NEWTON
Jurnal Matematika, Fisika, Algoritma dan Sains

NEWTON : Jurnal Matematika, Fisika, Algoritma dan Sains
Volume 1, No. 1 Desember 2023
ISSN 9999-9999 (media online)
Hal 73-83

Prediksi Cuaca Yang Akan Datang Menggunakan Metode Data Mining

Fitri Anggraeni, Niko Kristiawan, Rikha Lutfiati, Yudha Dirgantara, Perani Rosyani
Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia

e-mail: fitrianggra86@gmail.com, nikstrom31@gmail.com, rikhalutfiati@gmail.com, dirgantara.yudha75@gmail.com, dosen00837@unpam.ac.id

Abstrak-Klasifikasi adalah teknik data mining yang digunakan untuk memprediksi hubungan antar data dalam suatu dataset dengan mengelompokkan data ke dalam beberapa kelas berdasarkan kriteria tertentu. Prediksi dilakukan dengan mengklasifikasikan data ke dalam beberapa kelas berbeda dengan mempertimbangkan faktor-faktor tertentu. Klasifikasi adalah salah satu pendekatan empiris yang dapat digunakan untuk prediksi cuaca jangka pendek. Algoritma klasifikasi yang digunakan dalam penelitian ini adalah Classification Tree yang menggunakan perangkat lunak Orange Data Mining 3.3.12. Selanjutnya, algoritma ini digunakan untuk memprediksi hujan dengan parameter uji Confusion Matrix. Data input berupa data sinoptik dari Stasiun Meteorologi Kemayoran, Jakarta (96745) selama 10 tahun (2006-2015) sebanyak 3528 dataset dan terdiri dari 8 atribut. Berdasarkan serangkaian pemrosesan, pemilihan model, dan pengujian, hasil menunjukkan bahwa akurasi algoritma Classification Tree adalah 74,7% dengan kategori klasifikasi yang baik, di mana jumlah prediksi yang benar adalah 818 dataset dari total data yang diuji sebanyak 1095 dataset. Atribut cuaca yang dominan dalam pembentukan hujan secara berurutan adalah kelembaban (RHavg), suhu minimum (Tmin), suhu maksimum (Tmax), suhu rata-rata (Tavg), dan arah angin (ddd)..

Kata Kunci : Data Mining, Analisis, Diagnosa, Kecerdasan Buatan, Prediksi, Cuaca

Abstract- *Classification is a data mining technique employed to forecast the correlations within a dataset by categorizing data into distinct classes based on specific criteria. This empirical approach is particularly useful for short-term weather predictions. In this study, the Classification Tree algorithm, implemented using Orange Data Mining 3.3.12 software, is utilized to predict rainfall, employing the Confusion Matrix test parameters. The input data consists of synoptic data from the Kemayoran Meteorological Station, Jakarta (96745), spanning a decade (2006-2015) and comprising 3528 datasets with 8 attributes. Following a series of processing, model selection, and testing, the results indicate that the accuracy of the Classification Tree algorithm is 74.7%, falling within the fair classification category. This accuracy is based on 818 correct predictions out of the total tested dataset of 1095. The key weather attributes influencing rainfall prediction include humidity (RHavg), minimum temperature (Tmin), maximum temperature (Tmax), average temperature (Tavg), and wind direction (ddd).*

Keywords : Data Mining, Analysis, Diagnosis, Artificial Intelligence, Prediction, Weather

1. PENDAHULUAN

Cuaca adalah faktor yang sangat memengaruhi aktivitas manusia dan lingkungan sekitar. Keakuratan dalam meramalkan cuaca di masa depan memiliki dampak signifikan dalam perencanaan dan pengambilan keputusan sehari-hari. Untuk meningkatkan ketepatan prediksi cuaca, penelitian-penelitian terkini semakin menggali potensi teknologi data mining.

Data mining, sebagai suatu cabang dari ilmu komputer, menyediakan metode analisis data yang canggih untuk mengidentifikasi pola dan hubungan dalam dataset yang besar dan kompleks. Dengan menerapkan teknik data mining pada data sinoptik, seperti kecepatan angin, suhu, dan tutupan awan, kita dapat merancang model prediksi cuaca yang lebih akurat. Penelitian ini membahas potensi dan relevansi metode data mining dalam meramalkan cuaca, membuka pintu untuk pemahaman yang lebih baik terhadap dinamika atmosfer dan peningkatan signifikan dalam kemampuan prediksi cuaca masa depan. Terdapat dua metode dalam prediksi cuaca (Bhatkande and Hubballi, 2016):

1. Pendekatan Empiris

Pendekatan ini bergantung pada penelitian terhadap data yang lampau untuk memprakirakan keadaan di masa yang akan datang dan mencari hubungan antar atribut. Metode

NEWTON
Jurnal Matematika, Fisika, Algoritma dan Sains
NEWTON : Jurnal Matematika, Fisika, Algoritma dan Sains
Volume 1, No. 1 Desember 2023
ISSN 9999-9999 (media online)
Hal 73-83

yang banyak digunakan dalam pendekatan empiris untuk prediksi cuaca adalah regresi, pohon keputusan (decision tree), artificial neural network, fuzzy logic dan metode pengolahan data yang lain.

