

IMPLEMENTASI *DECISION TREE* ALGORITMA C4.5 PADA KLASIFIKASI PENDUDUK MISKIN DI KABUPATEN JEMBER

Deni Arifianto, M.Kom¹, dan Cahya Dwi Rahayu²

¹Universitas Muhammadiyah Jember, Jember, Indonesia

²Universitas Muhammadiyah Jember, Jember, Indonesia

deniariyanto@unmuhjember.ac.id, cahyadwir@gmail.com

Abstract. , Di Kabupaten Jember tingkat kemiskinan masih tergolong tinggi, keterangan dari Dinas Sosial, perkembangan penduduk miskin di Kabupaten Jember dalam kurun waktu 2 tahun terakhir yakni 2016 berjumlah 265,10 ribu jiwa. Dalam pengklasifikasiannya, penduduk kurang mampu didasarkan pada beberapa parameter yaitu pendidikan, pendapatan, luas bangunan rumah, status kepemilikan rumah, jumlah makan dalam sehari, fasilitas MCK dan sumber penerangan. Untuk mengklasifikasikan data penduduk miskin dengan menggunakan 7 parameter, digunakan metode algoritma C4.5 yang mempunyai kehandalan yang baik dalam membuat klasifikasi data. Sebelum diproses menggunakan algoritma C4.5 dilakukan *preprocessing* data dengan mengisi *missing value*. Setelah dilakukan uji coba klasifikasi data penduduk miskin didapatkan tingkat akurasi tinggi sebesar 99,33%. Model ini berhasil diterapkan pada sistem untuk mengklasifikasikan data penduduk miskin yang layak untuk mendapatkan bantuan maupun tidak.

LATAR BELAKANG

Kemiskinan merupakan salah satu masalah yang dialami oleh beberapa Negara berkembang, termasuk Indonesia. Keberadaan penduduk miskin di Indonesia selalu menjadi salah satu objek permasalahan bagi negeri ini. Bagi pemerintah Indonesia masalah kemiskinan merupakan masalah lama yang belum dan sulit untuk diselesaikan. Di wilayah Kabupaten Jember, seorang warga disebut keluarga miskin berdasarkan beberapa aspek seperti aspek pangan, sandang, papan, penghasilan, kesehatan, pendidikan, maupun jumlah jiwa. Begitu banyak bantuan bagi penduduk miskin baik dana yang bersumberkan dari Pemerintah Pusat, seperti jaminan Kesehatan Nasional (JKN), Kartu Indonesia Sehat (KIS), Kartu Indonesia Pintar (KIP) dan lain sebagainya.

Salah satu aspek penting untuk mendukung strategi penanggulangan kemiskinan adalah tersedianya data kemiskinan yang akurat dan tepat sasaran. Seperti yang diungkapkan pada beritasatu.com (Aries Sudiono, 2016) bahwasanya

banyak diantara rumah tangga sasaran di kabupaten Jember yang menerima kurungan dana dari pemerintah itu justru diterima banyak warga yang bukan termasuk kategori miskin sebagaimana dipersyaratkan. Pengukuran kemiskinan yang dapat dipercaya dapat menjadi instrumen tangguh bagi pengambil kebijakan dalam memfokuskan perhatian pada kondisi hidup orang miskin. Data kemiskinan yang baik dapat digunakan untuk mengevaluasi kebijakan pemerintah terhadap kemiskinan, membandingkan kemiskinan antar waktu dan daerah, serta menentukan target penduduk miskin dengan tujuan untuk memperbaiki posisi mereka. Keterangan dari Dinas Sosial, perkembangan penduduk miskin di Kabupaten Jember dalam kurun waktu 2 tahun terakhir yakni 2016 berjumlah 265,10 ribu jiwa. Kesulitan yang dihadapi oleh dinas sosial yaitu kurangnya efisiensi waktu dalam pengelompokan data penduduk yang layak mendapat bantuan maupun yang tidak layak karena dengan banyaknya parameter yang ada, serta permasalahan dana bantuan yang tidak tepat sasaran.

Dalam teknik klasifikasi dikenal banyak algoritma yang dapat digunakan untuk membuat

klasifikasi data. Salah satu algoritma yang terkenal dan mempunyai kehandalan yang baik dalam membuat klasifikasi data adalah Algoritma C4.5. terbukti pada beberapa penelitian menerangkan bahwa klasifikasi menggunakan Algoritma C4.5 memiliki tingkat akurasi tinggi seperti pada penelitian oleh Astuti, dkk yang hasilnya mampu membantu meringankan tugas PLKB dalam mengklasifikasi data kedalam jenis tingkatan kesejahteraan keluarga menggunakan algoritma C4.5, dan memiliki tingkat akurasi sebesar 95% dengan menggunakan data pada kecamatan Pringapus, kabupaten Semarang. Pada penelitian lain oleh M. Ture, dkk membandingkan algoritma *decision tree*, yaitu C&RT, CHAID, QUEST, C4.5 dan ID3 menggunakan 500 data pasien kanker payudara dimana hasil dari penelitian tersebut menunjukkan bahwa algoritma C4.5 memiliki tingkat akurasi yang lebih baik dari metode *decision tree* lainnya.

Oleh karena itu, diharapkan Algoritma C4.5 ini diharapkan dapat membuat keputusan dengan tepat dan sistem yang akan dibuat dapat membantu Dinas Sosial dalam pendataan penduduk miskin yang layak maupun yang tidak layak mendapatkan bantuan.

PENDUDUK MISKIN

Menurut Anshari (2009:3) mengatakan bahwa penduduk miskin merupakan masyarakat yang tidak memiliki akses ke sarana dan prasarana dasar lingkungan yang memadai, dengan kualitas perumahan dan pemukiman yang jauh dibawah standart kelayakan serta mata pencaharian yang tidak menentu.

