

KLASIFIKASI DATA PRODUKSI PADI PULAU JAWA

MENGGUNAKAN ALGORITME *CLASSIFICATION VERSION 4.5 (C4.5)*

Dwi Setyowati, Yuliana Susanti, Supriyadi Wibowo

Program Studi Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Sebelas Maret

ABSTRAK. Padi merupakan tanaman pangan yang banyak diproduksi masyarakat Indonesia. Jumlah penduduk Indonesia yang semakin meningkat mengakibatkan semakin tinggi kebutuhan konsumsi pangan khususnya padi. Kebutuhan pangan yang meningkat harus diimbangi dengan jumlah produksi yang meningkat pula. Untuk menghindari jumlah impor beras yang berlebihan, perlu dilakukan optimalisasi produksi padi di kota/kabupaten. Penelitian ini bertujuan mengklasifikasi data produksi padi Pulau Jawa. Metode yang digunakan adalah klasifikasi dengan Algoritme *C4.5*. Data yang digunakan bersumber dari Badan Pusat Statistik. Berdasarkan hasil penelitian, klasifikasi data produksi padi Pulau Jawa menghasilkan model pohon keputusan dengan 9 aturan klasifikasi dan 4 klasifikasi status kota/kabupaten. Tingkat keakuratan yang diperoleh dari pohon keputusan sebesar 83.19% sehingga pohon keputusan yang terbentuk cukup layak dalam mengklasifikasi produksi padi Pulau Jawa.

Kata kunci: *klasifikasi, produksi padi, algoritme C4.5*

1. PENDAHULUAN

Padi merupakan tanaman pangan yang dikonsumsi oleh mayoritas masyarakat Indonesia. Sampai saat ini ketergantungan masyarakat Indonesia terhadap tanaman pangan khususnya padi masih sangat besar. Menurut Las [9], padi merupakan makanan pokok yang dikonsumsi lebih dari setengah penduduk dunia dan padi mengandung sekitar 60-70% kalori. Produksi padi di Indonesia dipengaruhi oleh faktor luar (eksternal) dan dalam (internal). Menurut Ina [7], faktor luar yang mempengaruhi produksi padi yaitu keadaan iklim dan geografis. Keadaan iklim meliputi curah hujan, suhu, cahaya matahari, air, dan musim. Keadaan geografis meliputi ketinggian tempat dan luas lahan. Menurut Hanum [6], curah hujan yang cocok untuk pertumbuhan tanaman padi yaitu 1500-2000 mm/tahun. Padi tumbuh subur pada ketinggian 0-650 mdpl dengan suhu 22°C-27°C serta pada ketinggian 650-1500 mdpl dengan suhu 19°C-23°C.

Indonesia juga merupakan penghasil komoditas pangan terbesar terutama padi. Hampir di setiap daerah menghasilkan tanaman pangan padi, salah satunya

di Pulau Jawa. Badan Pusat Statistik (BPS) mencatat bahwa angka produksi padi tahun 2015 mengalami kenaikan dibandingkan tahun 2014, yaitu sebanyak 2,31 juta ton [1]. Produksi padi yang tidak diimbangi dengan banyaknya kebutuhan padi dapat berdampak pada impor beras. Oleh karena itu, untuk menghindari impor beras perlu dilakukan strategi yaitu mengoptimalkan produksi padi kota/kabupaten di Pulau Jawa dengan teknik klasifikasi. Klasifikasi merupakan salah satu teknik dalam data mining yang menggunakan pohon keputusan untuk membentuk data dalam grup atau kelas. Salah satu algoritme yang digunakan dalam pohon keputusan adalah Algoritme *C4.5*. Pada penelitian ini, dilakukan penerapan Algoritme *C4.5* untuk mengklasifikasi data produksi padi Pulau Jawa tahun 2015 yang menghasilkan pohon keputusan dan aturan klasifikasi, sehingga dapat mengetahui status kota/kabupaten di Pulau Jawa.