2. Pendekatan Dinamis

Pada pendekatan dinamis, diharapkan dapat menghasilkan pendekatan terhadap keadaan sebenarnya dengan pemodelan fisika untuk memprakirakan kondisi di masa yang akan datang.

Prediksi cuaca paling umum dilakukan pada unsur hujan. Hujan adalah fenomena jatuhnya hydrometeor, yaitu partikel-partikel air dengan diameter 0,5 mm atau lebih yang mencapai permukaan tanah (Soepangkat, 1994).

Data mining adalah proses penambangan atau penemuan informasi baru melalui pencarian pola atau aturan khusus dari kumpulan data yang sangat besar (Davies, 2004).

Para peneliti di bidang meteorologi telah menyelidiki cara mendapatkan metode prediksi yang tepat dan akurat menggunakan teknik data mining. Berdasarkan penelitian tersebut, disimpulkan bahwa penerapan teknik data mining untuk prediksi cuaca dengan menganalisis parameter cuaca dapat dilakukan dan menghasilkan nilai akurasi yang baik. Sebagai contoh, algoritma Decision Tree mencapai akurasi yang tinggi, yaitu 88.2%, ketika diterapkan pada data cuaca karena mampu mengklasifikasikan dengan baik (E. Manjula, 2016).

Penelitian yang berjudul *Use of Data Mining Techniques for Weather Data in Basra City* menerapkan teknik K Means Clustering dan Artificial Neural Network (ANN) untuk menentukan nilai akurasi dan Root Mean Square Error (RMSE) sebagai metode klasifikasi dalam prediksi cuaca (hujan, cerah, berawan). Parameter cuaca yang digunakan melibatkan kelembaban, suhu rata-rata, kecepatan angin, arah angin, waktu kejadian angin maksimum, dan curah hujan dengan periode data selama 9 tahun (2004-2013). Hasil penelitian menunjukkan bahwa metode tersebut cukup efektif untuk prediksi cuaca (Prasad and Nejres, 2015).

Data mining dapat dibagi menjadi beberapa bagian berdasarkan tugas yang dilakukan, dan salah satunya adalah prediksi, sesuai dengan Larose (2006). Prediksi memiliki kesamaan dengan klasifikasi dan estimasi, namun perbedaannya terletak pada fakta bahwa dalam prediksi, nilai hasil diproyeksikan ke masa mendatang. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi juga dapat diterapkan, sesuai keadaan yang tepat, untuk tujuan prediksi.

Klasifikasi dan prediksi merupakan dua bentuk analisis data yang dapat digunakan untuk mengekstrak model dari data yang memuat kelas-kelas tertentu atau untuk memprediksi tren data yang akan muncul (Han dan Kamber, 2006).

Classification and Regression Tree (CART) merupakan salah satu metode atau algoritma dari teknik pohon keputusan. CART adalah suatu metode statistik nonparametrik yang dapat menggambarkan hubungan antara variabel respon (variabel dependen) dengan satu atau lebih variabel predictor (variabel independen). Jika variabel respon berbentuk kontinu, metode yang digunakan disebut metode regresi pohon (regression tree). Sedangkan jika variabel respon memiliki skala kategorik, metode yang digunakan disebut metode klasifikasi pohon (classification tree), (Breiman, 1993).

Pembentukan classification tree melibatkan 3 tahap yang memerlukan sampel pembelajaran (learning sample) L. Tahap pertama adalah pemilihan pemilah, di mana setiap pemilahan hanya bergantung pada nilai yang berasal dari satu variabel independen. Metode pemilahan yang sering digunakan adalah indeks Gini, dengan fungsi sebagai berikut (Tobergte and Curtis, 2013):

$$i(t) = \sum_{i \neq j} p(i|t)p(j|t) \dots (1)$$

NEWTON
Jurnal Matematika, Fisika, Algoritma dan Sains
NEWTON : Jurnal Matematika, Fisika, Algoritma dan Sains
Volume 1, No. 1 Desember 2023
ISSN 9999-9999 (media online)
Hal 73-83

Dengan $i(t)$ adalah fungsi indeks keheterogenan Gini, $p(i|t)$ adalah proporsi kelas i pada simpul t, dan $p(j|t)$ adalah proporsi kelas j pada simpul t. Goodness of Split merupakan evaluasi pemilahan oleh pemilah s pada simpul t. Goodness of split $\emptyset(s,t)$ didefinisikan sebagai penurunan keheterogenan.

$$\emptyset(s,t) = \Delta i(s,t) = -p_l(t_l) - p_R(t_R) \dots (2)$$

Pengembangan pohon dilakukan dengan mencari semua kemungkinan pemilah pada simpul t_1, sehingga ditemukan pemilah s^* yang memberikan nilai penurunan keheterogenan tertinggi, yaitu,

$$\Delta i(s^*, t_1) = max_{s \in S}(s, t_1) \dots (3)$$

Dengan $\emptyset(s,t)$ sebagai kriteria goodness of split, $p_L i\ (t_L)$ adalah proporsi pengamatan dari simpul t menuju simpul kanan. Tahap ketiga melibatkan penandaan label untuk setiap simpul terminal berdasarkan aturan jumlah anggota kelas terbanyak, seperti berikut:

$$p(j_0|t) = max_j\, p(j|t) = max_j \frac{N_J(t)}{N(t)}$$

Pertanyaan pokok dalam penelitian ini mencakup bagaimana penerapan teknik data mining untuk meramalkan hujan, seberapa akurat hasil prediksi yang dapat diukur menggunakan parameter Confusion Matrix, dan bagaimana visualisasi hasil prediksi hujan dapat dilakukan melalui algoritma Classification Tree. Metode umum yang digunakan untuk mengevaluasi kinerja pengklasifikasi adalah Cross Validation, suatu bentuk sederhana dari teknik statistik. Dalam penelitian ini, penggunaan 10-fold cross-validation menjadi standar untuk memprediksi tingkat kesalahan dari data (Witten, Frank, dan Hall, 2011).

