Penerapan Algoritma Klasifikasi Data Mining C4.5 Pada Dataset Cuaca Wilayah Bekasi

Adhika Novandya[1], Isni Oktria[2]
Program Studi Manajemen Informatika, AMIK Bina Sarana Informatika[1]
Jurusan Teknik Informatika, Fakultas Teknologi Industri, Universitas Gunadarma[2]
Email: Adhika.avn@bsi.ac.id[1], isnioktria@staff.gunadarma.ac.id[2]

Abstrak

Cuaca adalah keadaan udara pada saat tertentu dan pada wilayah tertentu yang relatif sempit dan pada jangka waktu yang singkat. Prakiraan cuaca pada umumnya sering disebut peramalan cuaca yang merupakan penggunaan ilmu dan teknologi untuk memperkirakan atmosfer bumi pada masa akan datang untuk suatu tempat tertentu. Data yang digunakan pada penelitian didapat dari *World Weather Online,* merupakan sebuah situs yang memberikan data dan informasi mengenai kondisi cuaca sehari-hari. Data yang dipakai memiliki interval waktu setiap 3 jam terhitung mulai tanggal 12 Agustus 2016 pukul 01.00 hingga tanggal 20 Agustus 2016 pukul 22.00. Penelitian bertujuan untuk mendapatkan pola klasifikasi cuaca dengan menggunakan algoritma klasifikasi data mining yaitu algoritma C4.5. Hasil pengujian algoritma C4.5 menggunakan 10-*fold cross validation* dan dibuktikan dengan pembuatan aplikasi web untuk pengujian sehingga menghasilkan nilai akurasi sebesar 88.89%.

Kata Kunci: Cuaca, Prakiraan Cuaca, Data Mining, Algoritma Klasifikasi

Abstract

The weather is the state of the air at a certain point and to a particular region a relatively narrow and in a short time frame. Weather reports in general often called weather forecasting is the use of science and technology to estimate the atmosphere of the earth in the will come to a particular place. The data used to research obtained from world weather online, is a the site that providing data and information on weather conditions daily. Data used having intervals time every three hours starting from August 12, 2016 at 01.00 up to August 20, 2016 at 10 pm. Research aims to get pattern classifications weather with data mining algorithm classifications namely c4.5 algorithm. The results of testing algorithm c4.5 use 10-fold cross validation and provable to the creation of application web for testing so as to produce value accuracy of 88.89 %.

Keywords: Weather, Weather Forecasting, Data Mining, Classification Algorithm

1 PENDAHULUAN

Cuaca adalah keadaan udara pada saat tertentu dan pada wilayah tertentu yang relatif sempit dan pada jangka waktu yang singkat. Menurut *World Climate Conference*, cuaca adalah keadaan atmosfer secara menyeluruh termasuk perubahan, perkembangan, dan menghlangnya suatu fenomena.

Banyaknya parameter dalam menentukan suatu cuaca menyebabkan ketepatan dan kecepatan dalam memprediksikan cuaca kurang terpenuhi. Metode klasifikasi data mining merupakan sebuah teknik yang dilakukan untuk memprediksi *class* atau properti dari data itu sendiri. Adapun metode klasifikasi data mining memiliki beberapa algoritma salah satunya yaitu algoritma C4.5.

Penelitian ini bertujuan untuk mendapatkan pola klasifikasi cuaca yang terjadi dengan jumlah interval selama 3 jam sekali, sehingga nantinya dapat digunakan untuk memprediksi cuaca pada keesokan harinya atau dalam periode waktu tertentu.

2 STUDI LITERATUR

2.1 Penelitian Sebelumnya

Proses prakiraan cuaca memerlukan banyak komponen data cuaca, jumlah data yang besar serta kemampuan prakirawan. Hal tersebut menyebabkan ketepatan dan kecepatan prakiraan kurang terpenuhi. Untuk memecahkan masalah tersebut, dilakukan penelitian model prediksi menggunakan beberapa teknik data mining yaitu *Association rule*, C4.5, *Classification* dan *Random Forest*. Penelitian menghasilkan bahwa model prediksi C4.5 memiliki tingkat akurasi 68.5% [1].