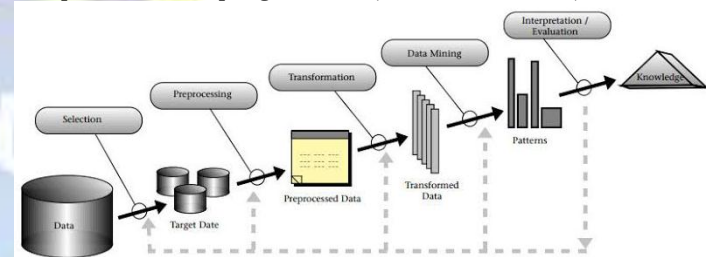
Penyebab kemiskinan menurut Kuncoro (Musa,2014:25) sebagai berikut:

1. Secara makro, penduduk miskin hanya memiliki sumber daya dalam jumlah terbatas dan kualitasnya rendah.
2. Kemiskinan muncul akibat perbedaan kualitas sumber daya manusia karena kualitas sumber daya manusia yang rendah berarti produktifitasnya juga akan rendah dan upahnya pun rendah.
3. Kemiskinan muncul karena adanya akses modal.

DATA MINING

Menurut Turban (Alexander,2005:2) dalam bukunya yang berjudul “*Decision Support Systems and Intelligent Systems*”, data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan didalam basis data. Data mining

merupakan sebuah teknik untuk menggali informasi tersembunyi untuk memperoleh manfaat lebih dari data yang ada. Diantara tujuan data mining adalah untuk melakukan klasifikasi, klasterisasi, menemukan pola asosiasi hingga melakukan peramalan (*predicting*). Istilah data mining lebih populer dengan sebutan KDD (*Knowledge Discovery from Database*). Proses KDD dapat dilihat pada gambar 2.1 mulai dari pemilihan atribut data hingga terciptalah sebuah pengetahuan. (Iskandar dkk, 2013)



Gambar 2.1 KDD Process (Iskandar dkk,2013)

KLASIFIKASI

Pengertian kalsifikasi menurut Bertalya (2009:5) merupakan suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Aturan- aturan tersebut digunakan pada data-data baru untuk diklasifikasi. Teknik ini menggunakan *supervised induction* yang memanfaatkan kumpulan pengujian dari data set yang terklasifikasi.

METODE DECISSION TREE

Decision Tree adalah algoritma yang paling banyak digunakan untuk masalah pengklasifikasian. Sebuah *decision tree* terdiri dari 3 *node*, yaitu :

- a. *Root Node*, merupakan *node* paling atas, pada *node* ini tidak ada input dan bisa tidak mempunyai output atau mempunyai *output* lebih dari satu.
- b. *Internal Node*, merupakan *node* percabangan, pada *node* ini hanya terdapat satu input dan mempunyai output minimal dua.
- c. *Leaf Node* atau *Terminal Node*, merupakan *node* akhir, pada *node* ini hanya terdapat satu *input* dan tidak mempunyai *output*.

Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain ID3, CART, dan C4.5. Algoritma C4.5 merupakan algoritma pengembang dari ID3, Information Gain

pada ID3 lebih mengutamakan pengujian yang menghasilkan banyak keluaran. Dengan kata lain, atribut yang memiliki banyak nilai yang dipilih sebagai splitting atribut. Sebagai contoh, pembagian terhadap atribut yang berfungsi sebagai unique identifier, seperti nomor ID, akan menghasilkan keluaran dalam jumlah yang banyak, dimana setiap keluaran hanya terdiri dari satu tuple, jelas sekali terlihat bahwa partisi semacam ini tidaklah berguna. Karena itu algoritma C4.5 yang merupakan suksesor dari ID3 menggunakan gain ratio untuk memperbaiki information gain.

Algoritma C4.5 dan pohon keputusan merupakan dua model yang tak terpisahkan, karena untuk membangun sebuah pohon keputusan dibutuhkan algoritma C4.5. Ada beberapa tahapan dalam membuat sebuah pohon keputusan dalam algoritma C4.5, yaitu :

- Mempersiapkan data training. Data training biasanya diambil dari data histori yang pernah terjadi sebelumnya atau disebut data masa lalu dan sudah dikelompokkan dalam kelas-kelas tertentu.
- Menghitung akar dari pohon. Akar akan diambil dari atribut yang terpilih, dengan cara menghitung nilai gain ratio dari masing-masing atribut, nilai gain ratio yang paling tinggi akan menjadi akar pertama. Sebelum menghitung nilai *gain ratio* dari atribut, hitung dahulu nilai *entropy*, *information gain*, dan *split info*. Untuk menghitung nilai *entropy* digunakan rumus :

$$Entropy(S) = \sum_{i=1}^n -p_i \log_2 p_i$$

Keterangan :

S = Himpunan Kasus

N = Jumlah partisi S

Pi = proporsi Si terhadap S

Kemudian hitung nilai info gain menggunakan rumus :

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i)$$

Keterangan :

S = Himpunan Kasus

A = Fitur

n = Jumlah partisi atribut A

|Si| = Proporsi Si terhadap S

|S| = Jumlah kasus dalam S

Untuk rumus split info dan gain ratio, sebagai berikut :

$$Split Info(S,A) = - \sum_{i=1}^n \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

$$Gain Ratio(S,A) = \frac{Gain(S,A)}{Split Info(S,A)}$$

c. Ulangi langkah ke 2 dan langkah ke 3 hingga semua record terpartisi.

d. Proses partisi pohon keputusan akan berhenti saat :

Semua record dalam simpul N mendapat kelas yang sama

Tidak ada atribut didalam record yang dipartisi lagi

Tidak ada record di dalam cabang yang kosong

e. *Gain ratio* tertinggi akan menjadi root diawal. Apabila terdapat beberapa atribut dengan *gain ratio* tertinggi, maka atribut yang lebih dulu diperiksa dianggap yang memiliki *gain ratio* tertinggi. (Astuti, 2015:4).

CONFUSION MATRIX

Dalam penelitian ini menggunakan *Confusion Matriks* untuk mengukur hasil klasifikasi. Diantaranya yaitu akurasi, akurasi merupakan pengukuran bagaimana kekakuratan dari suatu klasifikasi. Suatu klasifikasi dikatakan baik jika nilai akurasi mendekati 100%, baik dari percobaan pertama hingga seterusnya.

