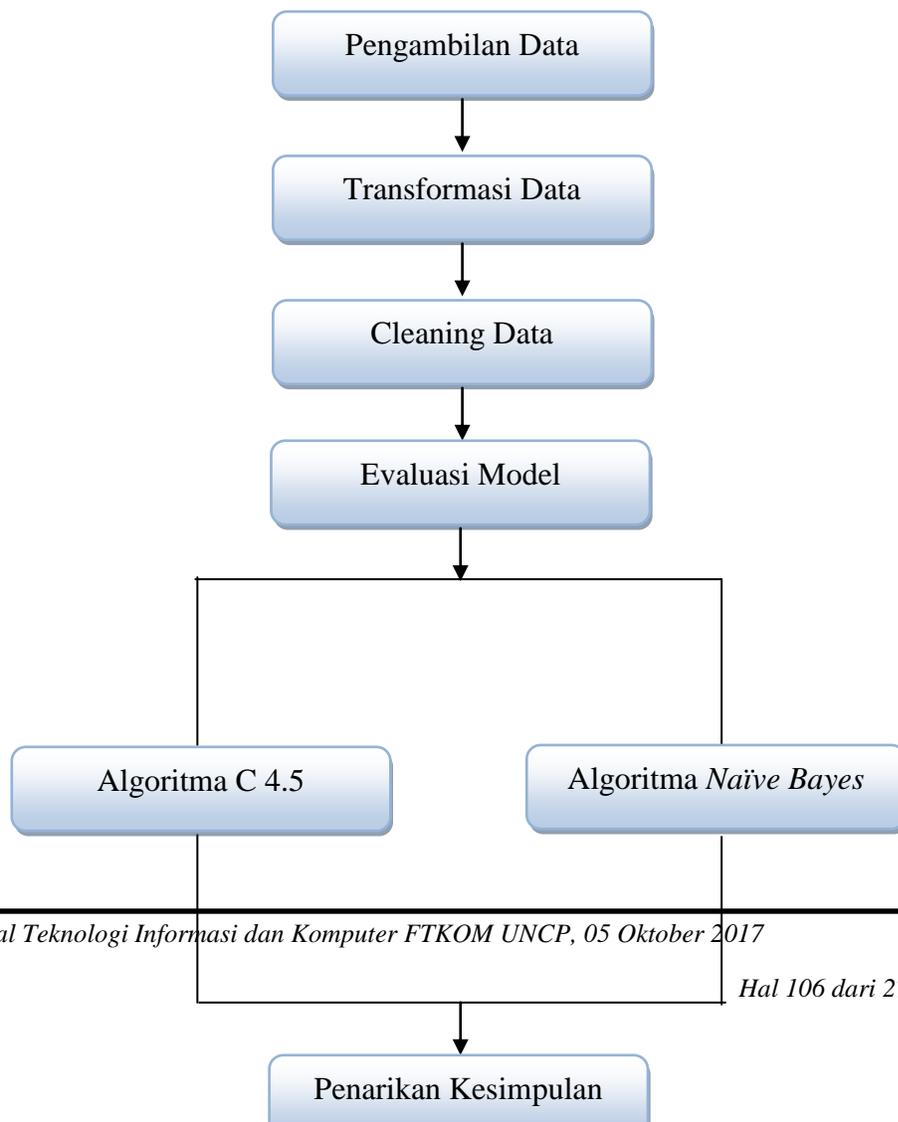
Berdasarkan hal tersebut, perlu dilakukan klasifikasi terhadap tingkat pendidikan anak miskin di Indonesia dengan menggunakan teknik *data mining*. Istilah *data mining* sering juga disebut sebagai *Knowledge Discovery in Database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar[4]. Singkatnya, data mining merupakan suatu

proses untuk mendapatkan pola dari sejumlah data dalam pengambilan keputusan. Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu: Deskripsi, estimasi, prediksi, klasifikasi, pengklusteran, dan asosiasi. Namun, dalam penelitian ini, teknik data mining yang digunakan ialah klasifikasi. Klasifikasi adalah sebuah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data dengan tujuan untuk memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Hal ini juga dapat dikatakan sebagai pembelajaran (klasifikasi yang dapat memetakan sebuah unsur (item) data kedalam salah satu dari beberapa kelas yang sudah didefinisikan [5]. Dalam memecahkan masalah klasifikasi data mining, ada beberapa teknik yang digunakan, seperti: Algoritma CART (*Classification and Regression Trees*), algoritma *k-nearest neighbor*, algoritma C4.5, dan algoritma *Naïve Bayes*, dll.