Peramalan cuaca merupakan suatu proses memprediksikan bagaimana kondisi atmosfir berubah. Untuk memprediksi suatu cuaca digunakan algoritma *decision tree* untuk mengklasifikasikan parameter cuaca seperti temperatur maksimum, temperatur minimum, curah hujan, penguapan, dan kecepatan angin dengan menggunakan data dari situs cuaca *wondergound* mulai dari tahun 2001 sampai 2013. Hasilnya didapat bahwa parameter tersebut mempunyai pengaruh yang berarti [2].

Peramalan cuaca adalah aplikasi yang paling penting dalam meteorologi dan telah menjadi salah satu yang paling ilmiah dan menjadi pemasalahan teknologi yang menantang. Algoritma klasifikasi pohon keputusan C5 digunakan untuk menghasilkan pohon keputusan dan aturan klasifikasi parameter cuaca pada data yang didapat dari stasiun meteorologi Ibadan dari tahun 2000 sampai 2009, Artificial Neural Networks (ANN) dapat mendeteksi hubungan antara variabel inpu dan menghasilkan output berdasarkan pola observasi data [3].

Algoritma *FP Growth* digunakan untuk menghasilkan pohon keputusan. Data yang digunakan didapat dari departemen cuaca Nagpur periode 2010 sampai dengan 2014. Algoritma *FP growth* dengan evaluasi MAE, MSE, dan SD menampilkan hasil yang lebih akurat dibandingkan dengan algoritma Neural Net (NN), dimana FP Growth menghasilkan prediksi curah hujan yang benar setiap bulannya [4].

2.2 Data Mining

Data mining adalah proses penting dimana metode kecerdasan diaplikasikan untuk mengekstrak pola data [5].

Data mining adalah analisis observasional sekumpulan data untuk menemukan hubungan tidak terduga dan untuk meringkas data dengan cara baru yang dapat dipahami dan berguna bagi pemilik data [6].

2.3 Klasifikasi

Klasifikasi merupakan suatu proses menemukan kumpulan pola atau fungsi yang mendeskripsikan serta memisahkan kelas data yang satu dengan yang lainnya untuk menyatakan objek tersebut masuk pada kategori tertentu yang sudah ditentukan.

Klasifikasi adalah bentuk analisis data yang mengekstrak model yang menggambarkan kelas data [5].

2.4 Algoritma C4.5

Algoritma C4.5 adalah ekstensi Quinlan untuk algoritma ID3 untuk menghasilkan pohon keputusan, algoritma C4.5 rekursif mengunjungi setiap node keputusan, memilih split optimal sampai tidak ada perpecahan lanjut yang memungkinkan [7].

Pada dasarnya konsep dari algoritma C4.5 adalahmengubah data menjadi pohon keputusan dan aturan-aturan keputusan (*rule*). C4.5 adalah algoritma yang cocok untuk masalah klasifikasi dan data mining. C4.5 memetakan nilai atribut menjadi kelas yang dapat diterapkan untuk klasifikasi baru [8].

Ada beberapa tahapan dalam membangun sebuah pohon keputusan dengan algoritma C4.5 yaitu [9].

1. Menyiapkan data *training*. Data *training* biasanya diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon. Akar akan diambil dari atribut yang terpilih, dengan cara menghitung nilai *gain* dari masing-masing atribut, nilai *gain* yang paling tinggi yang akan menjadi akar pertama.
3. Sebelum menghitung nilai *gain* dari atribut, hitung dahulu nilai entropi. Untuk menghitung nilai entropi digunakan rumus:

$$Entropy\ (S) = \sum_{i=1}^{N} - pi\ \log 2\ pi \quad (1)$$

Dimana:

S = Himpunan Kasus
n = Jumlah partisi S
Pi = Proporsi Si terhadap S

4. Kemudian hitung nilai *gain* yang menggunakan rumus:

$$Gain(S,A) = entropy(S) - \sum_{i=1}^{n} \frac{|si|}{s} * entropy(Si) \quad (2)$$

Dimana:

S = Himpunan Kasus
A = Fitur
n = Jumlah Partisi Atribut A
|Si| = Proporsi Si terhadap S
|S| = Jumlah Kasus dalam S

5. Ulangi langkah ke-2 hingga semua *record* terpartisi.
6. Proses partisi pohon keputusan akan berhenti saat:
   a. Semua *record* dalam simpul N mendapat kelas yang sama.
   b. Tidak ada atribut di dalam *record* yang dipartisi lagi.
   c. Tidak ada *record* di dalam cabang yang kosong.

