PENERAPAN TEKNIK *DATA MINING* DENGAN ALGORITMA *CLASSIFICATION TREE* UNTUK PREDIKSI HUJAN

Ratih Prasetya
Pusat Pendidikan dan Pelatihan, Badan Meteorologi Klimatologi dan Geofisika
ratih.prasetya@bmkg.go.id

ABSTRACT

Classification is a data mining technique used to predict the relationship between data in a dataset. Prediction is done by classifying data into several different classes by considering certain factors. Classification is one of the empirical approaches that can be used for short-term weather prediction. The classification algorithm used in this study is the Classification Tree utilizing software of Orange Data Mining 3.3.12. Furthermore, the algorithm is used to predict rain with the Confusion Matrix test parameters. The input data is a synoptic data from the Kemayoran Meteorological Station, Jakarta (96745) for 10 years (2006 - 2015) as many as 3528 datasets and consists of 8 attributes. Based on a series of processing, selection and testing of the model shows that the accuracy of the Classification Tree algorithm is 74.7% with a fair classification category where the number of correct predictions is 818 datasets out of the total amount of data tested that is 1095 datasets. The dominant weather attributes in the formation of rain respectively are humidity (RHavg), minimum temperature (Tmin), maximum temperature (Tmax), average temperature (Tavg) and wind direction (ddd).

***Keywords**: Classification, Supervised Learning, Confusion Matrix, Weather Attributes*

ABSTRAK

Klasifikasi adalah salah satu teknik *data mining* yang digunakan untuk memprediksi hubungan antar data pada suatu dataset. Prediksi dilakukan dengan mengklasifikasikan data menjadi beberapa kelas berbeda dengan mempertimbangkan faktor tertentu. Klasifikasi merupakan salah satu teknik pendekatan empiris yang dapat dimanfaatkan untuk prediksi cuaca jangka pendek. Algoritma Klasifikasi yang digunakan pada studi ini adalah *Classification Tree* menggunakan perangkat lunak *Orange Data Mining 3.3.12.* Selanjutnya algoritma tersebut digunakan untuk memprediksi hujan dengan parameter uji *Confusion Matrix*. Data masukan adalah data sinoptik Stasiun Meteorologi Kemayoran, Jakarta (96745) selama 10 tahun (2006 - 2015) sebanyak 3528 dataset dan terdiri dari 8 atribut. Berdasarkan serangkaian pengolahan, pemilihan dan pengujian model menunjukkan bahwa tingkat akurasi algoritma *Classification Tree* yaitu sebesar 74.7% dengan kategori *fair classification* dimana jumlah prediksi benar adalah 818 dataset dari total jumlah data yang diuji yaitu 1095 dataset. Atribut cuaca yang dominan dalam pembentukan hujan secara berturut-turut adalah kelembaban (*RHavg*), temperatur minimum (*Tmin*), temperatur maksimum (*Tmax*), temperatur rata-rata (*Tavg*) dan arah angin (*ddd*).

Kata kunci: Klasifikasi, *Supervised Learning, Confusion Matrix,* Atribut Cuaca

PENDAHULUAN

Prediksi cuaca merupakan salah tantangan besar dalam ilmu meteorologi yang telah banyak dijadikan subjek penelitian. Penelitian mengenai prediksi cuaca dilakukan dengan berbagai metode dimana setiap metode memiliki kekurangan dan kelebihan. Pendekatan dalam prediksi cuaca dapat dilakukan dengan metode empiris maupun dinamis. Prediksi cuaca jangka pendek dilakukan dengan penerapan metode dinamis yang merupakan sebuah pendekatan analitis yang didasarkan pada prinsip-prinsip dinamika fluida, sedangkan metode empiris yang dilakukan dengan pendekatan statistik dan matematis lebih banyak digunakan untuk prediksi cuaca jangka panjang.

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 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.

Unsur cuaca yang paling sering dilakukan prediksi adalah hujan. Hujan merupakan jatuhan hydrometeor yang merupakan partikel-partikel air yang memiliki diameter 0.5 mm atau lebih yang sampai ke tanah(Soepangkat, 1994).

Data mining adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar (Davies, 2004)

Para peneliti di bidang meteorologi telah mengkaji bagaimana mendapatkan metode prediksi yang tepat dan akurat dengan teknik data mining. Berdasarkan penelitian tersebut, disimpulkan bahwa penerapan teknik data mining untuk prediksi cuaca dengan menganalisis parameter cuaca. Berdasarkan penelitian tersebut, disimpulkan bahwa penerapan teknik data mining untuk prediksi cuaca dengan menganalisis parameter cuaca dapat dilakukan dan mendapatkan nilai akurasi yang baik. Algoritma *Decision Tree* menghasilkan akurasi yang baik 88.2% saat diaplikasikan dengan data cuaca karena dapat mengklasifikan dengan baik (E.Manjula, 2016).

Penelitian yang berjudul *Use of Data Mining Techniques for Weather Data in Basra City* menggunakan teknik *K Means Clustering* dan *ANN* untuk mencari nilai akurasi dan nilai *Root Means Square Error (RMSE)*sebagai metode klasifikasi untuk prediksi cuaca (hujan, cerah, berawan). Parameter cuaca yang digunakan adalah kelembaban, temperature rata-rata, kecepatan angin, arah angina, waktu kejadian angin maksimum dan curah hujan dengan periode data selama 9 tahun (2004-2013). Hasil yang penelitian adalah metode tersebut cukup baik untuk prediksi cuaca (Prasad and Nejres, 2015).