Berikut Tabel *Confusion Matriks*.

Tabel 2.1 Tabel Confusion Matriks

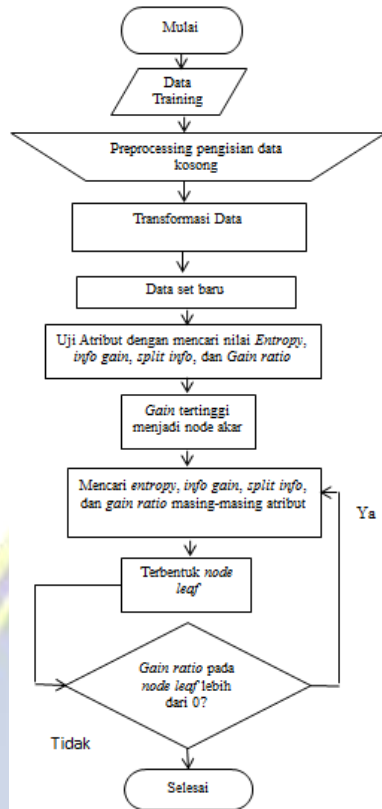
Classification	Predicted class	
	Class = Yes	Class = No
Class = Yes	a (true positive – TP)	c (false negative – FN)
Class = No	b (false positive – FP)	d (true negative – TN)

Dimana untuk mencari nilainya dengan rumus :

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$$

DESAIN

Proses desain merupakan proses penggambaran, perancangan dan pembuatan penerapan Algoritma C4.5 untuk klasifikasi data penduduk miskin. Berikut merupakan *flowchart* dari Algoritma C4.5:



Gambar 3. 1 Flowchart Algoritma C4.5

Berdasarkan flowchart gambar 3.3 dapat diketahui alur algoritma C4.5 yang digunakan. Pada persiapan awal yaitu menyiapkan data yang telah didapat dari hasil survey, kemudian dilakukan *preprocessing* mengisi data kosong dengan cara memberi nilai berdasarkan nilai yang paling banyak atau dominan dalam atribut tersebut, setelah tahapan *preprocessing* selesai dilanjutkan proses transformasi data, kemudian hitung entropy, info gain, split info dan gain ratio masing-masing atribut. Setelah dihitung, gain ratio tertinggi digunakan sebagai node akar/root. Mencari *entropy*, *info gain*, *split info* dan *gain ratio* dari masing-masing atribut yang akan terbentuk *node leaf*, apabila gain ratio pada leaf memiliki nilai lebih dari 0, hitung *entropy*, *info gain*, *split info* dan *gain ratio* masing-masing atribut lagi, jika bernilai 0 maka proses perhitungan selesai.

Proses Perhitungan Algoritma C4.5

Data yang digunakan dalam tugas akhir ini yaitu 100 data training penduduk miskin didapat dari Dinas Sosial Kabupaten Jember. Atribut yang digunakan dalam proses perhitungan Algoritma C4.5 yaitu pendapatan, pendidikan, program yang diterima, jumlah makan dalam sehari, status rumah,

luas bangunan rumah, fasilitas MCK, dan sumber penerangan. Sebelum diproses, lengkapi data dengan mengisi data kosong dengan memberi nilai dominan. Contoh seperti pada contoh data berikut :

No	pendidikan	Program	Status rumah	Luas bangunan	Fasilitas MCK	sumber penerangan	Jumlah makan	pendapatan	kesimpulan
1	sd	cukup	milik sendiri	sedang	tidak punya	450 milik sendiri	sedang	kurang	layak
2	sd		milik sendiri	sedang	punya tidak lay	450 milik sendiri	kurang	kurang	layak
3	sd	kurang	milik sendiri	sedang	punya tidak lay	tidak memiliki	sedang	kurang	layak
4	sd	kurang	milik sendiri	sempit		450 milik sendiri	kurang	kurang	layak
5	tidak sekolah	banyak	menumpang	sedang		450 milik sendiri	lebih	sedang	tidak layak
6	sd	kurang	milik sendiri	sempit	punya tidak lay	450 milik sendiri	cukup	kurang	layak
7	sd	kurang		luas	punya layak	450 milik sendiri	cukup	cukup	tidak layak
8	sd	kurang	milik sendiri	sedang	punya tidak lay	450 milik sendiri	cukup	kurang	layak
9	sma	cukup	sewa	sempit	umum	450 milik sendiri	lebih	sedang	tidak layak
10	sd	banyak	menumpang	sedang	umum	menumpang	cukup	cukup	layak
11	sd	banyak	milik sendiri	sempit	punya layak	450 milik sendiri	cukup	sedang	layak
12	sd	kurang	milik sendiri	sangat luas	punya layak	450 milik sendiri	sedang	tinggi	tidak layak
13	sma	banyak	sewa	luas	punya layak	>450 milik sendiri	sedang	tinggi	tidak layak
14	sd	banyak	sewa	sedang	punya tidak lay	450 milik sendiri	lebih	tinggi	tidak layak
15	sd	sedikit	milik sendiri	luas	punya layak	450 milik sendiri	sedang	sedang	tidak layak

Gambar 3.3 Data Kosong

Terdapat data kosong pada atribut program yang diterima dengan kesimpulan layak diberi bantuan. Maka hitung nilai dominan pada atribut dengan kesimpulan layak, yaitu jumlah data cukup yaitu 1, banyak yaitu 2 dan kurang yaitu 4. Maka data kosong tersebut diisi dengan nilai kurang. Begitu dengan data-data kosong yang lain. Setelah data lengkap lakukan proses transformasi data. Berikut adalah bentuk kategorial dari data penduduk miskin :