2.5 Model Validasi

Peneliti menggunakan *stratified 10-fold cross-validation* untuk melakukan pengetesan terhadap *dataset*. Peneliti melakukan 10 kali pengetesan terhadap data untuk melihat performa dari masing-masing algoritma klasifikasi yang digunakan. Adapun bentuk model *stratified 10 fold cross validation* dapat dilihat pada Tabel 1.

Tabel 1. Stratified 10 Fold Cross Validation [10].

n-validation	Dataset's Partiiton									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

2.6 Model Evaluasi

Model evaluasi yang digunakan oleh peneliti yaitu *Confusion Matrix* yang menghasilkan nilai akurasi dari validasi algoritma terhadap *dataset* yang ada. *Confusion Matrix* adalah alat visualisasi yang biasa digunakan pada *supervised learning*. Tiap kolom pada matriks adalah contoh kelas prediksi, sedangkan tiap baris mewakili kejadian di kelas yang sebenarnya [11].

Hasil dari proses perhitungan *confusion matrix* yaitu 4 keluaran diantaranya *recall*, *precision*, *accuracy*, dan *error rate*. Dapat dilihat pada Tabel 2.

Tabel 2. Confusion Matrix

		Prediksi	
Aktual	Negatif	A	C
	Positif	B	D

Keterangan:
1. A = jumlah prediksi yang tepat bersifat negatif.
2. B = jumlah prediksi yang salah bersifat positif.
3. C= jumlah prediksi yang salah bersifat negatif.
4. D = Jumlah prediksi yang tepat bersifat positif.

Beberapa persyaratan yang telah didefinisikan untuk matrik klasifikasi diantaranya sebagai berikut:
1. *Accuracy* merupakan proporsi jumlah prediksi benar. Rumus akurasi adalah:

$$\mathbf{AC = (A + D) / A + B + C + D}$$

2. *Recall* atau tingkat positif benar (TP) adalah proporsi kasus positif yang diidentifikasi dengan benar, yang dapat dihitung dengan persamaan:

$$\mathbf{TP = D / C + D}$$

3. Tingkat positif salah (FP) adalah proporsi kasus negatif yang salah diklasifikasikan sebagai positif, yang dapat dihitung dengan menggunakan persamaan:

FP = B / A+B

4. Tingkat negatif sejati (TN) didefinisikan sebagai proporsi kasus negatif yang diklasifikasikan dengan benar, dapat dihitung dengan menggunakan persamaan:

TN = A / A+B

5. Tingkat negatif palsu (FN) adalah proporsi kasus positif yang salah diklasifikasikan sebagai negatif, yang dihitung dengan menggunakan persamaan:

FN = C / C +D

6. *Precision* (P) adalah proporsi prediksi kasus positif yang benar, yang dihitung dengan menggunakan persamaan:

P = D / B + D

3 HASIL PENELITIAN

3.1 *Dataset*

Dataset yang digunakan pada penelitian ini dapat dilihat pada gambar 1.