2. DATA MINING

Data mining didefinisikan sebagai proses menemukan pola dalam data. Proses ini bekerja secara otomatis atau semi otomatis dan menghasilkan pola yang memberikan kemudahan bagi peneliti. Pola tersebut diidentifikasi, divalidasi, dan digunakan untuk membuat sebuah prediksi (Witten *et al.* [12]). Sebagaimana ditulis oleh Larose [8], salah satu teknik yang dimiliki data mining berdasarkan tujuan yang dicapai, yaitu klasifikasi. Klasifikasi merupakan proses yang bertujuan membedakan kelas data dalam beberapa kategori. Menurut Han dan Kamber [5], pohon keputusan adalah salah satu metode klasifikasi paling populer dan mudah untuk diinterpretasikan. Pohon keputusan juga berguna untuk mengeksplorasi data dan menemukan hubungan tersembunyi antara sejumlah variabel *input* (*independent*) dengan sebuah variabel target (*dependent*).

3. ALGORITME *C4.5*

Algoritme *Classification Version 4.5* (*C4.5*) merupakan pengembangan dari Algoritme *Iterative Dichotomiser 3* (*ID3*) yang diperkenalkan oleh Quinlan tahun 1993 (Quinlan [10]). Idenya yaitu membuat pohon keputusan dengan simpul awal sebagai atribut yang paling signifikan. Kelebihan dari Algoritme *C4.5* yaitu dapat menangani data diskrit dan kontinyu. Langkah-langkah menggunakan Algoritme *C4.5* dalam membuat sebuah pohon keputusan (Florin [4]) yaitu:

1. Mempersiapkan data *training*, yaitu data yang sudah dikategorikan ke dalam kelas-kelas tertentu.
2. Menghitung nilai *gain ratio* sebagai percabangan awal. Ada beberapa langkah untuk menghitung nilai *gain ratio* sebagai berikut.

- a. *Entropy*. *Entropy* adalah ukuran keberagaman dalam suatu kumpulan data. Semakin tinggi nilai *entropy* maka semakin tinggi tingkat keberagaman.

$$Entropy(S) = \sum_{i=1}^n -p_i \times \log_2(p_i) \quad (2.1)$$

dengan S adalah himpunan kasus, n adalah banyaknya kelas, dan p_i adalah proporsi S_i terhadap S .

- b. *Gain*. *Gain* adalah ukuran efektivitas dari atribut data.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (2.2)$$

dengan A adalah atribut, $|S_i|$ adalah banyaknya kasus dalam S_i , dan $|S|$ adalah banyaknya kasus dalam S .

- c. *Split Information*. Untuk menghitung *gain ratio* perlu diketahui suatu term baru yang disebut pemisahan informasi (*split information*).

$$Split Information = \sum_{j=1}^n -p_j \times \log_2(p_j) \quad (2.3)$$

dengan p_j adalah proporsi S_j terhadap S .

- d. *Gain Ratio*. *Gain ratio* digunakan dalam menentukan atribut sebagai simpul (Wei Ji, et al. [11]).

$$Gain Ratio = \frac{Gain}{Split Information} \quad (2.4)$$

3. Ulangi langkah ke 2 dan langkah ke 3 hingga semua *record* terpartisi.
4. Proses partisi pohon keputusan akan berhenti saat semua *record* dalam simpul A mendapat kelas yang sama.

4. CONFUSION MATRIX

Menurut Bramer [2], *confusion matrix* merupakan tabel yang mencatat hasil kerja klasifikasi. Metode ini menggunakan tabel matriks seperti pada Tabel 1.

Tabel 1. Model *Confusion Matrix*

Correct Classification	Classified Matrix	
	+	-
+	<i>true positive</i>	<i>false negative</i>
-	<i>false Positive</i>	<i>true negative</i>

Pengujian data berdasarkan nilai akurasi, yaitu ukuran dari seberapa baik model mengkorelasi antara hasil dengan atribut dalam data (Han dan Kamber [5]).

$$Akurasi (\%) = \frac{\text{Jumlah data yang benar}}{\text{jumlah data yang diuji}} \times 100\% \quad (2.5)$$

Semakin tinggi nilai akurasi maka semakin tinggi ketepatan model pohon keputusan dalam melakukan proses klasifikasi.