1. Pendapatan
 - >1.500.000 : Tinggi
 - 1.000.000-1.500.000 : Sedang
 - 600.000-1.000.000 : Cukup
 - <600.000 : Kurang
2. Jumlah makan dalam sehari
 - > 3 : Lebih
 - 3 : Sedang
 - 2 : Cukup
 - < 2 : Kurang
3. Program yang diterima
 - > 3 : Banyak
 - 3 : Cukup
 - 2 : Kurang
 - <2 : Sedikit
4. Luas bangunan rumah
 - > 12M2 : Sangat Luas
 - Antara 10M2-12M2 : Luas
 - Antara 8M2-10M2 : Sedang
 - < dari 8M2 : Sempit

Setelah semua dikategorikan dengan benar maka dilakukan proses perhitungan mining, dengan menghitung entropy, information gain, split info dan gain ratio. Setelah perhitungan mining selesai, akan menghasilkan suatu pohon keputusan dan rule. Untuk mencari atribut yang menjadi root maka dicari nilai gain ratio tertinggi dari semua atribut. Sebelum menghitung entropy, ingo gain, split info dan gain ratio dari masing-masing atribut maka dihitung jumlah kasus dan jumlah masing-

masing atribut penduduk miskin yang layak dan tidak layak. Terdapat pada tabel 3.2 berikut ini:

Tabel 3.2 Jumlah Kasus Penduduk Miskin

Atribut	Jumlah Kasus	Layak	Tidak Layak
Total	150	97	53
Pendapatan			
Tinggi (>1500000)	22	4	18
Sedang (1000000-1500000)	30	8	22
Cukup (600000-1000000)	25	16	9
Kurang (<600000)	73	69	4
Status Kepemilikan Rumah			
Sewa	15	11	4
Milik Sendiri	123	79	44
Menumpang	12	7	5
Program yang diterima			
Sedikit (<2)	37	19	18
Cukup (3)	29	13	16
Banyak (>3)	39	28	11
Kurang (2)	45	37	8
Luas Bangunan Rumah			
sangat luas >12	18	4	14
luas Antara 10 - 12 M2	14	5	9
sedang 8 M2-10	59	39	20
sempit <8	59	49	10
Pendidikan			
SD	93	59	34
SMP	15	9	6
SMA	12	4	8
tidak sekolah	30	25	5
Jumlah Makan dalam Sehari			
Cukup	8	5	3
Kurang	19	14	5
Sedang	113	74	39
Lebih	10	4	6
MCK			
punya tidak layak	58	52	6
punya layak	55	19	36
umum	15	6	9
tidak punya	22	17	5
Sumber penerangan			
>450 milik sendiri	31	10	21
450 milik sendiri	86	60	26
menumpang	27	23	4
tidak memiliki	6	4	2

Setelah menghitung jumlah kasus, maka selanjutnya menghitung entropy, info gain, split info dan gain ratio untuk masing-masing atribut. Tabel 3.3 menjelaskan perhitungan entropy, info gain, split info dan gain ratio.

Tabel 3.3 Pembentukan Node Akar

Atribut	Jumlah Kasus	Layak	Tidak Layak	Entropy	Information Gain	Split Info	Gain Ratio
Total	150	97	53	0,9370			
Pendapatan					0,3631	1,8070	0,2009
Tinggi (>1500000)	22	4	18	0,6840			
Sedang (1000000-1500000)	30	8	22	0,8366			
Cukup (600000-1000000)	25	16	9	0,9427			
Kurang (<600000)	73	69	4	0,3064			
Status Kepemilikan Rumah					0,0035	0,8585	0,0041
Sewa	15	11	4	0,8366			
Milik Sendiri	123	79	44	0,9408			
Menumpang	12	7	5	0,9799			
Program yang diterima					0,0729	1,9829	0,0368
Sedikit (<2)	37	19	18	0,9995			
Cukup (3)	29	13	16	0,9923			
Banyak (>3)	39	28	11	0,8582			
Kurang (2)	45	37	8	0,6752			
Luas Bangunan Rumah					0,1359	1,7454	0,0779
sangat luas >12	18	4	14	0,7642			
luas Antara 10 - 12 M²	14	5	9	0,9403			
sedang 8 M²-10	59	39	20	0,9238			
sempit <8	59	49	10	0,6565			
Pendidikan					0,0492	1,5157	0,0324
SD	93	59	34	0,9472			
SMP	15	9	6	0,9710			
SMA	12	4	8	0,9183			
tidak sekolah	30	25	5	0,6500			
Jumlah Makan dalam Sehari					0,0157	1,1714	0,0134
Cukup	8	5	3	0,9544			
Kurang	19	14	5	0,8315			
Sedang	113	74	39	0,9296			
Lebih	10	4	6	0,9710			
MCK					0,2000	1,7992	0,1112
punya tidak layak	58	52	6	0,4798			
punya layak	55	19	36	0,9299			
umum	15	6	9	0,9710			
tidak punya	22	17	5	0,7732			
Sumber penerangan					0,0970	1,5613	0,0621
>450 milik sendiri	31	10	21	0,9072			
450 milik sendiri	86	60	26	0,8841			
menumpang	27	23	4	0,8052			
tidak memiliki	6	4	2	0,9183			

Dari hasil perhitungan pada tabel 4.3 diatas, diketahui bahwa nilai gain ratio terbesar yaitu pada atribut pendapatan yaitu sebesar 0,2009 sehingga atribut pendapatan menjadi node akar. Berikut cara perhitungannya :

- a. Cara perhitungan *Node 1*:

Baris total kolom *Entropy* dihitung dengan persamaan :

Entropy (Total)

$$Entropy (S) = \sum_{i=0}^n -p_i \log_2 p_i$$

$$= ((-97/150) * \log_2(97/150)) + ((-53/150) * \log_2(53/150))$$

$$= 0,9370$$

- b. Cara perhitungan *Information Gain*:

Nilai *Information Gain* pada baris pendapatan dihitung :