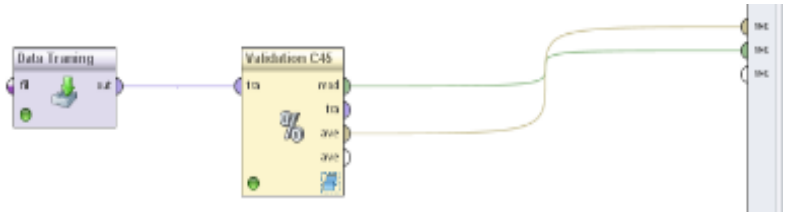
1	Date	Time	Desc	Weather	Temp (Celcius)	Rain (mm)	Wind (mph)	Dir	Cloud (%)	Humidity (%)	Pressure (mb)
2	12 Agustus 2016	1	Night	clear	26	0.0	5	S	5	80	1013
3	12 Agustus 2016	4	Morning	clear	26	0.0	4	S	2	83	1012
4	12 Agustus 2016	7	Morning	sunny	29	0.0	4	S	6	75	1013
5	12 Agustus 2016	10	Daylight	sunny	34	0.0	1	WSW	4	53	1013
6	12 Agustus 2016	13	Daylight	sunny	36	0.0	4	NNW	23	43	1011
7	12 Agustus 2016	16	Afternoon	partly cloudy	33	0.0	12	NNW	29	57	1010
8	12 Agustus 2016	19	Night	patchy rain nearby	30	0.4	7	NNW	21	74	1011
9	12 Agustus 2016	22	Night	patchy rain nearby	27	0.1	3	ENE	25	78	1013
10	13 Agustus 2016	1	Night	clear	27	0.0	4	SE	8	79	1012
11	13 Agustus 2016	4	Morning	clear	26	0.0	4	SE	18	80	1010
12	13 Agustus 2016	7	Morning	sunny	29	0.0	4	SE	24	74	1012
13	13 Agustus 2016	10	Daylight	sunny	34	0.1	3	NE	21	55	1013
14	13 Agustus 2016	13	Daylight	patchy rain nearby	35	0.8	6	N	30	49	1010
15	13 Agustus 2016	16	Afternoon	heavy rain shower	33	2.7	6	N	25	60	1008
16	13 Agustus 2016	19	Night	light rain shower	29	1.0	4	NNE	44	78	1010
17	13 Agustus 2016	22	Night	patchy rain nearby	27	2.3	5	ENE	17	81	1012
18	14 Agustus 2016	1	Night	partly cloudy	26	0.0	3	ESE	25	84	1011
19	14 Agustus 2016	4	Morning	patchy rain nearby	25	0.1	4	S	21	85	1010
20	14 Agustus 2016	7	Morning	patchy rain nearby	28	0.3	5	SSE	19	79	1012
21	14 Agustus 2016	10	Daylight	sunny	34	0.0	2	E	14	55	1012
22	14 Agustus 2016	13	Daylight	sunny	35	0.0	5	NNW	22	47	1009
23	14 Agustus 2016	16	Afternoon	patchy rain nearby	33	1.1	8	NNW	28	57	1008

Gambar 1. *Dataset*

Dataset dibuat oleh peneliti berdasarkan informasi yang ditampilkan pada situs *World Weather Online*. *Dataset* memiliki beberapa atribut diantaranya yaitu *Date*, *Time*, *Desc*, *Weather*, *Temp* (*Celcius*), *Rain* (*mm)*, *Wind* (*mph*), *Dir*, *Cloud* (%), *Humidity* (%), dan *Pressure* (*mdb*). Atribut yang menjadi *class* atau label pada *dataset* yaitu atribut *weather*.

3.2 Pemodelan Proses

Bentuk pemodelan proses yang digunakan pada penelitian menggunakan software Rapid Miner Studio dimana model proses dapat dilihat pada gambar 2.

