SIGMA
JURNAL TEKNOLOGI PELITA BANGSA

Vol. 13 No. 3 September 2022 p-ISSN: 2407-3903 e-ISSN: 2407-3903

Diterima, 13 September 2022 | Direvisi, 22 September 2022 | Dipublikasikan, 26 September 2022

DATA MINING IMPLEMENTATION ON JAVA NORTH COAST WEATHER FORECAST DATASET USING C4.5 ALGORITHM

Dendy K. Pramudito[1]
Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
[1]doktor.haji.dendy@pelitabangsa.ac.id

Abstrak

Cuaca merupakan salah satu hal yang paling berpengaruh dalam kehidupan manusia sehari-hari. Banyak aktivitas yang dilakukan manusia tidak lepas dari kondisi cuaca yang berlaku. Akhir-akhir ini sering terjadi penyimpangan pola cuaca yang tidak biasa atau bisa dikatakan ekstrim. Oleh karena itu, mengamati cuaca sangat diperlukan untuk membuat prediksi tentang cuaca. Garis pantai utara merupakan salah satu jalur penting di Pulau Jawa, khususnya jalur pantai utara di Jawa Tengah, oleh karena itu informasi prakiraan cuaca pada jalur ini sangat dibutuhkan. Tujuan dari penelitian ini adalah untuk mendapatkan faktor-faktor yang paling mempengaruhi perubahan cuaca. Pendekatan data mining yang digunakan dalam penelitian ini adalah metode pohon keputusan dan algoritma C4.5. Dari hasil pengujian 2.400 data prakiraan cuaca diambil dari situs accuweather dan dibagi menjadi 2 yaitu data latih sebanyak 1.680 data, selebihnya data pengujian sebanyak 720 data, hasil yang didapat dari pohon keputusan dengan root node adalah atribut kelembaban dengan tingkat akurasi 81,94% yang telah dibuktikan melalui alat rapid miner 9.10.

Kata kunci: Prakiraan Cuaca, Data Mining, Algoritma C4.5, Pohon Keputusan

Abstract

Weather is one of the most influential things in everyday human life. Many activities carried out by humans cannot be separated from the prevailing weather conditions. Lately, there have been frequent irregularities in weather patterns that are not usual or can be said to be extreme. Therefore, observing the weather is very necessary to make predictions about the weather. The northern coastline is one of the most important routes in Java, especially the northern coast route in the Central Java, due to that information about weather forecasts on this route is needed. The purpose of this research was to obtain the most influence factors of weather changes. The data mining approach used in this research is decision tree method and C4.5 algorithm. From the test results of 2,400 weather forecast data taken from the accuweather site and divided into 2, namely training data as much as 1,680 data, the rest of testing data as much as 720 data, the results obtained from a decision tree with the root node is the humidity attribute with an accuracy rate of 81.94% which has been proven through rapid miner 9.10 tools.

Keywords: Weather Forecast, Data Mining, C4.5 Algorithm, Decision Tree

1. Pendahuluan

Cuaca merupakan kondisi fisik udara sesaat pada suatu area yang sempit dan suatu waktu tertentu. Secara sederhana, cuaca dapat dimaknai sebagai apa yang terjadi saat ini dan dapat berubah-ubah dari waktu ke waktu [1]. Pada umumnya cuaca dipengaruhi pada beberapa faktor yaitu suhu, kelembaban, kecepatan angin, dan curah hujan. 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 [2].

Hal ini menandakan bahwa perlunya dari berbagai pihak yang membutuhkan informasi kondisi cuaca yang lebih akurat, cepat, dan lengkap. Peran prakiraan cuaca di Indonesia cukup penting, sebab wilayah Indonesia memiliki karakteristik yang berbeda-beda antar daerah, menyebabkan terjadinya ketidakseragamannya antara cuaca di daerah yang

dengan lainnya. Dalam hal ini kelancaran pekerjaan maupun kegiatan lainnya juga dapat saja dipengaruhi oleh cuaca.

Dalam hal transportasi, prediksi cuaca juga sangat penting, khususnya di wilayah pesisir, sehingga ketidaktepatan prediksi cuaca dapat menyebabkan ketidaklancaran jalur transportasi khususnya di jalur pantura. Jalur pantura atau yang sering disebut dengan jalur pantai utara merupakan jalan nasional rute 1 atau jalan utama yang ada di Pulau Jawa. Jalan ini melewati 5 provinsi sepanjang 1.316 km di sepanjang pesisir pantai utara Jawa, yaitu Banten, Jakarta, Jawa Barat, Jawa Tengah dan Jawa Timur.

Jalur ini menjadi urat nadi utama transportasi darat karena setiap harinya dilalui oleh 20.000-70.000 kendaraan [3]. Pembangunan infrastruktur di sepanjang pantura masih akan tetap memperoleh prioritas tertinggi dari pemerintah dikarenakan jalur ini berperan paling strategis dalam perekonomian Indonesia.

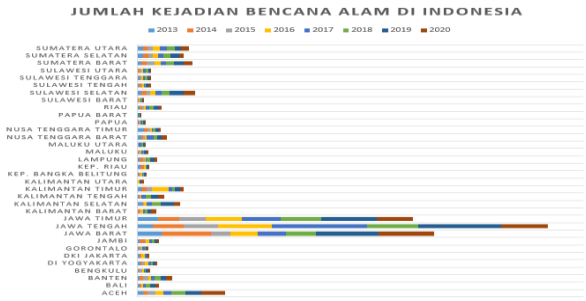
Belakangan ini sering terjadi penyimpangan pola-pola cuaca yang tidak biasanya, atau dapat dikatakan ekstrem, dengan frekuensi yang cenderung bertambah. Cuaca ekstrem yang biasa terjadi, seperti: angin kencang, suhu udara yang melewati ambang batas normalnya, ditambah dengan periodenya yang singkat kadang disertai dengan angin puting beliung dan curah hujan dengan intensitas tinggi atau disebut hujan ekstrem yang dapat mengakibatkan terjadinya bencana banjir dan longsor [4].