Information Gain (Pendapatan)

$$Information Gain = Entropy (S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy (S_i)$$

$$= 0,9370 - ((22/150) * 0,6840) + ((30/150) * 0,8366) + ((25/150) * 0,9427) + ((73/150) * 0,3064)$$

$$= 0,3631$$

- c. Cara perhitungan *Split info*:

$$Split Info = - \sum_{i=1}^n \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

$$= -((22/150 * (\log_2(22/150))) + ((30/150 * \log_2(30/150))) + ((25/150 * \log_2(25/150))) + ((73/150 * \log_2(73/150)))$$

$$= 1,8070$$

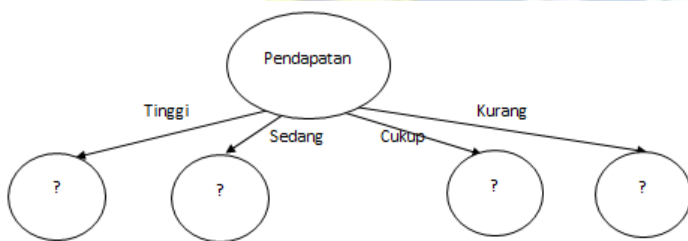
d. Cara perhitungan *Gain ratio* :

$$Gain Ratio = \frac{Info Gain}{Split Info}$$

$$= 0,3631/1,8070$$

$$= 0,2009$$

Dari hasil perhitungan maka didapatkan hasil tree sementara sebagai berikut :



Gambar 3.4 Root Pendapatan

Pada atribut pendapatan terdapat 4 nilai atribut yaitu tinggi, sedang, cukup, kurang. Nilai atribut tinggi, sedang, cukup dan kurang belum mengklasifikasikan kasus menjadi satu keputusan sehingga perlu dilakukan perhitungan lagi.

Interface Program

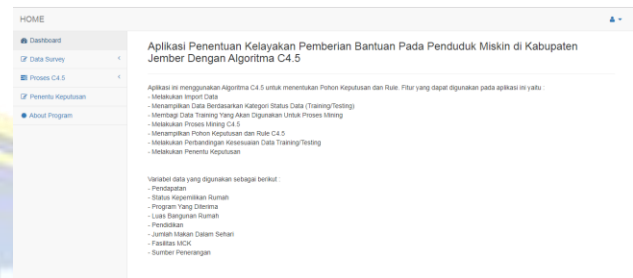
Interface program berisi potongan gambar dari Tugas Akhir Implementasi Algoritma C4.5 Pada Klasifikasi Penduduk Miskin di Kabupaten Jember. Berikut penggunaan program pada saat dijalankan :

Gambar 4.2 Tampilan Login

Untuk dapat mengakses menu pada web, maka user harus login terlebih dahulu, dengan memasukkan

username dan password seperti pada tampilan diatas.

Setelah login, maka akan tampil beberapa menu seperti pada gambar berikut :



Gambar 4.3 Halaman Utama

Pada halaman utama diatas menampilkan deskripsi sederhana tentang aplikasi, dan terdapat beberapa menu, diantaranya dashboard, data survey, data mining C4.5, penentu keputusan, dan about program.

Gambar 4.4 Halaman Import Data

Halaman import data berfungsi melakukan input data yang telah dipersiapkan sebelumnya. Data yang akan dimasukkan hanya file berekstensi xls. Saat proses import data asli yang berupa angka akan terkategoriikan. Data yang terkategoriikan yaitu pendapatan, program yang diterima, luas bangunan rumah dan jumlah makan dalam sehari.

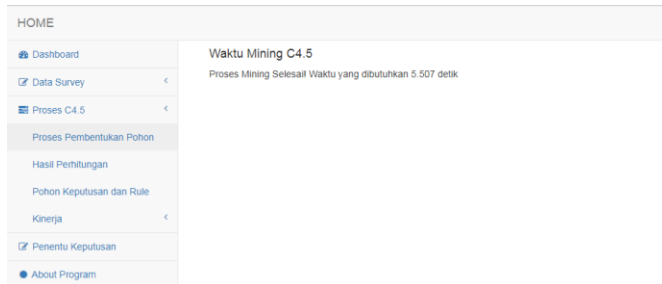
Semua Data Penduduk Miskin

No	Pendapatan	Status Kepemilikan Rumah	Program Yang Diterima	Luas Bangunan Rumah	Pendidikan	Jumlah Makan Dalam Sehari	Fasilitas MCK	Sumber Penerangan	Keputusan	Status Data
1	<600000	milik sendiri	3	8-10 M2	smk	3	punya tidak layak	450 mlk sendiri	layak	data training
2	>1500000	milik sendiri	2	10-12 M2	sma	4	punya layak	>450 mlk sendiri	tidak layak	data training

Gambar 4.5 Menu Lihat Data Survey

Pada menu lihat data survey, akan muncul data asli yang telah diimport pada menu import. Kemudian data yang terperses atau sudah terkategoriikan akan

muncul pada menu data hasil proses -> lihat semua data, seperti pada gambar berikut :



Gambar 4.7 Proses Pembentukan Pohon

Ketika klik menu proses pembentukan pohon, maka akan muncul hasil dari mining pada menu hasil perhitungan seperti pada gambar dibawah ini

HOME

Dashboard

Data Summary

Perhitungan C4.5

Proses Pembelajaran Pohon

Hasil Pembinaan

Polisi Kabupaten dan Kota

Kinerja

Presentasi Kependidikan

Audit Program

Perhitungan C4.5

	Attribut Gain Ratio	Attribut	Nilai Attribut	Jumlah Kasus Total	Jumlah Kasus Layak	Jumlah Kasus Tidak Layak	Entropy	Information Gain	Split Info	Gain Ratio
1	pendapatan	Total	Total	130	97	33	0.937			0
2	pendapatan	pendapatan	tinggi	22	4	18	0.684	0.3631	1.927	0.2009
3	pendapatan	pendapatan	sedang	30	8	22	0.696	0.3631	1.927	0.2009
4	pendapatan	pendapatan	cukup	25	16	9	0.9427	0.3631	1.927	0.2009
5	pendapatan	pendapatan	kurang	73	69	4	0.3064	0.3631	1.927	0.2009
6	pendapatan	status_rumah	sewa	15	11	4	0.686	0.0035	0.8585	0.0041
7	pendapatan	status_rumah	milik sendiri	123	79	44	0.9438	0.0035	0.8585	0.0041
8	pendapatan	status_rumah	menumpang	12	7	5	0.9799	0.0035	0.8585	0.0041
9	pendapatan	program_diterima	sedikit	37	19	18	0.9995	0.0739	1.9628	0.0368
10	pendapatan	program_diterima	cukup	29	13	16	0.9923	0.0739	1.9628	0.0368
11	pendapatan	program_diterima	banyak	39	28	11	0.6192	0.0739	1.9628	0.0368

Pada menu hasil perhitungan menampilkan nilai hasil perhitungan C4.5. Dapat dilihat jumlah data yang telah dihitung, *entropy*, *information gain*, *split info* dan *gain ratio*. Data tersebut akan terbentuk pohon keputusan, *gain ratio* tertinggi akan dijadikan sebagai *root*.