5. METODE PENELITIAN

Metode yang diterapkan dalam penelitian ini adalah Algoritme C4.5. Langkah-langkah menggunakan Algoritme C4.5 sebagai berikut.

1. Menghitung jumlah kasus untuk data produksi di atas median atau di bawah median pada setiap atribut.
2. Menghitung nilai *entropy*, *gain*, *split information*, dan *gain ratio* menggunakan persamaan (2.1), (2.2), (2.3), dan (2.4).
3. Menetapkan salah satu atribut dengan nilai *gain ratio* tertinggi sebagai akar dan atribut dengan nilai *gain ratio* tertinggi berikutnya sebagai cabang.
4. Proses akan berhenti apabila semua kasus pada cabang memiliki kelas yang sama.

Hasil yang diperoleh yaitu pohon keputusan dan menginterpretasikan pohon keputusan berupa aturan klasifikasi. Untuk mengetahui keakuratan dari pohon keputusan yang terbentuk, maka dilakukan pengujian akurasi terhadap data yang telah diklasifikasi menggunakan Algoritme C4.5.

6. HASIL DAN PEMBAHASAN

6.1. Deskripsi Atribut Data. Atribut yang digunakan terdiri dari suhu, curah hujan, luas panen, dan ketinggian wilayah. Atribut berfungsi sebagai variabel *input* yang digunakan dalam penentuan variabel target (produksi padi) yang ditunjukkan pada Tabel 2.

7. Tabel 2. Kategori Variabel Target

Jumlah Produksi (ton)	Produksi Padi
≥ 291378	di atas median
< 291378	di bawah median

Selain variabel target, variabel *input* juga dikategorikanke dalam kelas-kelas tertentu sebagai tahapan proses klasifikasi. Pengkategorian variabel *input* berdasarkan pendapat ahli yang ditunjukkan pada Tabel 3 (Hanum [6]).

Tabel 3. Kategori Variabel *Input*

Variabel	Kategori	Keterangan
Luas Panen	> 84803 Ha	Luas
	13044 Ha – 84803 Ha	Sedang
	< 13044 Ha	Sempit
Suhu	Ketinggian 0 – 650 mdpl	
	> 27°C	Tinggi
	22°C – 27°C	Cukup
Curah Hujan	Ketinggian 650 – 1500 mdpl	
	> 23°C	Tinggi
	19°C – 23°C	Cukup
Ketinggian	> 2000 mm/tahun	Tinggi
	1500 – 2000 mm/tahun	Cukup
	< 1500 mm/tahun	Rendah
Ketinggian	0 – 650 mdpl	Rendah
	650 – 1500 mdpl	Tinggi

7.1. Analisis Algoritme C4.5. Berikut adalah penjelasan dalam pembentukan pohon keputusan menggunakan Algoritme C4.5.

1. Menentukan *node* akar. Langkah awal yang dilakukan yaitu menghitung nilai *entropy*, *gain*, *split information*, dan *gain ratio*.

Iterasi	Atribut	Gain	Split Info	Gain Ratio
1	Suhu	0.0039	0.8598	0.0045
	Curah Hujan	0.0229	1.3714	0.0167
	Ketinggian	5.48E-06	0.4314	1.27E-06
	Luas Panen	0.4956	1.4955	0.3314

Perhitungan iterasi pertama diperoleh nilai *gain ratio* tertinggi yaitu luas panen sebesar 0.3314. Kategori luas dan sempit sudah mengklasifikasikan kasus ke dalam produksi padi di atas dan di bawah median. Kategori sedang perlu dilakukan iterasi karena belum dapat diklasifikasikan.

2. Menentukan *node* cabang. Penentuan *node* cabang berdasarkan nilai *gain ratio* tertinggi setelah menghapus atribut yang sudah terpilih sebagai *node* akar.