Atribut kualitas prediksi untuk kategori probabilistik, seperti yang diuraikan oleh Murphy dan Winkler (1992), mencakup sharpness, resolution, discriminant, bias, reliability (kalibrasi), akurasi, dan keterampilan. Dalam penelitian ini, atribut yang akan dibahas fokus pada akurasi. Harapannya, hasil penelitian ini dapat memberikan manfaat dengan menyumbangkan pengetahuan terkait aplikasi pendekatan empiris dalam teknik data mining untuk prediksi hujan. Selain itu, diharapkan penelitian ini dapat meningkatkan pemahaman mengenai teknik prediksi hujan berbasis probabilitas.

2. METODE PENELITIAN

Penelitian ini merupakan jenis penelitian kuantitatif. Pendekatan yang digunakan dalam penelitian ini adalah teknik data mining untuk mengolah seri data pengamatan cuaca dan menganalisis hasilnya guna menentukan algoritma terbaik untuk prediksi cuaca.

Sumber data yang digunakan dalam penelitian ini adalah data observasi meteorologi permukaan (synoptic) yang dikumpulkan selama 10 tahun (2006-2015). Data ini diperoleh dari Sub Bidang Database BMKG dan Stasiun Meteorologi Jakarta (96745). Observasi meteorologi permukaan dilakukan setiap jam, di mana data synop yang awalnya berupa kode kemudian diartikan dan dimasukkan ke dalam Ms Excel untuk analisis lebih lanjut. Berbagai unsur cuaca yang diamati dalam penelitian ini meliputi temperatur, tekanan atmosfer, jarak pandang (visibility), keadaan cuaca, arah dan kecepatan angin, titik embun, jenis awan, jumlah awan, radiasi matahari, lamanya penyinaran matahari, dan elemen-elemen cuaca lainnya.

Pada penelitian ini, pengolahan data menggunakan perangkat lunak Orange Data Mining versi 3.3.12 yang memiliki basis open source. Tahap preprocessing mengadopsi pendekatan Knowledge Data Discovery (KDD), yang melibatkan proses pengumpulan dan pemanfaatan data historis untuk mengungkapkan pola, keteraturan, serta hubungan yang terdapat dalam dataset

NEWTON
Jurnal Matematika, Fisika, Algoritma dan Sains
NEWTON : Jurnal Matematika, Fisika, Algoritma dan Sains
Volume 1, No. 1 Desember 2023
ISSN 9999-9999 (media online)
Hal 73-83

berukuran besar (Pramudiono, 2007). Knowledge Database Discovery (KDD) sendiri merupakan suatu proses identifikasi informasi berharga dan pola tersembunyi dalam basis data yang besar, sebelumnya tidak diketahui, dan memiliki potensi manfaat (Han dan Kamber, 2006). Tahapan dalam proses Knowledge Database Discovery (KDD) dapat diuraikan sebagai berikut (Fayyad, Piatetsky-Shapiro, dan Smyth, 1996).

1. **Data Collection**
   Langkah pertama dalam proses Knowledge Database Discovery (KDD) adalah pengumpulan data yang akan menjadi subjek teknik pengolahan data mining. Pemilihan atau seleksi data dari kumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data yang telah melalui tahap seleksi, dan akan digunakan untuk proses data mining, disimpan dalam suatu berkas terpisah, yang terpisah dari basis data operasional.
2. **Data Cleaning**
   Sebelum memulai proses data mining, langkah awal yang diperlukan adalah membersihkan data yang menjadi fokus Knowledge Database Discovery (KDD). Proses pembersihan (cleaning) melibatkan tindakan seperti menghapus data duplikat, memeriksa konsistensi data, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Selain itu, tahap cleaning juga melibatkan proses enrichment, di mana data yang sudah ada "diperkaya" dengan informasi tambahan yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.
   Proses cleaning memiliki fungsi utama, yakni menghilangkan duplikasi data, memeriksa data yang tidak konsisten, dan memperbaiki kesalahan tipografi pada data. Dengan melakukan langkah-langkah ini, data menjadi siap untuk melanjutkan proses selanjutnya, yaitu pemilahan data.
3. **Data Selection**
   Pada tahap ini, dilakukan pemilihan data yang akan digunakan untuk analisis dari dataset. Data meteorologi yang relevan untuk penelitian ini melibatkan temperatur rata-rata harian, temperatur minimum harian, temperatur maksimum harian, kelembaban, kecepatan angin, arah angin, dan lama penyinaran matahari. Atribut-atribut cuaca ini, yang disebut sebagai variabel, dipilih untuk dimasukkan ke dalam dataset baru menggunakan Microsoft Excel dengan format xls. Dataset tersebut mencakup 3528 entri data dengan 8 variabel, terdiri dari 7 variabel numerik dan 1 variabel teksual. Hasil pengolahan pada tahap ini menunjukkan bahwa data tersebut memiliki 9 atribut (lihat Tabel 1) dan tidak terdapat nilai yang hilang (missing value).