Sebagai negara kepulauan, Indonesia merupakan negara tropis yang sangat rentan terhadap dampak dari cuaca ekstrem. Jika dilihat dari dampak yang ditimbulkan maka kajian cuaca ekstrem perlu dikembangkan di Indonesia. Pengetahuan yang baik perihal cuaca beserta parameter-parameternya, terutama kejadian cuaca ekstrem sangat berguna untuk banyak orang pada berbagai bidang pekerjaan ekonomi agar dapat dimaksimalkan dan dapat meminimalkan kerugian.

Kondisi cuaca diyakini mempengaruhi berbagai aspek yang dapat mempengaruhi keselamatan jalan, yaitu keputusan untuk melakukan perjalanan atau dalam memilih moda transportasi. [3] Umumnya, manusia akan sangat memperhatikan efek negatif hujan dan suhu ekstrim pada permintaan transportasi.

Hal ini terutama berlaku untuk perjalanan yang dibuat untuk tujuan rekreasi, karena dapat dengan mudah dijadwal ulang atau dibatalkan. Oleh karena itu, kebutuhan akan akurasi prediksi cuaca juga diharapkan akan sangat membantu manusia dalam aktivitas keseharian yang membutuhkan informasi akan keakuratan prakiraan cuaca, untuk menghindari terjadinya hal-hal yang tidak diinginkan saat berkendara.

Berikut adalah grafik jumlah kejadian bencana alam karena cuaca yang terjadi di Indonesia.



Gambar 1. Jumlah Bencana Alam Akibat Cuaca di Indonesia

Seiring perkembangan kemajuan ilmu pengetahuan dan teknologi, dengan melihat adanya penyimpangan pola-pola cuaca yang tidak biasanya, atau dapat dikatakan ekstrem, dengan frekuensi yang cenderung bertambah, sehingga informasi prakiraan cuaca sangat diperlukan. Peramalan cuaca adalah aplikasi yang paling penting dalam meteorologi dan telah menjadi salah satu yang paling ilmiah dan menjadi pemasalahan teknologi yang semakin berkembang. 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 [6].

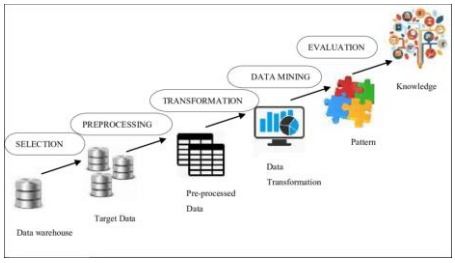
Algoritma C4.5 adalah salah satu algoritma yang digunakan untuk membentuk pohon keputusan. Peramalan cuaca merupakan suatu proses memprediksikan bagaimana perubahan kondisi atmosfir di waktu yang akan datang. Untuk memprediksi suatu cuaca digunakan algoritma decision tree untuk mengklasifikasikan parameter cuaca seperti suhu, kelembaban, arah angin, dan kecepatan angin. Hasil yang di dapat adalah parameter yang memiliki pengaruh yang berarti [2]. Dilihat dari sudut pandang machine learning, C4.5 tergolong sebagai algoritma supervised learning, yaitu teknik yang digunakan untuk mengenali pola dari dataset menggunakan sebuah atribut sebagai pengenal terhadap karakteristik data, dimana atribut tersebut pada umumnya diberi nama label atau kelas [7]. Karakteristik utama dari *supervised learning* selain memiliki atribut label/kelas adalah wajib memiliki dua jenis data, yaitu *data training* dan *data testing*.

2. Metode Penelitian

2.1. *Data Mining*

Merupakan suatu proses untuk menemukan pola yang menarik dan pengetahuan dari data dalam jumlah besar [12]. Dari definisi tersebut maka dapat disimpulkan bahwa *data mining* merupakan proses ekstraksi informasi dari basis data yang berukuran besar untuk mendapatkan pengetahuan yang tersimpan dari data tersebut. Pada beberapa tahun belakangan ini, kemajuan dari beberapa bidang ilmu pengetahuan seperti sains, bisnis dan lain-lain telah melahirkan koleksi basis daya yang terus meningkat. *Data mining* sering dianggap sebagai bagian dari *knowledge*

discovery in database (KDD) yaitu sebuah proses mencari pengetahuan yang bermanfaat dari data [13].



Gambar 2. Tahapan *Data Mining*

Data mining menganalisis data menggunakan *tools* untuk menemukan pola dan aturan dalam himpunan data yang diharapkan mampu mengenal pola ini dalam data dengan input minimal dari *user*. Menurut [14], *data mining* memiliki 5 (lima) fungsi untuk menemukan pola yaitu (1) fungsi deskripsi (*Description*), (2) fungsi estimasi (*estimation*), (3) fungsi prediksi (*prediction*), (4) fungsi klasifikasi (*Classification*), dan (5) fungsi asosiasi (*association*).

2.2. Algoritma C4.5

Merupakan salah satu algoritma yang digunakan untuk membentuk pohon keputusan berdasarkan data latih. Algoritma ini merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. [15] menjelaskan bahwa konsep dasarnya adalah mengubah data menjadi pohon keputusan dan menjadi aturan-aturan keputusan (*decision rules*). Banyak yang telah melakukan pengembangan terhadap algoritma C4.5 sehingga mampu mengatasi *mising value*, *continue data* dan *pruning*. Algoritma C4.5 mempunyai input *training samples* dan *samples* di mana *training samples* berupa data contoh yang digunakan untuk membangun *tree* yang telah diuji kebenaranya, sedangkan untuk *samples* merupakan *field-field* data yang akan digunakan sebagai parameter di dalam melakukan klasifikasi data.

Algoritma C4.5 ini pernah digunakan sebagai penelitian ilmiah yang dilakukan oleh [2] dengan data yang dipakai dalam penelitian ini memiliki interval waktu setiap 3 jam terhitung mulai tanggal 12 Agustus 2016 pukul 01.00 hingga tanggal 20 Agustus 2016 pukul 22.00. Penelitian ini 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%.