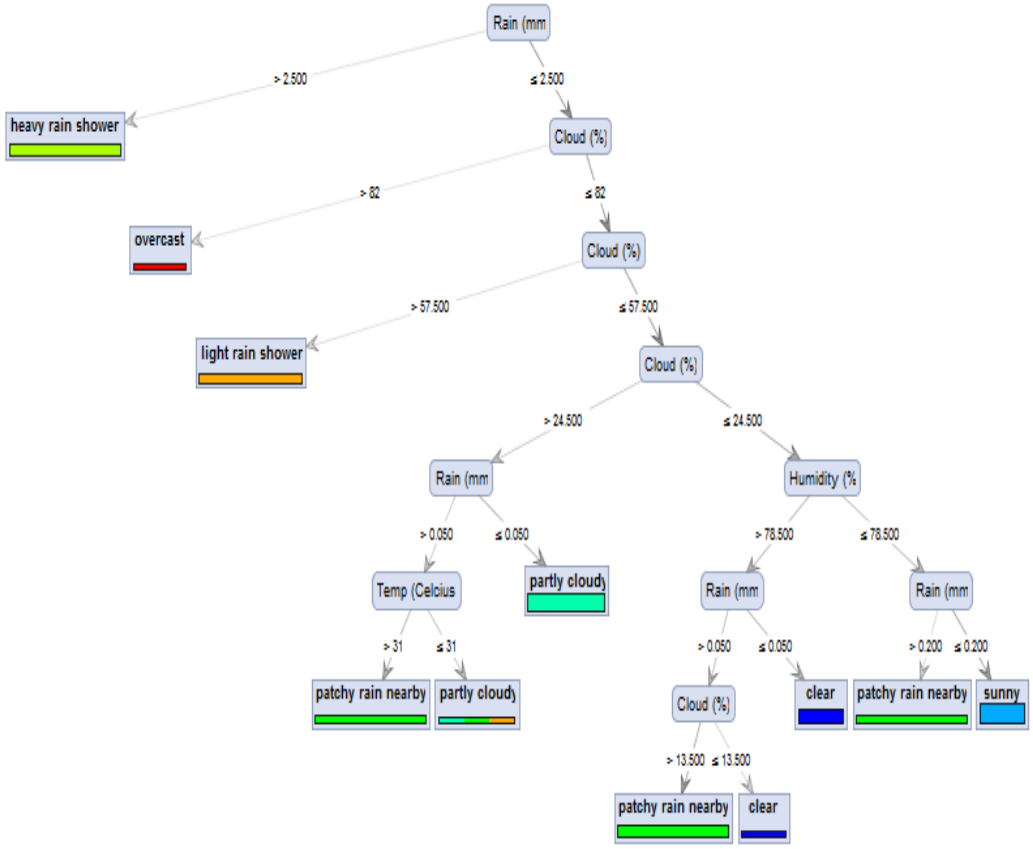


Gambar 2. Pemodelan Proses

Berdasarkan gambar 2 dapat dijelaskan dengan menggunakan *software Rapid Miner*, *dataset* yang digunakan diuji dengan menggunakan operator *X-Validation* dimana tipe sampling yaitu *stratified* dengan jumlah validasi yang dilakukan terhadap *dataset* sebanyak 10 kali.

3.3 Pola Pohon Keputusan

Setelah melewati proses pengujian, maka dihasilkan sebuah pola yang berbentuk pohon keputusan (*decision tree*) dikarenakan algoritma C4.5 termasuk ke dalam algoritma *decision tree*. Pohon keputusan yang dihasilkan dapat dilihat pada gambar 3.