Atribut	Tipe	Keterangan
Year	Numerical	Year Considered
Month	Numerical	Month Considered
Temperatur rata – rata	Numerical	Temperatur rata – rata harian
Temperature Minimum	Numerical	Temperatur Minimum Harian
Temperatur Maksimum	Numerical	Temperature Maksimum Harian
Kelembaban	Numerical	Kelembaban rata – rata dalam satu hari

NEWTON
Jurnal Matematika, Fisika, Algoritma dan Sains
NEWTON : Jurnal Matematika, Fisika, Algoritma dan Sains
Volume 1, No. 1 Desember 2023
ISSN 9999-9999 (media online)
Hal 73-83

Kecepatan Angin	Numerical	Kecepatan Angin rata – rata dalam satu Hari
Arah Angin	Numerical	Arah Angin Terbanyak dalam Satu Hari
Lama Penyinaran Matahari	Numerical	Lamanya Penyinaran Matahari dalam Satu Hari
Curah Hujan	Label	Jumlah Curah Hujan Harian

Tabel 1. Atribut Dataset Meteorologi

4. **Data Tranformation**

   Data transformasi, yang juga dikenal sebagai penggabungan data, merupakan tahap di mana data yang telah dipilih diubah menjadi format yang sesuai untuk teknik data mining.

5. **Validasi**

   Tahap validasi dilakukan dengan membagi data menjadi dua bagian, yaitu data training (data latih) sebanyak 70% dari total dataset dan data test (data uji) sebanyak 30% dari total dataset. Data training digunakan untuk menjalankan algoritma Classification Tree, sementara data test digunakan untuk proses validasi. Dalam pendekatan cross-validation, setiap catatan (record) digunakan beberapa kali dalam jumlah yang sama untuk pelatihan dan tepat sekali untuk pengujian. Metode ini membagi data menjadi dua subset yang berukuran sama, memilih salah satu sebagai data training dan yang lainnya sebagai data testing. Kemudian, fungsi antara dua subset tersebut ditukar sehingga yang sebelumnya menjadi set pelatihan menjadi set pengujian, dan sebaliknya. Pendekatan ini disebut two-fold-cross-validation. Total kesalahan diperoleh dengan menjumlahkan kesalahan-kesalahan untuk kedua proses tersebut.

6. **Evaluasi/Pengujian**

   Cross Validation adalah suatu metode umum yang digunakan untuk mengevaluasi kinerja pengklasifikasi. Metode ini merupakan bentuk sederhana dari teknik statistik. Jumlah lipatan (fold) standar yang umum digunakan untuk memprediksi tingkat kesalahan dari data adalah dengan menggunakan 10-fold cross validation, sesuai dengan Witten, Frank, dan Hall (2011).

   Dalam metode cross-validation, setiap catatan digunakan beberapa kali dengan jumlah yang sama untuk keperluan pelatihan, dan hanya sekali untuk pengujian. Pendekatan ini melibatkan pembagian data ke dalam dua subset yang memiliki ukuran yang sama. Salah satu subset dipilih sebagai data pelatihan, sementara subset lainnya digunakan untuk pengujian. Setelah itu, dilakukan pertukaran fungsi antara kedua subset, di mana subset yang sebelumnya berfungsi sebagai set pelatihan menjadi set pengujian, dan sebaliknya. Pendekatan ini dikenal sebagai two-fold cross-validation. Total kesalahan diperoleh dengan menjumlahkan kesalahan dari kedua proses tersebut.

NEWTON
Jurnal Matematika, Fisika, Algoritma dan Sains
NEWTON : Jurnal Matematika, Fisika, Algoritma dan Sains
Volume 1, No. 1 Desember 2023
ISSN 9999-9999 (media online)
Hal 73-83

	Model 1	Model 2	Model 3	Model 4
Fold 1	Test data	Training data	Training data	Training data
Fold 2	Training data	Test data	Training data	Training data
Fold 3	Training data	Training data	Test data	Training data
Fold 4	Training data	Training data	Training data	Test data

Gambar 1. *Fold Cross Validation*

Dalam Gambar 1, diberikan contoh perhitungan dengan menggunakan nilai fold sebanyak 4 pada metode cross-validation. Berikut ini, disajikan langkah-langkah pengujian data dengan 4-fold cross-validation.