Pohon Keputusan dan Rule

Pohon Keputusan:

```
pendapatan == tinggi (layak = 4, tidak layak = 18) : ?
|
| sumber_penerangan == >450 milik sendiri (layak = 0, tidak layak = 11) : tidak layak
|
| sumber_penerangan == 450 milik sendiri (layak = 3, tidak layak = 6) : ?
|
| | fasilitas_mck == punya tdk layak (layak = 2, tidak layak = 0) : layak
|
| | fasilitas_mck == punya layak (layak = 1, tidak layak = 5) : ?
|
| | pendidikan == sd (layak = 0, tidak layak = 4) : tidak layak
|
| | pendidikan == smp (layak = 0, tidak layak = 0) : Null
|
| | pendidikan == sma (layak = 0, tidak layak = 1) : tidak layak
|
| | pendidikan == tidak sekolah (layak = 1, tidak layak = 0) : layak
|
| | fasilitas_mck == umum (layak = 0, tidak layak = 1) : tidak layak
|
| | fasilitas_mck == tidak punya (layak = 0, tidak layak = 0) : Null
|
| sumber_penerangan == menumpang (layak = 1, tidak layak = 0) : layak
|
| sumber_penerangan == tidak memiliki (layak = 0, tidak layak = 1) : tidak layak
pendapatan == sedang (layak = 8, tidak layak = 22) : ?
|
| fasilitas_mck == punya tdk layak (layak = 3, tidak layak = 0) : layak
|
| fasilitas_mck == punya layak (layak = 5, tidak layak = 18) : ?
|
| | program_diterima == sedikit (layak = 1, tidak layak = 8) : ?
|
| | pendidikan == sd (layak = 0, tidak layak = 5) : tidak layak
|
| | pendidikan == smp (layak = 0, tidak layak = 0) : Null
|
| | pendidikan == sma (layak = 0, tidak layak = 3) : tidak layak
|
| | pendidikan == tidak sekolah (layak = 1, tidak layak = 0) : layak
```

Gambar 4.9 Pohon Keputusan

Halaman ini menampilkan hasil pembentukan pohon keputusan yang tela dibentuk dari perhitungan C4.5.

No	Pendapatan	Status Kepemilikan Rumah	Program Yang Diterima	Luas Bangunan Rumah	Pendidikan	Jumlah Makan Dalam Sehari	Fasilitas MCK	Sumber Penerangan	Keputusan	Status Data
1	kurang	milik sendiri	3	sedang	smp	sedang	punya tidak layak	400 milik sendiri	layak	data training
2	tinggi	milik sendiri	kurang	luas	sma	lebih	punya layak	>450 milik sendiri	tidak layak	data training

Selain itu juga menampilkan *rule* yang diperoleh dari nilai pohon keputusan. Untuk keputusan layak akan ditampilkan dengan warna hijau dan keputusan tidak layak akan ditampilkan dengan warna merah.

Rule:

```
1. IF (pendapatan == tinggi AND sumber_penerangan == >450 milik sendiri) THEN tidak layak (id = 2)
2. IF (pendapatan == tinggi AND sumber_penerangan == 450 milik sendiri AND fasilitas_mck == punya tdk layak) THEN layak (id = 4)
3. IF (pendapatan == tinggi AND sumber_penerangan == 450 milik sendiri AND fasilitas_mck == punya layak AND pendidikan == sd) THEN tidak layak (id = 6)
4. IF (pendapatan == tinggi AND sumber_penerangan == 450 milik sendiri AND fasilitas_mck == punya layak AND pendidikan == smp) THEN Null (id = 7)
5. IF (pendapatan == tinggi AND sumber_penerangan == 450 milik sendiri AND fasilitas_mck == punya layak AND pendidikan == sma) THEN tidak layak (id = 8)
6. IF (pendapatan == tinggi AND sumber_penerangan == 450 milik sendiri AND fasilitas_mck == punya layak AND pendidikan == tidak sekolah) THEN layak (id = 9)
7. IF (pendapatan == tinggi AND sumber_penerangan == 450 milik sendiri AND fasilitas_mck == umum) THEN tidak layak (id = 10)
8. IF (pendapatan == tinggi AND sumber_penerangan == 450 milik sendiri AND fasilitas_mck == tidak punya) THEN Null (id = 11)
9. IF (pendapatan == tinggi AND sumber_penerangan == menumpang) THEN layak (id = 12)
10. IF (pendapatan == tinggi AND sumber_penerangan == tidak memiliki) THEN tidak layak (id = 13)
11. IF (pendapatan == sedang AND fasilitas_mck == punya tdk layak) THEN layak (id = 15)
12. IF (pendapatan == sedang AND fasilitas_mck == punya layak AND program_diterima == sedikit AND pendidikan == sd) THEN tidak layak (id = 18)
13. IF (pendapatan == sedang AND fasilitas_mck == punya layak AND program_diterima == sedikit AND pendidikan == smp) THEN Null (id = 19)
14. IF (pendapatan == sedang AND fasilitas_mck == punya layak AND program_diterima == sedikit AND pendidikan == sma) THEN tidak layak (id = 20)
15. IF (pendapatan == sedang AND fasilitas_mck == punya layak AND program_diterima == sedikit AND pendidikan == tidak sekolah) THEN layak (id = 21)
16. IF (pendapatan == sedang AND fasilitas_mck == punya layak AND program_diterima == cukup) THEN tidak layak (id = 22)
17. IF (pendapatan == sedang AND fasilitas_mck == punya layak AND program_diterima == banyak AND sumber_penerangan == >450 milik sendiri AND status_rumah == sewa) THEN Null (id = 25)
```