Pada klasifikasi tingkat pendidikan anak miskin di Indonesia, digunakan algoritma klasifikasi, baik itu algoritma C 4.5 dan *Naive Bayes*. Dalam penelitian ini, kedua metode dibandingkan untuk melihat algoritma klasifikasi yang paling baik digunakan untuk mengklasifikasikan tingkat pendidikan anak miskin di Indonesia.

## 2. Pembahasan

Dalam penelitian ini, terdapat 6 tahapan metodologi yang digunakan, mulai dari pengambilan data, transformasi data, *cleaning* data, evaluasi model, dan penarikan kesimpulan. Lebih jelasnya, hal ini dapat digambarkan dalam bentuk diagram alir sebagai berikut.



Gambar 1. Diagram Alir Metodologi Penelitian

**Pengambilan Data**

Pada tahapan ini dilakukan proses pengambilan data yang didapatkan dari sebuah situs yang dikenal dengan nama Satu Data Indonesia, dengan alamat website [data.go.id](http://data.go.id). Satu Data Indonesia adalah portal resmi data terbuka Indonesia, yang berisi data kementerian, lembaga pemerintahan, pemerintahan daerah, dan semua instansi lain yang terkait yang menghasilkan data yang berhubungan dengan Indonesia. Satu Data adalah sebuah inisiatif Pemerintah Indonesia untuk meningkatkan interoperabilitas dan pemanfaatan data pemerintah. Pemanfaatan data pemerintah tidak terbatas pada penggunaan internal antar instansi, tetapi juga sebagai bentuk pemenuhan kebutuhan data publik bagi masyarakat. [6].

Pada situs Satu Data ini terdapat sekumpulan *dataset* yang ada di Indonesia dan dapat diakses oleh publik. Adapun *dataset* yang dipilih dalam penelitian ini ialah *dataset* tentang tingkat pendidikan anak miskin dengan kesejahteraan 30% di Indonesia. *Dataset* dengan jumlah data sebanyak 463, terdiri atas lima atribut, yakni nama provinsi, jenis kelamin, tingkat pendidikan, jumlah individu, dan kode provinsi.

**Transformasi Data**

Pada tahapan ini dilakukan transformasi *dataset* yang telah didapatkan dari [data.go.id](http://data.go.id) untuk diolah dan dianalisis dengan menggunakan *software WEKA Tools*. *Dataset* yang semula masih berekstensi csv ditransformasikan ke dalam bentuk *dataset* yang berekstensi arff, sehingga *dataset* dapat diproses oleh *WEKA Tools*.

**Cleaning Data**

Sebelum data diolah dan dianalisis dengan menggunakan *software WEKA Tools*, dilakukan *cleaning* data, yakni menghapus beberapa atribut yang tidak diperlukan atau tidak relevan dengan apa yang akan dianalisis. Untuk *dataset* tingkat pendidikan anak miskin, beberapa atribut seperti kode provinsi dan nama provinsi dihapus karena kedua atribut tersebut tidak berpengaruh terhadap proses analisis.

**Evaluasi Model**

Pada tahapan ini *dataset* mulai diolah dan dianalisis dengan menggunakan *software WEKA Tools*. Pengolahan *dataset* dilakukan dengan menggunakan dua metode algoritma klasifikasi yang berbeda, yakni metode algoritma C4.5 dan *Naïve Bayes*. Hasil pengolahan *dataset* berupa tabel informasi, kemudian dilakukan perbandingan terhadap tabel informasi tersebut sehingga dapat ditemukan perbedaan dari kedua metode, baik itu metode algoritma C 4.5 atau *Naïve Bayes*.

**Penarikan Kesimpulan**

Setelah dilakukan evaluasi dari kedua metode yang digunakan, dilakukan penarikan kesimpulan, yakni dengan melihat tingkat akurasi, *presisi*, ataupun *recall* dari kedua metode.

Berdasarkan hasil analisis *dataset* dengan menggunakan software WEKA Tools, hasil klasifikasinya ditampilkan dalam bentuk *confusion matrix* 7 x 7. *Confusion matrix* adalah suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep data mining [7]. Adapun model *confusion matrix*nya ditunjukkan pada tabel 1.