Dalam mengevaluasi perbandingan algoritma, parameter yang digunakan melibatkan Confusion Matrix, yang mencakup akurasi, presisi, dan recall. Evaluasi menggunakan confusion matrix memberikan informasi mengenai tingkat akurasi dan tingkat kesalahan. Analisis akurasi, terutama dalam prediksi dikotomi seperti hujan atau tidak, memungkinkan penentuan sejauh mana tingkat ketepatan algoritma klasifikasi yang digunakan. Nilai akurasi dihasilkan dari matriks kontingensi, yang juga dikenal sebagai "error matrix" atau "confusion matrix". Akurasi mencerminkan persentase total data yang diprediksi dengan benar, sementara tingkat kesalahan menggambarkan persentase total data yang diprediksi secara keliru. Laju error merupakan persentase dari total data yang diprediksi secara salah, dihitung dengan rumus:

$$Akurasi = \frac{Jumlah\ data\ yang\ diprediksi\ secara\ benar}{Total\ jumlah\ prediksi\ yang\ dilakukan}$$
$$= \frac{a+d}{a+b+c-d} \times 100\%$$

$$Laju\ Error = \frac{Jumlah\ data\ yang\ diprediksi\ secara\ salah}{Total\ jumlah\ prediksi\ yang\ dilakukan}$$
$$= \frac{b+c}{a+b+c+d} \times 100\%$$

Setelah membuat confusion matrix, langkah selanjutnya adalah menghitung nilai precision, recall, dan accuracy. Untuk mengukur tingkat akurasi, precision, dan recall, biasanya digunakan hasil dari confusion matrix.

Nilai precision, recall, dan accuracy dapat diperoleh melalui perhitungan sebagai berikut:

$$Percision(p) = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall(r) = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

NEWTON
Jurnal Matematika, Fisika, Algoritma dan Sains
NEWTON : Jurnal Matematika, Fisika, Algoritma dan Sains
Volume 1, No. 1 Desember 2023
ISSN 9999-9999 (media online)
Hal 73-83

$$Accuracy = \frac{Jumlah\ klasifikasi\ benar}{Total\ sample\ testing\ yang\ diuji}$$

7. **Interpretasi**

Pola informasi yang dihasilkan dari proses data mining perlu disajikan dalam bentuk yang dapat dengan mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses Knowledge Database Discovery (KDD) yang disebut sebagai interpretasi. Pada tahap ini, dilakukan pemeriksaan apakah pola atau informasi yang ditemukan konsisten dengan fakta atau hipotesis yang ada sebelumnya. Tahap interpretasi bertujuan untuk memahami pola informasi yang dihasilkan dari proses data mining. Informasi yang dihasilkan melalui perangkat lunak Orange akan ditampilkan dan memberikan informasi terkait kinerja masing-masing algoritma dalam metode klasifikasi.

3. HASIL DAN PEMBAHASAN

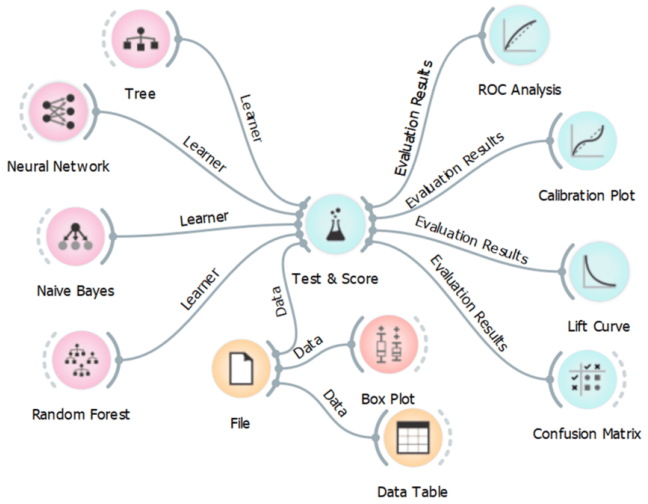
Pada tahap Pengumpulan Data, digunakan data observasi meteorologi permukaan (synoptic) rata-rata harian selama 10 tahun (2006-2015) yang terdiri dari 3528 dataset. Lokasi penelitian berada di Stasiun Meteorologi Kemayoran Jakarta (96745) yang terletak di Jalan Angkasa I No. 2, Kemayoran, Jakarta. Unsur-unsur cuaca yang diamati melibatkan Temperatur, Tekanan, Visibility (Jarak Pandang), keadaan cuaca, arah angin, kecepatan angin, titik embun, jenis awan, jumlah awan, radiasi matahari, lamanya penyinaran matahari, dan lain-lain.

Pada tahap *Data Cleaning*, data meteorologi yang digunakan dalam penelitian ini memiliki 9 atribut (lihat Tabel 1) dan tidak terdapat nilai yang hilang (missing value).

Pada tahap *Data Selection*, dilakukan analisis data dengan memilih parameter cuaca yang relevan untuk penelitian ini, yaitu temperatur rata-rata harian, temperatur minimum harian, temperatur maksimum harian, kelembaban, kecepatan angin, arah angin, dan lama penyinaran matahari. Atribut-atribut cuaca ini, yang selanjutnya disebut sebagai variabel, dipilih untuk dimasukkan ke dalam dataset baru menggunakan Microsoft Excel dengan format xls. Dataset tersebut terdiri dari 3528 entri data dengan 8 variabel, yang terdiri dari 7 variabel numerik dan 1 variabel teksual.

Pada tahap *Data Tranformation*, dilakukan perubahan format dari xls ke csv sesuai dengan tipe data input yang dibutuhkan oleh perangkat lunak Orange.