HOME

Dashboard

Data Survey

Proses C4.5

Proses Pembentukan Pokon

Hasil Pembentukan

Pohon Keputusan dan Rule

Kinerja

Kinerja C4.5

Tabel Penilaian

Penentu Keputusan

About Program

Tabel Perbandingan Keputusan

Lakukan Proses Training | Hapus Semua Data

No	Pendapatan	Status Kepemilikan Rumah	Program Yang Diterima	Luas Bangunan Rumah	Pendidikan	Jumlah Makan Dalam Sehari	Fasilitas MCK	Sumber Penerangan	Keputusan Asli	Keputusan [tree] ID Rule
1	sedang	milik sendiri	sedikit	sedang	tidak sekolah	sedang	punya layak	400 mlak sendiri	tidak layak	tidak layak 11
2	cukup	milik sendiri	kurang	luas	sd	sedang	punya layak	400 mlak sendiri	tidak layak	tidak layak 11
3	cukup	milik sendiri	sedikit	sempit	sd	sedang	punya layak	400 mlak sendiri	tidak layak	tidak layak 18
4	kurang	milik sendiri	kurang	sempit	sd	sedang	punya layak	400 mlak sendiri	layak	layak 28
5	sedang	milik sendiri	kurang	luas	sma	sedang	punya layak	>400 mlak sendiri	tidak layak	tidak layak 6

Gambar 4.11 Halaman Kinerja

Halaman kinerja berfungsi untuk menghitung kesesuaian data berdasarkan keputusan yang telah dibentuk dari proses mining.

Tabel Penilaian Akurasi

	Diidentifikasi Tidak Layak oleh C4.5	Diidentifikasi Layak oleh C4.5
Keputusan Asli: Tidak Layak = 53	52	1
Keputusan Asli: Layak = 97	0	97

Akurasi = $(52 + 97) / (52 + 53) * 100\% = 99,33\%$

Gambar 4.12 Tabel Penilaian

Halaman tabel penilaian menampilkan nilai akurasi kinerja data testing atau data *training*. Data tersebut dibandingkan dengan *rule* dari proses mining yang telah dilakukan.

Penentu Keputusan

Penentu Keputusan

Pendapatan: Sedang (1000000-1500000)

Status Kepemilikan Rumah: Milik Sendiri

Program Yang Diterima: Sedikit (<2)

Luas Bangunan Rumah: Sedang (5-10 M2)

Pendidikan: Tidak Sekolah

Jumlah Makan Dalam Sehari: Sedang

Fasilitas MCK: Punya Layak

Sumber Penerangan: 400 Watt Milik Sendiri

Proses | Hapus Semua Data

Data Keputusan:

No	Pendapatan	Status Kepemilikan Rumah	Program Yang Diterima	Luas Bangunan Rumah	Pendidikan	Jumlah Makan Dalam Sehari	Fasilitas MCK	Sumber Penerangan	Keputusan C4.5 ID Rule	Ops
1	sedang	milik sendiri	sedikit	sedang	tidak sekolah	sedang	punya layak	400 mlak sendiri	tidak layak 11	Hapus

Gambar 4.13 Penentu Keputusan

Halaman penentu keputusan berfungsi untuk memberi keputusan apa yang sesuai terhadap data *training* yang diuji.

K-Fold Validation

Dalam melakukan pengujian dengan menggunakan *k-fold validation*, pembagian data *traning* dan data testing dilakukan dengan mengambil k sama dengan 2, 3, 5, 6 dan 10. Jumlah

data yang digunakan adalah sebanyak 300 data. Berikut hasil dari proses validasi yang telah dilakukan:

2 Fold

Menggunakan 150 data testing dan 150 data training

Tabel 4.1 2 Fold

Data		TP	TN	FP	FN	Akurasi
150	150	97	52	1	0	99,33 %
150	150	98	50	2	0	98,67 %



= Data Testing

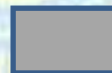
= Data Training

3 Fold

Menggunakan 100 data testing dan 200 data *training*

Tabel 4.2 3 Fold

Data			TP	TN	FP	FN	Akurasi
100	100	100	128	66	5	1	97 %
100	100	100	117	73	6	4	95 %
100	100	100	138	58	2	2	98 %



= Data Testing



= Data Training

5 Fold

Menggunakan 60 data testing dan 240 data *training*

Tabel 4.3 5 Fold

Data					TP	TN	FP	FN	Akurasi
60	60	60	60	60	116	48	5	1	96,47%
60	60	60	60	60	169	64	7	0	97,08 %
60	60	60	60	60	131	96	9	4	94,58 %
60	60	60	60	60	167	67	5	1	97,5 %
60	60	60	60	60	155	75	10	0	95,83 %



= Data Testing



= Data Training

6 Fold

Menggunakan 60 data testing dan 240 data *training*

Tabel 4.4 6 Fold

Data						TP	TN	FP	FN	Akurasi
50	50	50	50	50	50	153	84	9	4	94,8 %
50	50	50	50	50	50	166	78	5	1	97,6 %
50	50	50	50	50	50	162	81	6	1	97,2 %
50	50	50	50	50	50	152	92	5	1	97,6 %
50	50	50	50	50	50	168	70	9	3	95,2 %
50	50	50	50	50	50	162	75	11	2	94,8 %