Tabel 1 Model *Confusion Matrix* 7 x 7

		Kelas Hasil Prediksi						
		Kelas 1	Kelas 2	Kelas 3	Kelas 4	Kelas 5	Kelas 6	Kelas 7
Kelas Asli	Kelas 1	K11	K12	K13	K14	K15	K16	K17
	Kelas 2	K21	K22	K23	K24	K25	K26	K27
	Kelas 3	K31	K32	K33	K34	K35	K36	K37
	Kelas 4	K41	K42	K43	K44	K45	K46	K47
	Kelas 5	K51	K52	K53	K54	K55	K56	K57
	Kelas 6	K61	K62	K63	K64	K56	K66	K67
	Kelas 7	K71	K72	K73	K74	K57	K76	K77

Untuk penggambarannya, didapatkan tabel *Confussion Matrix* untuk metode C4.5 dan Naïve Bayes. Hasilnya dapat dilihat pada gambar 2 untuk metode C4.5 dan gambar 3 untuk metode Naïve Bayes.

```

=== Confusion Matrix ===
      a  b  c  d  e  f  g  <-- classified as
63  0  1  2  0  0  0 | a = M. Aliyah
 0 64  1  1  0  0  0 | b = M. Ibtidaiyah
 6  4 55  1  0  0  0 | c = M. Tsanawiyah
 6  8  6 46  0  0  0 | d = Perguruan Tinggi
 0  0  0  0 66  0  0 | e = SD/SDLB/Paket A
 0  3  1  0  0 61  1 | f = SMA/SMK/SMALB/Paket C
 0  0  0  0  1  2 63 | g = SMPT/SMPLB/Paket B
    
```

Gambar 2 *Confussion Matrix* C4.5

```

=== Confusion Matrix ===
      a  b  c  d  e  f  g  <-- classified as
40  0  6 20  0  0  0 | a = M. Aliyah
16  0 10 34  0  2  4 | b = M. Ibtidaiyah
28  0  6 26  0  6  0 | c = M. Tsanawiyah
17  0  5 44  0  0  0 | d = Perguruan Tinggi
 0  1  2  0 18 25 20 | e = SD/SDLB/Paket A
 0  9 27 15  0  7  8 | f = SMA/SMK/SMALB/Paket C
 0  2 27  3  6 26  2 | g = SMPT/SMPLB/Paket B
    
```

Gambar 3 Confussion Matrix Naïve Bayes

Berdasarkan evaluasi dari *confussion matrix*, maka dapat diketahui tingkat akurasi dari masing-masing metode, baik itu metode C4.5 ataupun metode *Naïve Bayes*. Adapun hasil perbandingan tingkat akurasi dari kedua metode tersebut dapat dilihat pada tabel 2.

Tabel 2 Tingkat Akurasi Metode C4.5 dan Naïve Bayes

Metode	Akurasi
C4.5	90.4762 %
Naïve Bayes	25.3247 %

Selain tingkat akurasi, didapatkan pula nilai *presisi* dan *recall* dari masing-masing metode. *Presisi* didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item terpilih. *Presisi* merupakan probabilitas bahwa sebuah item yang dipilih adalah relevan. Dapat diartikan sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan itu [8]. Sementara itu, *recall* adalah probabilitas informasi relevan yang didapatkan kembali oleh sistem dibandingkan jumlah informasi-informasi yang relevan [9]. *Recall* dapat dihitung dengan jumlah rekomendasi yang relevan yang dipilih oleh *user* dibagi dengan jumlah semua rekomendasi yang relevan baik dipilih maupun rekomendasi yang tidak terpilih [8]. Hasil nilai *presisi* dan *recall* berkisar antara 0-1, dimana semakin tinggi nilainya, maka semakin baik pula metode tersebut.