Pada tahap Validasi, proses perancangan model pada perangkat lunak Orange dapat dijabarkan sebagai berikut:



Gambar 2. Proses Klasifikasi Pada Software Orange Vers. 3.3.1.2

NEWTON
Jurnal Matematika, Fisika, Algoritma dan Sains
NEWTON : Jurnal Matematika, Fisika, Algoritma dan Sains
Volume 1, No. 1 Desember 2023
ISSN 9999-9999 (media online)
Hal 73-83

Data observasi cuaca permukaan (synoptic) diproses menggunakan algoritma klasifikasi, yaitu Classification Tree. Model keluaran dari setiap metode tersebut diuji dengan sebagian data masukan untuk mengevaluasi kehandalan model.

Pada tahap Evaluasi/Pengujian, dilakukan pembagian data menjadi dua bagian, yaitu data latih (training) sebanyak 70%, yang digunakan untuk proses data mining dan perolehan nilai probabilitas. Sementara itu, data uji (test data) sebanyak 30% digunakan untuk menguji nilai probabilitas yang telah terbentuk. Pengujian menggunakan confusion matrix dilakukan untuk mendapatkan nilai precision, recall, dan akurasi dari hasil pengujian. Hasil pengujian ini bertujuan untuk mengukur tingkat akurasi dan Area Under Curve (AUC) dari penentuan dengan metode 10-fold Cross Validation. Berikut adalah hasil pengujian dari masing-masing algoritma:

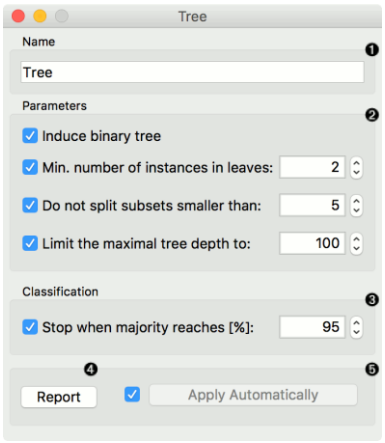
Accuracy : 74,7%		
	True Normal (Hujan)	**True Anomaly (Tidak Hujan)**
Prediksi Normal (Hujan)	286	532
Prediksi Anomaly (Tidak Hujan)	140	137
Proportion of Pedicted	67,1%	79,5%

Tabel 2. Confusion Matrix pada Classification Tree

Dari Tabel 2, nilai-nilai dihitung sebagai berikut:

$$Precision = 0{,}74;\ Recall = 0{,}74;\ Akurasi = 0{,}74;\ dan\ AUC = 0{,}73$$

Berdasarkan hasil tersebut, dapat dilihat bahwa tingkat akurasi menggunakan algoritma Classification Tree mencapai 74,7%. Jumlah prediksi benar adalah sebanyak 818 dataset dari total data yang diuji, yaitu 1095 dataset.



Gambar 3. Pengaturan Classification Tree pada Software Orange

Pada tahap interpretasi dalam algoritma Classification Tree, proses ini menjadi lebih sederhana karena metode ini terkenal sebagai salah satu pendekatan klasifikasi yang mudah dipahami. Dengan menggunakan struktur pohon atau hierarki, Classification Tree menjadi model prediksi yang mempermudah pengguna untuk menginterpretasikan hasilnya.

Model Classification Tree memiliki keunggulan dalam mengubah data menjadi struktur pohon keputusan dan aturan-aturan keputusan. Kemampuan ini mempermudah proses pengambilan keputusan, terutama ketika berurusan dengan atribut yang kompleks. Pengaturan pada widget

NEWTON
Jurnal Matematika, Fisika, Algoritma dan Sains
NEWTON : Jurnal Matematika, Fisika, Algoritma dan Sains
Volume 1, No. 1 Desember 2023
ISSN 9999-9999 (media online)
Hal 73-83

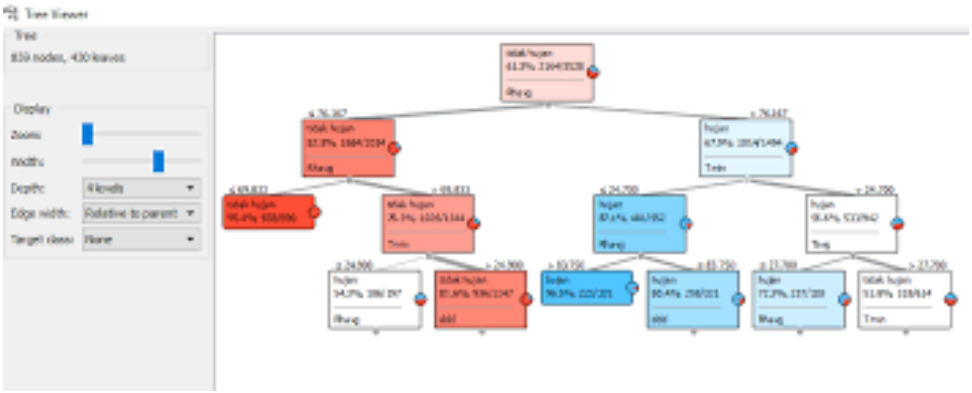
Classification Tree pada Orange Software, seperti yang terlihat pada gambar 3, memberikan kontrol tambahan terhadap proses pembentukan pohon keputusan.