Penelitian lain dalam mempredikasi hujan yang berjudul juga dilakukan oleh [8] dengan menggunakan data cuaca harian pertahun dari tahun 2005 sampai dengan tahun 2009 pada salah satu stasiun pemantau di Jakarta yang didapat dari *World Meteorological Organization* (WMO). Tujuan penelitian ini adalah untuk melihat pola prediksi dari setiap atribut yang terdapat pada data cuaca dengan menggunakan algoritma C4.5. Akurasi pola prediksi yang didapat mampu mencapai 79%. Akurasi tersebut dihasilkan dari uji coba dengan mengunakan data cuaca tahun 2007 sebagai *data training* serta data cuaca tahun 2008 dan 2009 sebagai *data testing*.

Lebih lanjut mengenai algoritma C4.5. maka [11] melakukan penelitian membahas informasi pada web pariwisata. Teknik klasifikasi yang akan diterapkan untuk mengklasifikasikan data tweet yang menginformasikan cuaca di kota Yogyakarta. Hasil pengujian dengan perangkat lunak Rapid Miner 5.3 diperoleh nilai akurasi terkecil 71% dengan sampel sebanyak 100 dan nilai akurasi tertinggi 95,58% dengan sampel 15106 dengan algoritma C4.5. Hal tersebut senada dengan [8] yang menjelaskan mengenai peranan klasifikasi di dalam *data mining* dalam mencari pola tertentu, di mana Algoritma C4.5 ini mampu melakukan pencarian pola prediksi dari setiap atribut yang terdapat pada data cuaca.

Sedangkan [10] membahas tentang konsep dasar dari pohon keputusan adalah mengubah data menjadi sebuah model pohon keputusan, kemudian mengubah pohon menjadi aturan dan menyederhanakan aturan yang ada. Data dalam pohon keputusan dinyatakan dalam bentuk tabel dengan atribut dan *record*. Terdapat beberapa cara untuk mengontruksikan pohon keputusan salah satunya menggunakan algoritma C4.5.

2.3. Pohon Keputusan

[16] bahwa pohon keputusan merupakan salah satu metode klasifikasi yang kuat dan terkenal. Metode ini mengubah fakta besar menjadi pohon keputusan yang merepresentasikan aturan, aturan tersebut dapat dengan mudah untuk diinterpretasi oleh manusia. Pohon keputusan juga berguna untuk mengeksplorasi data, dan juga menemukan hubungan tersembunyi antara sejumlah variabel masukkan dengan sebuah variabel target. [13] menjelaskan pada pohon keputusan terdapat 3 (tiga) jenis *node*, yaitu (1) *Root node* yang merupakan *node* paling atas di mana pada *node* ini tidak ada masukkan dan bisa tidak mempunyai keluaran atau mempunyai keluaran namun lebih dari satu, (2) *Internal node*, merupakan *node* percabangan, pada *node* ini hanya terdapat satu masukkan dan mempunyai keluaran minimal dua, (3) *Leaf node* atau *terminal node* merupakan *node* akhir di mana pada *node* ini hanya terdapat satu masukkan dan tidak mempunyai keluaran.

2.4. Prediksi

Suatu peramalan secara sains akan melibatkan pengambilan data historis dan memproyeksikan data historis tersebut ke masa yang akan datang dengan menggunakan model sistematis [17]. Prediksi yang disebut juga dengan peramalan merupakan pernyataan mengenai nilai yang akan datang dari variabel. Prediksi yang baik dapat menjadi keputusan dengan menggunakan banyak informasi [18]. Berdasarkan definisi-definisi tersebut maka dapat disimpulkan

bahwa peramalan adalah sebuah perkiraan di masa yang akan datang dengan melibatkan pengambilan data dari masa lalu pada periode waktu tertentu.

Dalam suatu prediksi jangka pendek biasanya membutuhkan metode yang tidak bervariasi, berbeda halnya dengan prediksi jangka menengah dan panjang. [19] menjelaskan bahwa terdapat beberapa metode yang digunakan dalam prediksi data yaitu (1) *Smoothing* yaitu suatu metode yang biasa digunakan untuk peramalan perencanaan keuangan dan berfungsi untuk meminimalisir data masa lalu yang tidak beraturan atau musiman. (2) *Box jenknis* merupakan suatu metode berisi model matematis yang berfungsi untuk meramalkan data *time series* pada jangka pendek, dan (3) Proyeksi *trend* merupakan suatu metode berisi persamaan matematis yang berupa garis *trend* dan berfungsi untuk prediksi jangka panjang dan pendek.

3. Metode Penelitian

Penelitian ini dilakukan pada tanggal 20 November 2021 - 9 Januari 2022. Dimulai sejak pengumpulan data, eksperimen, penyusunan laporan, hingga penyelesaian laporan penelitian. Prakiraan cuaca sangat dibutuhkan untuk menentukan suatu keputusan pada suatu kegiatan yang berhubungan dengan cuaca [2]. Prakiraan cuaca memungkinkan orang untuk merencanakan dan mengambil tindakan pencegahan terhadap berbagai bencana alam, seperti banjir dan angin topan sehingga dapat meminimalkan dampaknya. Cuaca buruk seperti hujan deras atau angin kencang dapat merusak properti dan menyebabkan kematian. Oleh karena itu dengan adanya prediksi bahwa cuaca buruk akan terjadi, orang dapat mengambil tindakan pencegahan seperti, mengungsi dari daerah yang terkena bencana atau tinggal di dalam rumah. Dalam hal transportasi, prediksi cuaca juga sangat penting, khususnya di wilayah pesisir, sehingga ketidaktepatan prediksi cuaca dapat menyebabkan ketidaklancaran jalur transportasi khususnya di jalur pantura.

Jalur pantura atau yang sering disebut dengan jalur pantai utara merupakan jalan nasional rute 1 atau jalan utama yang ada di Pulau Jawa [3]. Jalur pantura di area Jawa tengah meliputi Kecamatan Losari, Kecamatan Bulakamba, Kecamatan Brebes, Kota Tegal, Kecamatan Suradadi, Kecamatan Pemalang, Kecamatan Petarukan, Kota Pekalongan, Kecamatan Batang, Kecamatan Gringsing, Kecamatan Weleri, Kecamatan Kendal, Kecamatan Kaliwungu, Kota Semarang, Kecamatan Demak, Kecamatan Cangkring, Kecamatan Kudus, Kecamatan Pati, Kecamatan Rembang, dan Kecamatan Lasem. Data pengamatan ini sangat penting untuk melihat karakteristik cuaca setempat dalam pembuatan informasi prakiraan beberapa hari kedepan [20].