 = Data Testing  = Data Training

10 Fold

Menggunakan 10 data testing dan 90 data *training*

Tabel 4.5 10 Fold

Data										TP	TN	FP	FN	Akurasi
30	30	30	30	30	30	30	30	30	30	168	92	8	2	96,3%
30	30	30	30	30	30	30	30	30	30	174	85	7	4	95,93%
30	30	30	30	30	30	30	30	30	30	176	84	8	2	96,3 %
30	30	30	30	30	30	30	30	30	30	184	77	7	2	96,67 %
30	30	30	30	30	30	30	30	30	30	164	99	6	1	97,41 %
30	30	30	30	30	30	30	30	30	30	163	94	11	2	95,19%
30	30	30	30	30	30	30	30	30	30	179	85	5	1	97,78 %
30	30	30	30	30	30	30	30	30	30	180	78	9	3	95,56 %
30	30	30	30	30	30	30	30	30	30	172	87	10	1	95,93 %
30	30	30	30	30	30	30	30	30	30	173	83	10	4	94,81 %

 = Data Testing
 = Data Training

Dari hasil perhitungan diatas dengan menggunakan 2,3,5,6 dan 10 *fold*, terdapat akurasi tertinggi pada 2k -*fold* yang terletak pada percobaan ke 2 yaitu dengan 150 data *training* dan 150 data testing, mendapatkan hasil dengan TP = 97, TN = 52, FP = 1,

dan FN = 0 dan mendapatkan akurasi senilai 99,33 %. *Tree* yang dihasilkan dapat dilihat pada lampiran.

Hasil Validasi

Setelah mendapatkan hasil dari masing-masing *k-fold*, maka hasil tersebut divalidasi dengan data yang ditetapkan sebagai data validasi yaitu sebanyak 100 data Berikut hasilnya :

Tabel 4.6 Hasil Validasi

Klasifikasi	Tidak Layak	Layak
Tidak Layak	47	5
Layak	0	48

Maka akurasi yang didapatkan dari hasil perhitungan yaitu sebesar 95%.

Kesimpulan

Berdasarkan hasil pengujian tugas akhir yang berjudul Implementasi Decission Tree Algoritma C4.5 Pada Klasifikasi Penduduk Miskin di Kabupaten Jember secara fungsional berjalan dengan baik. Untuk pengujian dengan menggunakan 300 data penduduk miskin dan 100 data untuk data validasi, didapatkan rule dan pohon keputusan dari hasil proses perhitungan mining C4.5. Hasil uji coba sistem mengolah 300 data dengan 150 data training dan 150 data testing tingkat akurasi tertinggi didapat pada skenario 2 fold uji ke 1 yaitu sebesar 99,33%. Dan terdapat 53 penduduk yang diidentifikasi tidak layak serta 97 penduduk yang layak diberi bantuan sedangkan untuk hasil dari validasi data mendapat tingkat akurasi sebesar 95%.

Saran

Adapun saran untuk penelitian selanjutnya yang akan membahas atau mengembangkan topik yang sama dengan penelitian ini agar lebih baik lagi, antara lain :

1. Disarankan untuk menambahkan variabel lain yang lebih mendukung untuk diproses dalam penentuan

2. Untuk penelitian lebih lanjut disarankan untuk mencoba algoritma lain, membandingkan dengan algoritma lain atau meningkatkan jumlah data kasusnya dengan algoritma yang sama.



DAFTAR PUSTAKA

Abidatul Izzah, dkk. 2014. *Imputasi Missing Data Menggunakan Metode K-Nearest Neighbour Dengan Optimasi Algoritma Genetika*. Malang: Universitas Muhammadiyah Malang.

Acuna, dkk. 2004. *The treatment of missing values and its effect on classifier accuracy, Classification, Clustering, and Data Mining Applications*. Springer Berlin Heidelberg.

Aldi Nurzahputra, dkk. 2017. *Peningkatan Akurasi Pada Algoritma C4.5 Menggunakan Adaboost Untuk Meminimalkan Resiko Kredit*. Semarang: Universitas Negeri Semarang.

Alexander Setiawan. 2009. *Implementasi Aplikasi Decision Support System Dengan Metode AHP Untuk Penentuan Jenis Supplier*. Surabaya: Universitas Kristen Petra Surabaya.

Anhar. 2010. *PHP dan MySql Secara Otodidak*. Jakarta : PT TransMedia.

Anshari, Akhnis Rasyid. 2009. *Potret Masyarakat Miskin Sebagai Kritik Sosial Dalam Karya Seni Gambar*. Semarang: Universitas Negeri Semarang.

Aries Sudiono. 2016. *Penerima dana PSKS di Kabupaten Jember*. Jember

Astuti, Lilik Dwi dan Hanny Haryanto. 2015. *Metode Pohon Keputusan Menggunakan Algoritma C4.5 Untuk Pengelompokan Data*

Penduduk Pada Tingkatan Kesejahteraan Keluarga. Semarang: Universitas Dian Nuswantoro.

Bertalya. 2009. *Konsep Data Mining*. Depok: Universitas Gunadarma

Iskandar, Derick dan Yoyon K. Suprpto. 2013. *Perbandingan Akurasi Klasifikasi Tingkat Kemiskinan antara Algoritma C4.5 dan Naive Bayes Classifier*. Surabaya: Institut Teknologi Sepuluh November.

Jogiyanto, HM. 2005. *Analisa dan Desain Sistem Informasi: Pendekatan Terstruktur Teori dan Praktik Aplikasi Bisnis*. Andi: Yogyakarta

Jundi, Musa Al. 2014. *Analisis Faktor Yang Mempengaruhi Tingkat Kemiskinan Provinsi-Provinsi di Indonesia*. Semarang: Universitas Diponegoro.

Melvut Ture, dkk. 2009. *Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients*. Turki: Trakya University

Nursela, Dwi Ayu. 2014. *Penerapan Algoritma C4.5 Untuk Klasifikasi Tingkat Keganasan Kanker Payudara*. Semarang: Universitas Dian Niswantoro.

Turban, E. 2005. *Decision Support Systems and Intelligent Systems Edisi Bahasa Indonesia Jilid I*, Andi, Yogyakarta

Wahyu Dwi Lesmono.2016. *Analisis dan Eksplorasi Data-Transformasi Data dan Statistika Inferensia* versi 1.01.Bogor:Universitas Pakuan Bogor.