Beberapa parameter pengaturan, seperti Min. Number of Instances in Leaves, memungkinkan pengguna untuk menentukan jumlah minimum instance dalam satu simpul. Fungsi Do Not Split Subsets Smaller Than mencegah algoritma memecah simpul dengan jumlah instance yang kurang dari yang telah ditentukan. Sementara Limit the Maximal Tree Depth membantu membatasi kedalaman pohon keputusan agar tidak melebihi nilai tertentu.

Fitur Stop When Majority Reaches (%) memungkinkan penghentian simpul setelah mencapai threshold tertentu, yang berguna dalam mengontrol kompleksitas model. Dalam proses klasifikasi, masalah dapat diatasi dengan menyajikan sejumlah atribut dari test record, sehingga model dapat memberikan prediksi dengan akurasi yang tinggi.

Dalam setiap langkah pembentukan pohon keputusan menggunakan algoritma Classification Tree, sejumlah atribut diproses untuk menghasilkan node, dan proses ini berulang hingga akhirnya diperoleh sebuah kesimpulan mengenai label kelas dari record yang sedang diproses. Rangkaian proses ini dapat direpresentasikan dalam bentuk pohon keputusan, yang merupakan struktur hirarki terdiri dari node dan simpul.

Pohon keputusan ini mencerminkan keputusan-keputusan yang diambil berdasarkan aturan-aturan yang didefinisikan selama proses pembentukan model. Gambar 4 mengilustrasikan pohon keputusan yang dihasilkan setelah pengolahan menggunakan algoritma Classification Tree. Melalui struktur ini, dapat dengan jelas dilihat bagaimana algoritma membuat keputusan berdasarkan atribut-atribut tertentu untuk mencapai kesimpulan mengenai kelas dari suatu record.

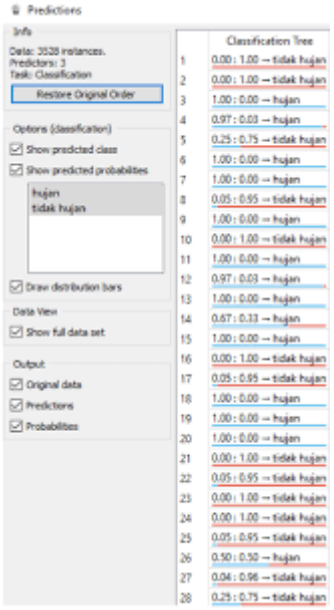


Gambar 4. Classification Tree untuk prediksi hujan

Hasil pengolahan data pada Classification Tree mengungkapkan adanya 859 node dan 430 daun (leaves) dengan kedalaman (depth) sebanyak 5 level. Node akar (root node) pada model ini menggunakan atribut Receiver Humidity rata-rata (RHavg) untuk memisahkan kondisi cuaca menjadi hujan atau tidak hujan. Dari total dataset sebanyak 3528, terdapat 2164 dataset atau 61.3% yang menunjukkan peluang cuaca tidak hujan. Selanjutnya, dari root node, pohon keputusan dibagi menjadi dua internal node, yaitu atribut RHavg dan temperatur minimum (Tmin). Jika nilai RH $\leq$ 76.167%, maka peluang cuaca tidak hujan, sedangkan jika nilai RH > 76.167%, maka peluang cuaca hujan. Dari internal node Tmin, proses pemisahan dilanjutkan dengan membaginya menjadi dua internal node lagi, yaitu RHavg dan temperatur rata-rata (Tavg). Jika Tmin $\leq$ 24.70°C, maka peluang cuaca hujan, dan jika Tmin > 24.70°C, maka peluang cuaca kecil hujan. Dengan struktur hierarki ini, Classification Tree memberikan gambaran yang jelas tentang bagaimana model membuat keputusan berdasarkan kombinasi atribut-atribut tertentu, sehingga memfasilitasi interpretasi dan pemahaman terhadap prediksi cuaca yang dihasilkan.

NEWTON
Jurnal Matematika, Fisika, Algoritma dan Sains
NEWTON : Jurnal Matematika, Fisika, Algoritma dan Sains
Volume 1, No. 1 Desember 2023
ISSN 9999-9999 (media online)
Hal 73-83

Pada langkah terakhir dari proses pembentukan Classification Tree, internal node RH mengarahkan ke leaf node yang memberikan informasi lebih lanjut tentang peluang terjadinya hujan berdasarkan nilai RH. Jika nilai RH > 83.75%, maka peluang terjadinya hujan cukup besar, mencapai 96.5%. Sebaliknya, jika nilai RH ≤ 69.83%, maka peluang tidak hujan cukup besar, yakni sebesar 95.4%. Dengan demikian, informasi ini memberikan pemahaman yang lebih rinci tentang hubungan antara kelembapan relatif (RH) dan peluang terjadinya hujan atau tidak hujan, yang merupakan hasil interpretasi dari pohon keputusan Classification Tree. Setiap leaf node pada pohon tersebut memberikan prediksi yang spesifik terkait dengan kondisi cuaca berdasarkan kombinasi nilai atribut yang diobservasi.