Iterasi	Atribut	Gain	Split Info	Gain Ratio
2	Suhu	0.0294	0.7746	0.0379
	Curah Hujan	0.0565	1.2108	0.0466
	Ketinggian	0.0133	0.3666	0.0362

Perhitungan iterasi kedua diperoleh nilai *gain ratio* tertinggi yaitu curah hujan sebesar 0.0466. Kategori rendah sudah mengklasifikasikan kasus ke dalam produksi di bawah

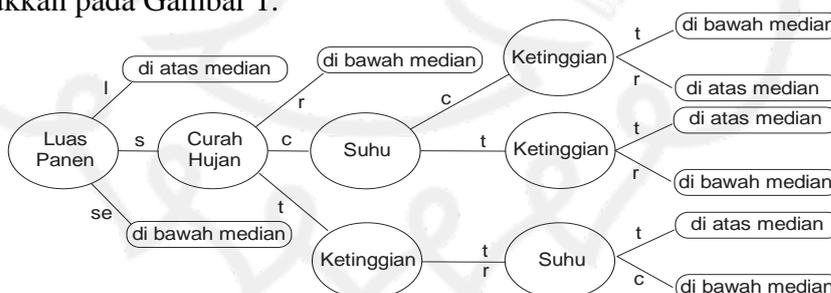
median.

Iterasi	Atribut	Gain	Split Info	Gain Ratio
3	Suhu	0.1461	0.9852	0.1483
	Ketinggian	0	0	0

Perhitungan iterasi ketiga diperoleh nilai *gain ratio* tertinggi yaitu suhu sebesar 0.1483. Suhu kategori cukup perlu dilakukan iterasi sehingga diperoleh atribut ketinggian sebagai *node* terakhir. Ketinggian kategori rendah dan tinggi sudah mengklasifikasikan kasus ke dalam produksi di atas dan di bawah median. Suhu kategori tinggi perlu dilakukan iterasi sehingga diperoleh atribut ketinggian sebagai *node* terakhir. Ketinggian kategori rendah dan tinggi sudah mengklasifikasikan kasus ke dalam produksi di bawah dan di atas median.

Iterasi	Atribut	Gain	Split Info	Gain Ratio
4	Suhu	0.0044	0.4395	0.0099
	Ketinggian	0.0358	0.5329	0.0673

Perhitungan iterasi keempat diperoleh nilai *gain ratio* tertinggi yaitu ketinggian sebesar 0.0673. Ketinggian kategori tinggi dan rendah perlu dilakukan iterasi sehingga diperoleh atribut suhu sebagai *node* terakhir. Suhu kategori cukup dan tinggi sudah mengklasifikasikan kasus ke dalam produksi di bawah dan di atas median. Pada iterasi ketiga dan keempat menggunakan atribut yang sama karena berada pada kategori curah hujan yang berbeda. Hasil pohon keputusan ditunjukkan pada Gambar 1.



Gambar 1. Pohon Keputusan Produksi Padi

Dari Gambar 1, diperoleh klasifikasi status kota/kabupaten di Pulau Jawa berdasarkan atribut luas panen yang ditunjukkan pada Tabel 4.

Tabel 4. Klasifikasi Status Kota/Kabupaten

Klasifikasi	Status Kota/Kabupaten
1	Luas Panen Luas dan Produksi Padi Tinggi
2	Luas Panen Sedang dan Produksi Padi Tinggi
3	Luas Panen Sedang dan Produksi Padi Rendah
4	Luas Panen Sempit dan Produksi Padi Rendah

Klasifikasi Data Produksi Padi ...

Selain itu, pohon keputusan yang terbentuk dapat diinterpretasikan berupa aturan klasifikasi yang ditunjukkan pada Tabel 5.

Tabel 5. Aturan Klasifikasi

No.	Aturan Klasifikasi
1	Jika luas panen sempit maka produksi padi di bawah median.
2	Jika luas panen sedang, dan curah hujan rendah maka produksi padi di bawah median.
3	Jika luas panen sedang, curah hujan cukup, suhu cukup, dan ketinggian rendah maka produksi padi di atas median.
4	Jika luas panen sedang, curah hujan cukup, suhu cukup, dan ketinggian tinggi maka produksi padi di bawah median.
5	Jika luas panen sedang, curah hujan cukup, suhu tinggi, dan ketinggian rendah maka produksi padi di bawah median.
6	Jika luas panen sedang, curah hujan cukup, suhu tinggi, dan ketinggian tinggi maka produksi padi di atas median.
7	Jika luas panen sedang, curah hujan tinggi, ketinggian tinggi atau rendah, suhu cukup maka produksi padi di bawah median.
8	Jika luas panen sedang, curah hujan tinggi, ketinggian tinggi atau rendah, suhu tinggi maka produksi padi di atas median.
9	Jika luas panen luas maka produksi padi di atas median.