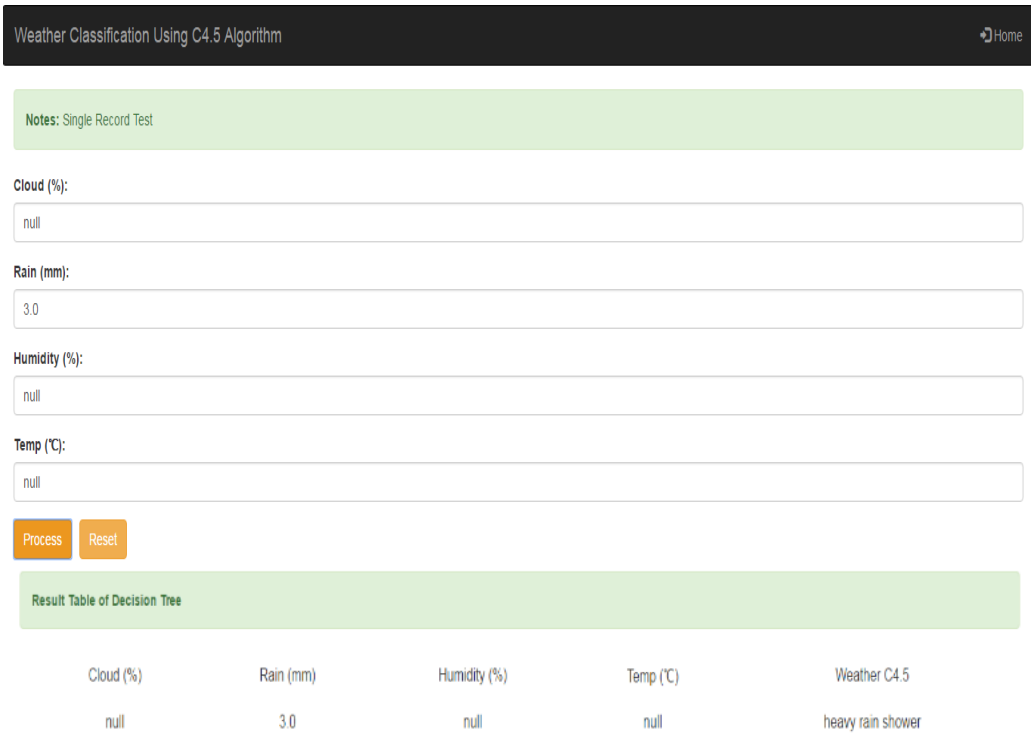
Gambar 3. Pohon Keputusan

Berdasarkan gambar 3 dapat diketahui bahwa atribut-atribut apa saja yang dapat memberikan pengaruh dalam melakukan proses pengklasifikasian *dataset* sehingga menyebabkan terbentuknya pola pengetahuan yang dapat digunakan sebagai acuan dalam memprediksikan cuaca yang akan datang. Atribut tersebut adalah *Rain*, *Cloud*, *Humidity*, dan *Temp*.

3.4 Pengujian dengan Aplikasi Web

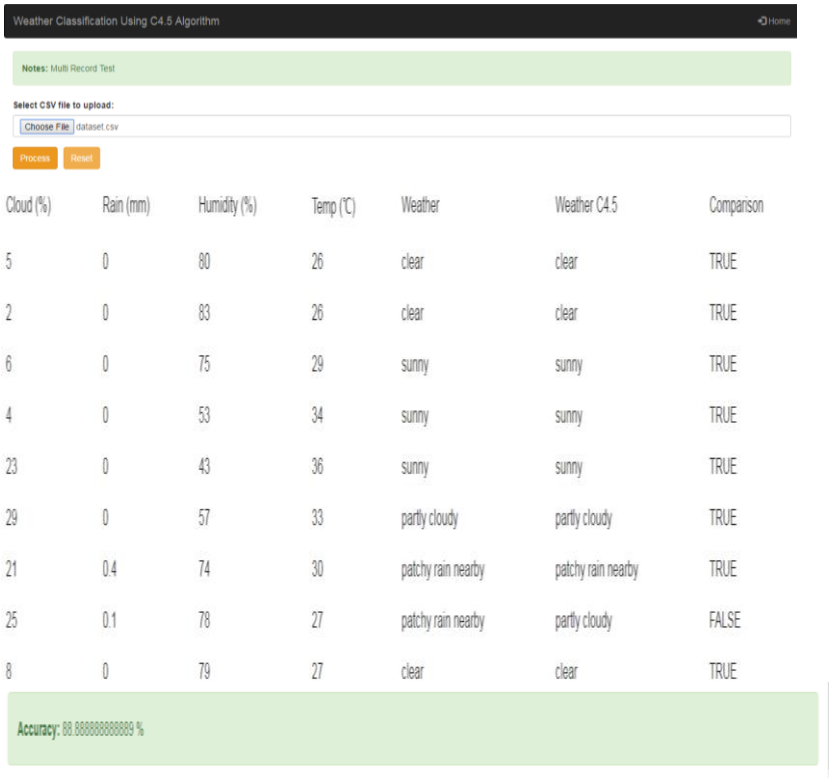
Setelah mendapatkan pola pengetahuan klasifikasi cuaca, penelitian dibuktikan dalam bentuk aplikasi berbasis web menggunakan bahasa pemrograman PHP. Aplikasi web tersebut terdiri dari 2 konten halaman yaitu halaman *single record test* dan halaman *multi record test*. Halaman *single record test* digunakan untuk membandingkan klasifikasi cuaca *user* dengan klasifikasi cuaca hasil dari algoritma C4.5 dengan hanya menggunaka satu *record* data yang dapat dilihat pada gambar 4.