Adapun nilai *presisi* dan *recall* yang diperoleh dari model klasifikasi dari metode C4.5 ditunjukkan pada tabel 3 dan metode Naïve Bayes ditunjukkan pada tabel 4

Tabel 3 Nilai *Presisi* dan *Recall* C4.5

Klasifikasi	<i>Presisi</i>	<i>Recall</i>
M. Aliyah	0.840	0.955
M. Ibtidaiyah	0.810	0.970
M. Tsanawiyah	0.859	0.833
Perguruan Tinggi	0.920	0.697
SD/SDLB/Paket A	0.985	1.000
SMA/SMK/SMALB/Paket C	0.968	0.924
SMPT/SMPLB/Paket B	0.984	0.955

Tabel 4 Nilai *Presisi* dan *Recall* Naïve Bayes

Klasifikasi	<i>Presisi</i>	<i>Recall</i>
M. Aliyah	0.396	0.606
M. Ibtidaiyah	0.000	0.000
M. Tsanawiyah	0.072	0.091
Perguruan Tinggi	0.310	0.667
SD/SDLB/Paket A	0.750	0.273

SMA/SMK/SMALB/Paket C	0.106	0.106
SMPT/SMPLB/Paket B	0.059	0.030

Berdasarkan nilai *presisi* dan *recall* C4.5 dan *Naïve Bayes* seperti yang ditunjukkan pada tabel 3 dan 4, maka dapat dilihat bahwa metode C4.5 memiliki nilai *presisi* dan *recall* yang lebih baik dibandingkan *Naïve Bayes*.

### 3. Kesimpulan

Berdasarkan hasil analisis *dataset* tingkat pendidikan anak miskin dengan menggunakan software WEKA Tools, dihasilkan 7 kelompok kelas klasifikasi, diantaranya: M. Aliyah, M. Ibtidaiyah, M.Tsanawiyah, Perguruan Tinggi, SD/SDLB/Paket A, SMA/SMK/SMALB/Paket C, dan SMPT/SMPLB/Paket B. Selanjutnya hasil perbandingan *dataset* dengan menggunakan kedua metode klasifikasi, yakni algoritma C4.5 dan *Naïve Bayes*, menunjukkan bahwa algoritma C4.5 memiliki tingkat akurasi klasifikasi 65.1515% yang lebih tinggi dibandingkan *Naïve Bayes*. Dari hasil analisis, dapat dilihat pula hasil nilai *presisi* dan *recall* yang cukup jauh dari kedua metode tersebut. Jika nilai *presisi* dan *recall* pada algoritma C4.5 berkisar antara 0,9 – 0,8, maka pada algoritma *Naïve Bayes*, nilai *presisi* dan *recall*nya berkisar antara 0,0-0,3. Sehingga hal tersebut menunjukkan bahwa algoritma C4.5 lebih baik dibandingkan *Naïve Bayes* dalam mengklasifikasi tingkat pendidikan anak miskin.

### Daftar Pustaka

- [1] [BPS] Badan Pusat Statistik. 2016. “*Indikator Pendidikan 1994-2016*”. Jakarta. 1 hal.
- [2] Portal Pendidikan Indonesia. 2017. “*Pendidikan Indonesia Beradadi Peringkat ke 57 Dunia Versi OECD*” <http://edupost.id/internasional/pendidikan-indonesia-berada-di-peringkat-ke-57-dunia-versi-oecd/> Diakses pada 09 Juli 2017
- [3] DW Akademie. 2017. “*Rangking Pendidikan Negara-Negara ASEAN*” <http://www.dw.com/id/rangking-pendidikan-negara-negara-asean/g-37594464> Diakses pada 09 Juli 2017
- [4] Ridwan, Mujib, Suyono, Hadi, dan Sarosa, M. 2013. “*Penerapan Data Mining Untuk Evaluasi Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes Classifier*” Jurnal EECCIS Vol 7 (1): 59-64
- [5] Kusriani, dan Luthfi, Emha Taufiq, 2009. “*Algoritma Data Mining*”. Yogyakarta: C.V Andi Offset.
- [6] Satu Data Indonesia. 2015. “*Tentang Satu Data Indonesia*” <http://data.go.id/tentang/> Diakses pada 09 Juli 2017
- [7] Mayadewi, Paramita dan Rosely, Ely. 2015. “*Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritma Klasifikasi Data Mining*”. Seminar Nasional Sistem Informasi Indonesia. Surabaya. 2-3 November 2015.
- [8] Swastina, Liliana. 2013. “*Penerapan Algoritma C4.5 untuk Penentuan Jurusan Mahasiswa*”. Jurnal GEMA AKTUALITA Vol 2 (1): 93-98
- [9] Anggaraini, Ratih Nur Esti, Zinni, Mohammad Ahmaluddin, dan Rochimah, Siti. 2016. “*Kakas Bantu Pendeteksi Kesalahan Tanda Baca pada Karya Tulis Ilmiah*” Jurnal Ilmiah Teknologi Informasi Vol 14(1): 117-125