Klasifikasi status kota/kabupaten digunakan untuk menggambarkan peta wilayah Pulau Jawa. Gambar klasifikasi produksi padi tiap kota/kabupaten di Pulau Jawa ditunjukkan pada Gambar 2.



Gambar 2. Klasifikasi produksi padi tiap kota/kabupaten di Pulau Jawa

Berdasarkan Gambar 2, diketahui bahwa kota/kabupaten yang produksi padinya perlu dioptimalkan yaitu kota/kabupaten yang masuk dalam Klasifikasi 2 karena memiliki luas panen sedang dan menghasilkan produksi padi tinggi.

7.2. Pengujian Akurasi. Menurut Chauhan dan Anu [3], pengujian akurasi pada

pohon keputusan menggunakan tabel *confusion matrix* ditunjukkan pada Tabel 6.

Tabel 6. *Confusion Matrix*

<i>Correct Classification</i>	<i>Classified Matrix</i>		Persentase
	Di atas median	Di bawah median	
Di atas median	40	16	71.43%
Di bawah median	3	54	94.74%

Berdasarkan Tabel 6, diketahui bahwa dari jumlah data sebanyak 113 terdapat 94 data dengan klasifikasi benar. Perhitungan nilai akurasi menggunakan persamaan (2.5) dan diperoleh persentase akurasi sebesar 83.19%. Persentase untuk mengklasifikasi produksi padi di atas median secara tepat yaitu sebesar 71.43% dan 94.74% untuk mengklasifikasi produksi padi di bawah median secara tepat.

8. KESIMPULAN

Berdasarkan hasil pembahasan diperoleh kesimpulan bahwa pohon keputusan dengan Algoritme *C4.5* digunakan mengklasifikasi data produksi padi di kota/kabupaten Pulau Jawa sehingga diperoleh kota/kabupaten yang termasuk dalam klasifikasi 2 merupakan kota/kabupaten yang produksi padinya perlu dioptimalkan.

DAFTAR PUSTAKA

- [1] Badan Pusat Statistik, *Statistik Indonesia Tahun 2016*, Jakarta Pusat: Badan Pusat Statistik, 2016.
- [2] Bramer, M., *Principles of Data Mining*, London: Springer, 2007.
- [3] Chauhan, H. and A. Chauhan, *Implementation of Decision Tree Algorithm C4.5*, International Journal of Scientific and Research Publications, 2013.
- [4] Florin, G., *Data Mining Concepts Models and Technique*, Berlin: Springer, 2011.
- [5] Han, J. and M. Kamber, *Data Mining Concept and Tehniques*, San Fransisco: Morgan Kauffman, 2006.
- [6] Hanum, C., *Teknik Budidaya Tanaman Jilid 2*, Departemen Pendidikan Nasional, Jakarta, 2008.
- [7] Ina, H., *Bercocok Tanam Padi*, Jakarta: Azka Mulia Media, 2007.
- [8] Larose, D. T., *Discovering Knowledge in Data*, New Jersey: John Willey and Sons, 2005.
- [9] Las, L., *Inovasi Teknologi untuk Peningkatan Produksi Padi dan Kesejahteraan Petani*, Balai Penelitian Tanaman Padi, Sukamandi, 2004.
- [10] Quinlan, J. R., *Induction of Decision Trees*, Machine Learning 1, 1986, 81-106.
- [11] Wei, D. and Wei Ji, *A Map Reduce Implementation of C4.5 Decision Tree Algorithm*, International Journal of Database Theory and Application, 2014.
- [12] Witten, I. H., Frank, E., and Hall, M. A., *Data Mining Practical Machine Learning Tool and Techniques (3rd edition)*, USA: Elsevier, 2011.