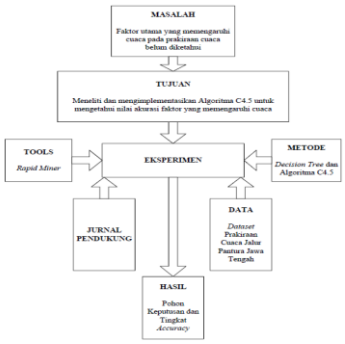
Data yang digunakan dalam penelitian ini adalah data valid prakiraan cuaca wilayah jalur pantura Jawa Tengah dari website accuweather, terhitung mulai dari tanggal 10 Januari 2022 sampai dengan 14 Januari 2022. Data tersebut memiliki 23 jumlah 2.400 dengan 7 atribut dan 1 label.

Tabel 1. Konten dari Prakiraan Cuaca

Konten	Detail Penggunaan
Wilayah	ID
Tanggal	Atribut
Jam	Atribut
Suhu	Atribut
Arah Angin	Atribut
Angin	Atribut
Kelembaban	Atribut
Prakiraan Cuaca	Label

3.1. Kerangka Pemikiran

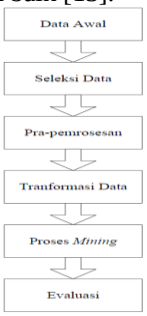
Di dalam penelitian ini menggunakan kerangka pemikiran yang dapay memberikan gambaran tentang masalah dalam penelitian ini yaitu tentang faktor utama yang berpengaruh pada prakiraan cuaca belum diketahui, oleh karena itu tujuan dari penelitian adalah untuk meneliti dan mengimplementasikan algoritma C4.5 dalam mengetahui faktor yang paling berpengaruh pada cuaca. Berdasarkan masalah yang ada, metode yang sesuai untuk dipakai dalam penelitian ini yaitu decision tree dengan algoritma C4.5, dengan *tools* berupa *rapid miner* sebagai aplikasi perangkat lunak untuk pengujiannya.



Gambar 3. Kerangka Pemikiran Penelitian

3.2. Tahapan Penelitian *Data Mining*

Tahapan diawali dari pengumpulan data. Kumpulan data tersebut akan diseleksi dari data sumber ke data target, tahap pre-processing untuk memperbaiki kualitas data, transformasi, data mining serta tahap interpretasi dan evaluasi yang menghasilkan output berupa pengetahuan baru yang diharapkan memberikan kontribusi yang lebih baik [13].



Gambar 4. Tahapan Penelitian *Data Mining*

p-ISSN: 2407-3903 e-ISSN: 2407-3903

Tabel 2. *Dataset* Prakiraan Cuaca

No	Wilayah	Tanggal	Pukul (WIB)	Suhu (C)	Arah Angin	Angin	Kelembaban	Prakiraan Cuaca
1	Losari, Jawa Tengah	10 Jan 2022	00.00	24	B	4	87	Berawan
2	Losari, Jawa Tengah	10 Jan 2022	01.00	24	BL	4	86	Berawan
3	Losari, Jawa Tengah	10 Jan 2022	02.00	24	BL	4	85	Berawan
4	Losari, Jawa Tengah	10 Jan 2022	03.00	24	BL	4	84	Berawan
5	Losari, Jawa Tengah	10 Jan 2022	04.00	24	BL	4	90	Berawan
6	Losari, Jawa Tengah	10 Jan 2022	05.00	24	BL	4	92	Berawan
7	Losari, Jawa Tengah	10 Jan 2022	06.00	24	BL	4	94	Berawan
8	Losari, Jawa Tengah	10 Jan 2022	07.00	25	BD	4	90	Berawan
9	Losari, Jawa Tengah	10 Jan 2022	08.00	27	BD	6	83	Berawan
10	Losari, Jawa Tengah	10 Jan 2022	09.00	28	BD	7	75	Berawan
…	…	…	…	…	…	…	…	…
2400	Losari, Jawa Tengah	14 Jan 2022	23.00	25	B	7	90	Berawan

Tabel 3. Data Hasil Transformasi

No	Wilayah	Suhu	Arah Angin	Angin	Kelembaban	Prakiraan Cuaca
1	Losari, 10 Jan 2022, 00.00 WIB	Sedang	B	Tenang	Sedang	Berawan
2	Losari, 10 Jan 2022, 01.00 WIB	Sedang	BL	Tenang	Sedang	Berawan
3	Losari, 10 Jan 2022, 02.00 WIB	Sedang	BL	Tenang	Sedang	Berawan
4	Losari, 10 Jan 2022, 03.00 WIB	Sedang	BL	Tenang	Sedang	Berawan
5	Losari, 10 Jan 2022, 04.00 WIB	Sedang	BL	Tenang	Sedang	Berawan
6	Losari, 10 Jan 2022, 05.00 WIB	Sedang	BL	Tenang	Tinggi	Berawan
7	Losari, 10 Jan 2022, 06.00 WIB	Sedang	BL	Tenang	Tinggi	Berawan
8	Losari, 10 Jan 2022, 07.00 WIB	Sedang	BD	Tenang	Sedang	Berawan
9	Losari, 10 Jan 2022, 08.00 WIB	Sedang	BD	Sedikit hembusan	Sedang	Berawan
10	Losari, 10 Jan 2022, 09.00 WIB	Sedang	BD	Sedikit hembusan	Rendah	Berawan
…	…	…	…	…	…	…
2400	Lasem, 14 Jan 2022, 23.00 WIB	Sedang	B	Sedikit hembusan	Sedang	Berawan

Data awal yang digunakan pada penelitian ini menggunakan data sekunder karena diperoleh dari dataset prakiraan cuaca. Tujuan utama dari penelitian ini adalah mendapatkan data yang akurat, sehingga tanpa mengetahui teknik pengumpulan data peneliti tidak akan mendapatkan data yang memenuhi standar yang ditetapkan [21]. Tabel 2 menunjukkan serangkaian *dataset* dari prakiraan cuaca yang diperoleh dari data sekunder. Berdasarkan *dataset* tersebut selanjutnya adalah melakukan seleksi data karena tidak semua data penelitian sudah memiliki atribut dan variabel sendiri sebelum dilakukan proses mining, maka seleksi data dilakukan untuk menentukan atribut dan variabel yang dibutuhkan untuk mengolah data [13]. Tabel 4 menunjukan atribut yang akan digunakan dalam penelitian ini. Indikator "o" menandakan atribut yang akan digunakan, sedangkan pada indikator "x" menandakan atribut tersebut akan dieliminasi pada tahap data awal. Pada penelitian ini terdapat pemindahan atribut tanggal dan jam ke dalam atribut wilayah, sehingga atribut wilayah menjadi unique dan dapat digunakan sebagai ID.