Gambar 5. Prosentase prediksi hujan pada tiap dataset

Dari Gambar 5, terlihat bahwa Classification Tree mampu memberikan informasi tentang prosentase prediksi hujan pada setiap dataset, dengan total 3528 dataset. Model ini berhasil mengklasifikasikan parameter-parameter yang paling berpengaruh terhadap prediksi curah hujan, dan urutannya berdasarkan tingkat pengaruhnya secara berturut-turut adalah RHavg, Tmin, RH, ddd (arah angin), LPM (lamanya penyinaran matahari), dan Tmax.

Selain itu, pohon klasifikasi yang terbentuk juga memberikan informasi tentang peluang intensitas hujan yang akan terjadi. Hal ini dapat diobservasi dari warna yang lebih kuat pada masing-masing node. Warnanya mencerminkan tingkat keyakinan atau kepastian model terkait dengan prediksi hujan pada setiap cabang pohon. Dengan demikian, visualisasi tersebut memberikan pemahaman yang lebih mendalam tentang faktor-faktor yang memengaruhi prediksi curah hujan dan sejauh mana tingkat keyakinan model terhadap hasil prediksi tersebut.

4. KESIMPULAN

Berdasarkan penelitian tentang prediksi cuaca jangka pendek menggunakan algoritma Classification Tree dengan parameter uji Confusion Matrix, dapat ditarik beberapa kesimpulan:

1. **Kualitas Klasifikasi yang Baik (Fair Classification)**
   Hasil analisis parameter uji Confusion Matrix menunjukkan bahwa algoritma Classification Tree dapat diaplikasikan untuk prediksi hujan dengan kategori yang cukup baik, mencapai tingkat klasifikasi yang adil (fair classification).
2. **Interpretasi Pengetahuan yang Signifikan**

NEWTON
Jurnal Matematika, Fisika, Algoritma dan Sains

NEWTON : Jurnal Matematika, Fisika, Algoritma dan Sains
Volume 1, No. 1 Desember 2023
ISSN 9999-9999 (media online)
Hal 73-83

Algoritma Classification Tree berhasil menghasilkan model yang dapat mengklasifikasikan data uji sebanyak 1095 dataset menjadi 287 nodes dan 144 leaves. Hal ini menunjukkan bahwa model dapat memahami pola-pola kompleks dalam data cuaca untuk membuat prediksi yang akurat.

3. **Parameter Cuaca yang Signifikan**
   Interpretasi pengetahuan dari model menunjukkan bahwa parameter cuaca yang paling signifikan terhadap pembentukan hujan adalah kelembaban relatif (RHavg), temperatur minimum (Tmin), temperatur maksimum (Tmax), temperatur rata-rata (Tavg), dan arah angin (ddd). Informasi ini memberikan wawasan yang berharga tentang faktor-faktor utama yang mempengaruhi prediksi hujan dalam konteks cuaca jangka pendek.

Dengan demikian, hasil penelitian ini tidak hanya mengukur kinerja model, tetapi juga memberikan pemahaman yang lebih dalam tentang faktor-faktor yang relevan dalam prediksi cuaca, yang dapat menjadi dasar untuk perbaikan dan pengembangan lebih lanjut pada model prediksi hujan menggunakan algoritma Classification Tree.

DAFTAR PUSTAKA

Prasetya, R. (2020). *Penerapan Teknik Data Mining Dengan Algoritma Classification Tree Untuk Prediksi Hujan*. Jurnal Widya Climago, 2(2), 13–23.

Nursobah, N., Lailiyah, S., Harpad, B., & Fahmi, M. (2022). *Penerapan Data Mining Untuk Prediksi Perkiraan Hujan dengan Menggunakan Algoritma K-Nearest Neighbor*. Building of Informatics, Technology and Science (BITS), 4(3). https://doi.org/10.47065/bits.v4i3.2564

Lareno, B. (2015). *Penerapan Algoritma MKNN-X Untuk Prediksi Curah Hujan*. GEMA AKTUALITA, 4(2).

Subhan Panji Cipta. (2016). *Penerapan Algoritma Evolving Neural Network Untuk Prediksi Curah Hujan*. Jurnal Teknologi Informasi Universitas Lambung Mangkurat (JTIULM), 1(1), 1–8. https://doi.org/10.20527/jtiulm.v1i1.2

Rofiq, H., Pelangi, K. C., & Lasena, Y. (2020). *Penerapan Data Mining Untuk Menentukan Potensi Hujan Harian Dengan Menggunakan Algoritma Naive Bayes*. Jurnal Manajemen Informatika Dan Sistem Informasi, 3(1), 8–15. Retrieved from http://mahasiswa.dinus.ac.id/docs/skripsi/jurnal/19417.pdf

Astuti, R. (2018). *Data Mining untuk Klasifikasi dengan Algoritma Cart (Classification and Regression Trees)*. Media Informatika, 17(3), 114–124. https://doi.org/10.37595/mediainfo.v17i3.15

Sofyan, F. M. A., Riyandoro, A. P., Maulana, D. F., & Jaman, J. H. (2023). *Penerapan Data Mining dengan Algoritma C5.0 Untuk Prediksi Penyakit Stroke*. Jurnal Teknologi Sistem Informasi Dan Sistem Komputer TGD, 6(2), 619–625.

Derisma, D. (2020). *Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining*. Journal of Applied Informatics and Computing, 4(1), 84–88. https://doi.org/10.30871/jaic.v4i1.2152