Berdasarkan gambar 4, terdapat 4 input data yang dimasukkan oleh *user*. Sebagai contoh, *user* menginput nilai atibut *rain* sebesar 3.0, dan atibut lainnya dengan nilai *null*. Sesuai dengan pola pengetahuan yang didapatkan maka algoritma menghasilkan cuaca yaitu *heavy rain shower*.



Gambar 4. *Single Record Test*

Berikutnya yaitu pengujian untuk *multi record test* yang dapat dilihat pada gambar 5.



Gambar 5. *Multi Record Test*

Berdasarkan gambar 5, *user* dapat memasukkan sebuah file bertipe CSV yang didalamnya terdiri dari banyak *record* data, dimana data tersebut nantinya akan diuji dengan algoritma yang ada dan didapatkan nilai akurasinya.

4 PENUTUP

4.1 Kesimpulan

Penelitian menggunakan *dataset* yang dibentuk dari informasi yang dihasilkan pada situs peramalan cuaca yaitu *World Weather Online* terhitung sejak tanggal 12 Agustus 2016 pukul 01:00 sampai dengan 20 Agustus 2016 pukul 22:00. Akurasi dari algoritma klasifikasi C4.5 menghasilkan nilai sebesar 88.89% yang telah dibuktikan melalui program yang dibuat.

4.2 Saran

Pengembangan pekerjaan yang akan datang dapat mempertimbangkan tidak hanya nilai *accuracy* dan *kappa* dari algoritma tersebut, tetapi memperhatikan nilai AUC yang dihasilkan. Untuk meningkatkan *accuracy* maka dapat digunakan metode optimasi salah satunya dengan menggunakan metode PSO (*Particle Swarn Optimization*) agar hasil yang didapat lebih akurat.

REFERENSI

[1] S. Mujiasih, "Utilization Of Data Mining For Weather Forecasting`," *J. Meteorol. dan Geofis.*, vol. 12, no. 2, pp. 189–195, 2011.

[2] A. Joshi, B. Kamble, V. Joshi, K. Kajale, and N. Dhange, "Weather Forecasting and Climate Changing Using Data Mining Application Rain Effects on Speed," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 3, pp. 19–21, 2015.

[3] F. Olaiya, "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies," *I.J. Inf. Eng. Electron. Bus.*, vol. 1, no. February, pp. 51–59, 2012.

[4] A. A. Taksande and P. S. Mohod, "Applications of Data Mining in Weather Forecasting Using Frequent Pattern Growth Algorithm," *Int. J. Sci. Res.*, vol. 4, no. 6, pp. 3048–3051, 2015.

[5] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques 3rd Edition*. USA: Morgan Kauffman, 2012.

[6] D. T. Larose, *Data Mining Methods and Models*. New Jersey, 2006.

[7] D. T. Larose, *Discovering knowledge in data*. New Jersey, 2005.

[8] W. K. V. Xindong, *The Top Ten Algorithms in Data Mining*, vol. XXXIII, no. 2. USA: Taylor & Francis Group, LLC, 2009.

[9] E. T. L. Kusrini, *Algoritma Data Mining*. Yogyakarta: Penerbit Andi, 2009.

[10] R. S. Wahono, N. S. Herman, and S. Ahmad, "A comparison framework of classification models for software defect prediction," *Adv. Sci. Lett.*, vol. 20, no. 10–12, pp. 1945–1950, 2014.

[11] F. Goronescu, *Data Mining: Concepts, Models and Techniques*. Verlag Berlin Heidel: Springer, 2011.