Tabel 4. Seleksi Data

No	Atribut	Indikator	Detail Penggunaan
1	Wilayah	O	ID
2	Tanggal	X	Tidak Digunakan
3	Jam	X	Tidak Digunakan
4	Suhu	O	Nilai Model
5	Arah Angin	O	Nilai Model
6	Angin	O	Nilai Model
7	Kelembaban	O	Nilai Model
8	Prakiraan Cuaca	O	Nilai Model

Setelah dilakukan seleksi data maka dilakukan pembersihan data yang berguna untuk membuang data yang redundan maupun mengisi data yang memiliki missing value sehingga data yang akan diolah bersih dan siap digunakan. Pada tahap ini, data prakiraan cuaca dalam proses yang digunakan tidak memiliki *missing value* dan data yang redundan, sehingga jumlah data yang digunakan adalah 2.400 data dengan 4 atribut, 1 ID, dan 1 label. Setelah data sudah dipilih maka dilakukan tahapan untuk melakukan transformasi terhadap atribut, transformasi akan dilakukan untuk memodifikasi sumber data ke format yang berbeda agar dapat diterima oleh proses data mining pada tahap selanjutnya. Transformasi nilai-nilai dari atribut juga perlu dilakukan sehingga dapat mengakibatkan proses pengenalan pola data dan pembentukan keputusan menjadi lama. Data prakiraan cuaca diklasifikasikan menjadi 3 (tiga) bagian atribut yang dilakukan transformasi data, yaitu pada atribut suhu, kelembaban, dan angin, untuk atribut lain tidak ditransformasi karena sudah bisa digunakan untuk proses *mining* sesuai hasil yang ditampilkan pada Tabel 3.

Selanjutnya dilakukan proses *mining* di mana *dataset* yang sudah disiapkan untuk klasifikasi dibagi menjadi 2 (dua) data, untuk data latih sebanyak 70% dan untuk data ujicoba sebanyak 30%. Perhitungan untuk mengambil data latih dan data ujicoba dengan acuan:

- Jumlah data keseluruhan (N) = 2.400
- Jumlah data latih: 70% × 2.400 = 1.680
- Jumlah data ujicoba: 30% × 2.400 = 720

Setelah semua data siap dan sudah sesuai dengan tahapan pengolahan sebelumnya, data yang sudah melalui proses pengolahan kemudian akan dilakukan perhitungan dengan menggunakan *tools rapid miner*.

Dua langkah yang dilakukan pada tahap ini adalah:

1) Perhitungan Algoritma C4.5 secara manual

Data yang akan digunakan dalam perhitungan secara manual yaitu 1680 sampel *data training* yang diambil 70% dari *dataset* prakiraan cuaca. Selanjutnya, hitung nilai *entropy* untuk mendapatkan nilai *gain. Entropy* adalah nilai informasi yang menyatakan ukuran ketidakpastian (*impurity*) dari atribut dari suatu kumpulan objek data dalam satuan bit. Untuk menghitung nilai *entropy* digunakan rumus [6]:

$$Entropy(S)\Sigma - pi \log 2piki = 1$$

S = Himpunan kasus
N = Jumlah partisi *S*
Pi = Proporsi *Si* terhadap *S*

Lalu, menghitung akar dari pohon, akar akan diambil dari atribut yang akan dipilih, dengan cara menghitung nilai *gain* dari masing-masing atribut. Nilai *gain* adalah ukuran efektifitas suatu atribut dalam melakukan klasifikasikan data. Nilai *gain* yang paling tinggi yang akan menjadi akar pertama, menghitung nilai *gain* dapat dengan rumus berikut [6]:

$$Gain\ (S, A) = Entropy\ (S)\ \Sigma|Si|Sni = 1 * Entropy(Si)$$

S = Himpunan kasus
A = Fitur
N = Jumlah partisi atribut *A*
|*Si*| = Proporsi *Si* terhadap *S*
|*S*| = Jumlah kasus dalam *S* [6]

2) Pengujian menggunakan *Rapid Miner*

Aplikasi piranti lunak ini menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. *Rapid miner* merupakan piranti lunak yang berdiri sendiri untuk analisis data dan sebagai mesin *data mining* yang dapat diintegrasikan pada produknya sendiri [22]. Pengujian dengan menggunakan *rapid miner* bertujuan untuk memudahkan dan membandingkan dalam pemrosesan data secara manual.

Langkah terakhir dari tahapan penelitian *data mining* adalah melakukan evaluasi yang dapat dilakukan dengan cara mengamati dan menganalisa hasil dari algoritma yang digunakan untuk memastikan bahwa hasil pengujian itu benar atau tidak sesuai dengan pembahasan [13]. Sedangkan, validasi dilakukan

dengan mengukur hasil prediksi untuk mengetahui tingkat *accuracy*. Maka dari itu, peneliti menggunakan metode *accuracy* sebagai parameter uji untuk mengukur kinerja dari algoritma

4. Hasil Pengujian

Sesuai dengan tahapan evaluasi di mana terdapat 2 (dua) cara perhitungan yaitu manual dan menggunakan piranti lunak *rapid miner*. Adapun langkah-langkah perhitungan nilai *entropy* dan nilai *gain* pada setiap atribut menggunakan data yang dihitung sebagai data latih yaitu sebanyak 1.680 data. Langkah pertama dalam algoritma C4.5 adalah dengan mencari nilai *entropy*. Pertama ditentukan dahulu nilai *entropy* keseluruhan dalam suatu kasus dengan perhitungan:

Diketahui:

Total jumlah kasus (S) = 1.680
Jumlah kasus Berawan (S1) = 1.369
Jumlah kasus Hujan Singkat (S2) = 26
Jumlah kasus Badai Petir (S3) = 285

Maka, perhitungan Entropy total adalah:

= ((-S1/S) * Log2(S1/S)) + ((-S2/S) * Log2(S2/S)) + ((-S3/S) * Log2(S3/S))
= ((-1369/1680) * Log2(1369/1680)) + ((-26/1680) *Log2 (26/1680)) + ((-285/1680) * Log2 (285/1680))
= 0,767925

Langkah kedua yaitu menghitung atribut pada setiap atribut berdasarkan pada jumlah kasus per *subset* atribut untuk mendapatkan nilai *gain.*

4.1. Hasil Pengujian Manual

Mengacu pada tabel 5 maka diketahui bahwa hasil dari perhitungan *entropy* secara manual pada atribut suhu yaitu pada variabel tinggi diketahui S=91, S1=81, S2=0, S3=10, menghasilkan entropy 0,49958. Pada variabel sedang diketahui S=1537, S1=1241, S2=24, S3=272, menghasilkan entropy 0,78502. Terakhir, pada variabel rendah diketahui S=52, S1=47, S2=2, S3=3, menghasilkan entropy 0,55004.

Tabel 5. Hasil *Entropy* dan *Gain* pada Atribut Suhu

Atribut	Variabel	S	S1	S2	S3	*Entropy*	*Gain*
Suhu	Tinggi	91	81	0	10	0,49958	0,0052
	Sedang	1.537	1.241	24	272	0,78502	
	Rendah	52	47	2	3	0,55004	

Berdasarkan tabel 6 maka diketahui bahwa hasil dari perhitungan *entropy* secara manual pada atribut kelembaban yaitu pada variabel tinggi diketahui S=317, S1=267, S2=10, S3=40, menghasilkan *entropy* 0,74271. Pada variabel sedang diketahui S=1.297, S1=1.038, S2=16, S3=243, menghasilkan *entropy* 0,78809. Terakhir, pada variabel rendah diketahui S=66, S1=64, S2=0, S3=2, menghasilkan entropy 0,19590.

Tabel 6. Hasil *Entropy* dan *Gain* pada Atribut Kelembaban

Atribut	Variabel	S	S1	S2	S3	*Entropy*	*Gain*
Suhu	Tinggi	317	267	10	40	0,74271	0,01165
	Sedang	1.297	1.038	16	243	0,78809	
	Rendah	66	64	0	2	0,19590	

Berdasarkan tabel 7 maka diketahui bahwa hasil dari perhitungan *entropy* manual pada atribut arah angin yaitu pada variabel B diketahui S=473, S1=383 S2=4, S3=86, menghasilkan *entropy* 0,75195. Pada variabel BD diketahui S=539, S1=461, S2=16, S3=62, menghasilkan *entropy* 0,70238. Pada variabel BL diketahui S=557, S1=431, S2=4, S3=122, menghasilkan *entropy* 0,81728. Pada variabel S diketahui S=16, S1=13, S2=0, S3=3, menghasilkan *entropy* 0.696212. Pada variabel TG diketahui S=19, S1=18, S2=0, S3=1, menghasilkan *entropy* 0,07389. Pada variabel TL diketahui S=35, S1=31, S2=0, S3=3, menghasilkan *entropy* 0,51270. Pada variabel U diketahui S=40, S1=30, S2=2, S3=8, menghasilkan *entropy* 0,99176. Terakhir, pada variabel *calm* diketahui S=3, S1=2, S2=1, S3=0, menghasilkan *entropy* 0,38997.

Tabel 7. Hasil *Entropy* pada Atribut Kelembaban

Atribut	Variabel	S	S1	S2	S3	*Entropy*	*Gain*
Arah Angin	B	473	383	4	86	0,75195	0,01080
	BD	539	461	16	62	0,70238	
	BL	557	431	4	122	0,81728	
	S	16	13	0	3	0,69621	
	TG	19	18	0	1	0,07389	
	TL	35	31	0	4	0,51270	
	U	40	30	2	8	0,99176	
	Calm	3	2	1	0	0,38997	

Berdasarkan tabel 8 maka diketahui bahwa hasil dari perhitungan *entropy* manual pada atribut angin yaitu pada variabel tenang diketahui S=44, S1=44, S2=0, S3=0, menghasilkan *entropy* 0. Pada variabel sedikit hembusan diketahui S=1.202, S1=972, S2=24, S3=206, menghasilkan *entropy* 0,79663. Pada variabel angin pelan diketahui S=374, S1=305, S2=24, S3=67, menghasilkan *entropy* 0,72472. Pada variabel angin sedang diketahui S=52, S1=42, S2=0, S3=10, menghasilkan *entropy* 0,70627. Terakhir, pada variabel angin sejuk diketahui S=8, S1=6, S2=0, S3=2, menghasilkan *entropy* 0,81127.

Tabel 8. Hasil *Entropy* pada Atribut Angin

Atribut	Var	S	S1	S2	S3	*Entropy*	*Gain*
Angin	Tenang	44	44	0	0	0	0,1089
	Sedikit Hembus	1.202	972	24	206	0,79663	
	Angin Pelan	374	305	2	67	0,72472	
	Angin Sedang	52	42	0	10	0,70627	
	Angin Sejuk	8	6	0	2	0,81127	

4.2. Hasil Pengujian *Rapid Miner*

Pada tahap ini metode *data mining* diterapkan untuk menemukan pengetahuan tersembunyi dan berharga dari data. Metode yang digunakan adalah klasifikasi pohon keputusan dengan algoritma C4.5. Langkah pertama yang dilakukan adalah melakukan impor data ke dalam area proses *rapid miner*. Piranti lunak ini mempunya tampilan GUI (*Graphical User Interface*) sehingga memudahkan pengguna dalam menggunakan perangkat lunak ini dan lisensi berupa bersifat *open*

source dan dibuat dengan menggunakan program *java* di bawah lisensi *GNU public licence* dan *rapid miner* dapat dijalankan di sistem operasi manapun. *Rapid miner* dikhususkan untuk penggunaan *data mining*. Model yang disediakan juga cukup banyak dan lengkap, seperti *model bayesian*, *modeling*, *tree induction*, *neural network,* dan lain-lain. [23].

4.2.1.Pohon Keputusan

Berikut merupakan langkah-langkah dalam membuat pohon keputusan dengan masukkan data latih saja serta pohon keputusan dengan masukkan berupa data latih dan data ujicoba.

Gambar 5. Hasil Pohon Keputusan Berdasarkan Data Latih

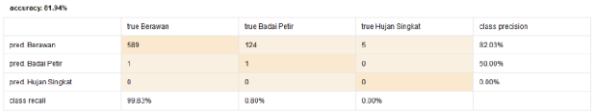
Dapat terlihat bahwa yang menjadi *node root* adalah pada atribut kelembaban, hal tersebut sesuai dengan perhitungan secara manual bahwa *gain* tertinggi pada data latih adalah atribut kelembaban juga yaitu sebesar 0,01165.

Gambar 6. Hasil Pohon Keputusan Berdasarkan Data Latih dan Data Ujicoba

Hasil yang serupa terjadi pada perhitungan pohon keputusan berdasakan data latih dan data ujicoba yang menunjukkan hasil yang sama dengan perhitungan secara manual di mana atribut kelembaban menjadi *node root* yang menunjukan bahwa atribut tersebut merupakan hal yang berpengaruh terhadap perubahan cuaca dari pada atribut-atribut yang lainnya

4.2.2. Nilai Ketepatan

Nilai ketepatan dihitung dengan cara membagi jumlah data benar yang bernilai positif (*true positive*) ditambah dengan jumlah benar yang bernilai negatif (*true negative*) dibagi dengan jumlah data benar yang bernilai positif (*true positive*), ditambah dengan jumlah data yang benar yang bernilai negatif (*true negative*), ditambah data yang salah yang bernilai positif (*false positive*) dan data salah yang bernilai negatif *(false negative*).



Gambar 7. Hasil Tingkat Ketepatan dari Perhitungan di Rapid Miner

5. Pembahasan Hasil Penelitian

Data yang digunakan pada penelitian ini berjumlah 2,400 data yang kemudian dibagi menjadi 2 (dua) bagian yaitu data latih dan data ujicoba. Untuk keperluan data latih maka diambil dari 70% dari keseluruhan *dataset* prakiraan cuaca sehingga data yang digunakan sebagai data latih sebanyak 1.680 data. Sedangkan untuk keperluan data ujicoba maka diambil dari 30% keseluruhan *dataset* prakiraan cuaca sehingga data ujicoba yang digunakan adalah 720 data. Hasil pengujian dari perhitungan *entropy* dan *gain* manual pada data latih menghasilkan kelembaban sebagai atribut yang memiliki nilai *gain* tertinggi yaitu sebesar 0,011654 sehingga kelembaban menjadi atribut utama yang berpengaruh dalam perubahan cuaca sebagai acuan dalam memprediksikan cuaca.

Hasil pengujian dari penelitian yang telah dilakukan dalam pengujian *rapid miner*, dapat diketahui bahwa terdapat atribut yang paling berpengaruh dalam melakukan proses pengklasifikasian *dataset* sehingga menyebabkan terbentuknya pola pengetahuan yang dapat digunakan sebagai acuan dalam memprediksikan cuaca yang akan datang. Atribut tersebut adalah kelembaban yang menjadi *node root* pada data latih dan juga eksekusi data latih dan data ujicoba pada *tools rapid miner*. Maka dari itu dapat disimpulkan bahwa kelembaban adalah atribut yang sangat berpengaruh dalam perubahan cuaca di sepanjang jalur pantura Jawa Tengah dibandingkan dengan atribut lain.

Performa akurasi dari eksekusi data latih dan data ujicoba menggunakan metode klasifikasi pohon keputusan menghasilkan nilai 81.94% dengan presisi prediksi berawan sebesar 82.03%, presisi prediksi badai petir sebesar 50.00%. *Class recall* yang dihasilkan dari performa akurasi ini adalah prediksi benar pada berawan 99.83%, prediksi benar pada badai petir 0.80%. Performa akurasi yang dihasilkan oleh penelitian pada *dataset* prakiraan cuaca ini termasuk pada *good classification* yang dimana nilainya ada pada kisaran 0.80-0.90 [24].

5.1. Implikasi

Penelitian ini menghasilkan informasi mengenai faktor utama yang berpengaruh terhadap perubahan cuaca, yaitu pada atribut kelembaban yang menjadi node *root* sekaligus *gain* tertinggi pada hasil perhitungan algoritma C4.5 dan hasil pengujian pohon keputusan, sehingga dapat disimpulkan bahwa yang berpengaruh besar terhadap perubahan cuaca di sepanjang jalur pantura Jawa Tengah adalah kelembaban. Dalam penelitian prediksi cuaca ini pengujian model dilakukan dengan menggunakan algoritma C4.5. Akurasi dari algoritma klasifikasi C4.5 dengan pengujian melalui *tools rapid miner* 9.10 ini menghasilkan nilai sebesar 81.94% yang termasuk pada kategori *good classification.*

Dengan demikian maka hasil perhitungan prediksi cuaca di pantai utara pulau Jawa menggunakan algoritma C4.5 baik secara manual maupun dengan

bantuan piranti lunak berupa rapid miner dapat memberikan peringatan awal bagi para penggerak bisnis yang berkaitan dengan transportasi baik darat dan laut. Dengan diketahuinya pola cuaca yang terjadi maka para pebisnis dapat melakukan perencanaan yang lebih matang terkait distribusi produk yang dimilikinya. Tidak hanya bagi pemilik bisnis namun hasil penelitian ini membuktikan hasil prediksi cuaca sehingga membantu bagi para pelintas di pulau Jawa untuk memperhatikan perubahan cuaca yang terjadi yang dapat berpengaruh terhadap keselamatan dan kenyamanan perjalanan.

5.2. Keterbasan Penelitian

Penelitian ini dilakukan untuk memperoleh prediksi cuaca berdasarkan *dataset* di pantai pulau Jawa. Karena adanya keterbatasan waktu maka *dataset* yang diperoleh berupa data sekunder dalam rentang waktu yang cukup singkat yaitu tanggal 10 Januari hingga 14 Januari 2022 dengan dataset sebanyak 2,400 dan hanya di wilayah pantai utara pulau Jawa di daerah pesisir Jawa Tengah. Lebih lanjut, penelitian lainnya dapat mengambil *dataset* berupa data primer dengan rentang lebih lama dan di wilayah yang lebih luas lagi atau berpindah ke pantai selatan pulau Jawa disebabkan banyaknya aktifitas warga dan bisnis yang berpengaruh terhadap perubahan cuaca.

Daftar Pustaka

[1] Kaho. (2014). "Panduan Interpretasi dan Respon Informasi Iklim dan Cuaca untuk Petani dan Nelayan," *PIKUL Society*. PIKUL Society.

[2] Novandya, A. (2017) "Penerapan Algoritma Klasifikasi Data Mining C4.5 pada Dataset Cuaca Wilayah Bekasi," KNiST, vol. 6, no. 2, pp. 368–372.

[3] Cahyanto, K. A., Muhamad, F. P. B. , dan Mulyani, E. (2019). "Penerapan Dizcretize By Frequency Dalam Meningkatkan Akurasi Algoritma C4.5 Dalam Memprediksi Cuaca Pada Jalur Pantura Tegal-Pekalongan-Semarang," JTT (Jurnal Teknol. Ter., vol. 5, no. 2, pp. 78. doi: 10.31884/jtt.v5i2.195.

[4] Putra, I. M. D. U., Gandhiadi, I. M. D. U. , dan Harini, L. P. I. (2016). "Implementasi Backpropagation Neural Network Dalam Prakiraan Cuaca Di Daerah Bali Selatan," *E-Jurnal Mat.*, vol. 5, no. 4, p. 126. doi: 10.24843/mtk.2016.v05.i04.p131.

[5] BPS, (2018). "Banyaknya Provinsi Menurut Jenis Bencana Alam," https://www.bps.go.id/indicator/168/954/1/banyaknya-desa-kelurahan-menurut-jenis-bencana-alam-dalam-tiga-tahun-terakhir.html.

[6] Larose, D. T. (2006). Data Mining Methods and Models. Hoboken New Jersey: John Willey & Sons, Inc.

[7] Musu, W., Ibrahim, A. dan Heriadi. (2021). "Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4.5" *Pros. Semin. Ilm. Sist. Inf. Dan Teknol. Inf.*, vol. X, no. 1, pp. 186–195.

[8] Raditya, A. (2014). "Implementasi Data Mining Classification Untuk Mencari Pola Prediksi Hujan Dengan Menggunakan Algoritma C45"

[9] Kerlinger, F.N. (1978). "Similarities and differences in social attitudes in four Western countries," Int. J. Psychol., vol. 13, no. 1, pp. 25–37.

[10] Bimo, P., Setio, N., Retno, D., Saputro, S., dan Winarno, B. (2020). "Klasifikasi dengan Pohon Keputusan Berbasis Algoritme C4.5," Prism. Pros. Semin. Nas. Mat., vol. 3, pp. 64–71.

[11] Y. Astuti, "Klasifikasi Posting Twitter Cuaca Provinsi Diy Menggunakan Algoritma C4.5 Untuk Informasi Pada Web Pariwisata," Pros. Semnastek, no.11, pp. 1–9.

[12] Turban, E., Aronson, J.E., dan. Liang, T.P. (2005). "Decision Support Systems and Intelligent Ssystems", Pearson/Prentice Hall, New Jersey.

[13] Rahmayuni, I. (2014)."Perbandingan Performansi Algoritma C4.5 dan Cart Dalam Klasifikasi Data Nilai Mahasiswa Prodi Teknik Komputer Politeknik Negeri Padang". Teknoif, vol. 2, no. 1, pp. 40–46, 2014.

[14] Jollyta, D., Ramdhan, W., and Zarlis, M. (2020), "Konsep Data Mining Dan Penerapan", Deepublish, Yogyakarta.

[15] Basuki, A., dan Syarif, I. (2003). "Modul Ajar Decision Tree". PENS-ITS, Surabaya.

[16] Berry, M.J., dan Linoff, G.S. (2004). "Data Mining Techniques for Marketing, Sales, and Customer Relationship Management". Wiley Publishing, Inc., 2nd Edition. New Jersey.

[17] Heizer, J. dan Render, B. (2014). "Operations Management: Sustainability and Supply Chain". Global Edition, Pearson Education, Inc. London.

[18] Stevenson, W.J., Choung, D., dan Chee, S. (2014). "Manajemen Operasi", McGraw-Hill Education 9th edition. New York.

[19] Putro, B., Furqon, M.T., dan Wijoyo, S.H. (2018). "Prediksi Jumlah Kebutuhan Pemakaian Air Menggunakan Metode Exponential Smoothing (Studi Kasus : PDAM Kota Malang

),” J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya, vol. 2, no. 11, pp. 4679–4686.

[20] Mujiasih, S. (2011). “Pemanfatan Data Mining Untuk Prakiraan Cuaca,” J. Meteorol. dan Geofis., vol. 12, no. 2, pp. 189–195. doi: 10.31172/jmg.v12i2.100

[21] Sugiyono, (2011). “Metode Penelitian, Kuantitatif, Kualitatif dan R&D”. Alfabeta, CV. Bandung.

[22] Rahma B., et al., (2017). “Implemetasi k-means clustering pada rapidminer untuk analisis daerah rawan kecelakaan,” Semin. Nas. Ris. Kuantitatif Terap. no. April, pp. 58–60.

[23] Haryati, S., Sudarsono, A., dan Suryana, E. (2015). “Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu),” J. Media Infotama, vol. 11, no. 2, pp. 130–138.

[24] Gorunescu, F. (2011). ”Data Mining: Concepts, Models and Techniques”. *Springer*, Berlin.