Data mining dibagi menjadi beberapa bagian berdasarkan tugas yang dilakukan, (Larose, 2006) salah satunya adalah prediksi. Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

Klasifikasi dan prediksi adalah dua bentuk analisis data yang bisa digunakan untuk mengekstrak model dari data yang berisi kelas-kelas atau untuk memprediksi *trend* data yang akan datang (Han dan Kamber, 2006).

Classification dan Regression Tree (CART) adalah salah satu metode atau algoritma dari teknik pohon keputusan. *CART* adalh suatu metode statistic nonparametric yang dapat menggambarkan hubungan antara variable respon (variabel dependen) dengan

satu atau lebih variabel predictor (variabel independen). Apabila variabel respon berbentuk kontinu maka metode yang digunakan adalah metode regresi pohon (*regression tree*), sedangkan apabila variabel respon memiliki skala kategorik maka metode yang digunakan adalah metode klasifikasi pohon (*classification tree)* (Breiman, 1993).

Pembentukan *classification tree* terdiri atas 3 tahap yang memerlukan *learning sample* L. Tahap pertama adalah pemilihan pemilah. 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):

$$i(t) = \sum_{i \neq j} p(i|t)p(j|t) \quad \text{..... (1)}$$

dengan *i (t)* adalah fungsi keheterogenan indeks gini, *p* (i|t) adalah proporsi kelas *i* pada simbul *t,* dan *p* (j|t) adalah proporsi kelas *j* pada simpul *t*. *Goodness of Split* merupakan suatu evaluasi pemilahan oleh pemilah *s* pada simpul *t. Goodness of split* ϕ (*s,t*) didefinisikan sebagai penurunan keheterogenan.

$$\emptyset(s,t) = \Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) \quad \text{..... (2)}$$

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

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

Dengan $\phi(s,t)$ adalah kriteria *goodness of split,* $P_L i$ (t_L) adalah proporsi pengamatan dari simpul *t* menuju simpul kanan. Tahap ketiga adalah penandaan label tiap simpul terminal berdasar aturan jumlah anggota kelas terbanyak, yaitu:

$$p(j_0|t) = \max_j p(j|t) = \max_j \frac{N_j(t)}{N(t)} \quad \text{.....(4)}$$

Selanjutnya, rumusan masalah pada penelitian ini adalah bagaimana penerapan teknik data mining untuk prediksi hujan, berapa akurasi hasil prediksi dengan parameter uji *Confusion Matrix* dan bagaimana hasil prediksi hujan dapat divisualisasikan menggunakan algoritma *Classification Tree. Cross Validation* merupakan metode umum yang digunakan untuk mengevaluasi kinerja *classifier. Cross Validation* adalah bentuk sederhana dari teknik statistic. Jumlah *fold* standar untuk memprediksi tingkat *error* dari data adalah dengan menggunakan 10-*fold cross validation* (Witten, Frank and Hall, 2011).

Atribut kualitas prediksi (*attributes of forecast quality*) untuk kategori *probabilistic* yaitu *sharpness*, *resolution*, *discriminant*, *bias*, *reliability (calibration)*, *accuracy* dan *skill* (Murphy and Wrinkler, 1992). Atribut yang akan dibahas pada penelitian ini adalah *accuracy*.

Diharapkan hasil penelitian ini dapat memberikan manfaat yaitu pengetahuan terkait aplikasi pendekatan empiris dalam teknik data mining untuk prediksi hujan dan menambah wawasan pengetahuan teknik prediksi hujan berbasis probabilistic.

METODE

Jenis penelitian ini adalah penelitian kuantitatif (*quantitative research*). Penelitian ini menggunakan teknik data mining untuk mengolah *series* data pengamatan cuaca dan menganalisis hasilnya untuk mendapatkan serta menentukan algoritma apa yang paling baik digunakan untuk prediksi cuaca.

Data yang digunakan dalam penelitian ini adalah data observasi meteorologi permukaan (*synoptic*) selama 10 tahun (2006-2015) yang diperoleh dari Sub Bidang Database BMKG dan Stasiun Meteorologi Jakarta (96745). Data observasi mateorologi permukaan (*synoptic)* diamati setiap jam, data *synop* yang berbentuk sandi kemudian diterjemahkan dan di input kedalam Ms Excel. unsur unsur cuaca yang diamati adalah 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.

Perangkat lunak yang digunakan untuk mengolah data adalah software *Orange Data Mining* versi 3.3.12 yang berbasis *open source.* Tahap *preprocessing* yang digunakan dalam penelitian ini menggunakan tahapan KDD (*Knowledge Data Discovery*) yang meliputi kegiatan pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Pramudiono, 2007). *Knowledge Database Discovery (KDD)* adalah proses menentukan informasi yang berguna serta pola-pola yang ada dalam data. Informasi ini terkandung dalam basis data yang berukuran besar yang sebelumnya tidak diketahui dan potensial bermanfaat (Han dan Kamber, 2006). Salah satu tahapan dalam *Knowledge Database Discovery* (*KDD*)*,* prosesnya dapat dijelaskan sebagai berikut (Fayyad, Piatetsky-Shapiro and Smyth, 1996).

1. Data Collection

Yaitu pengumpulan data yang akan digunakan pada processing teknik data mining. Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. Data Cleaning

Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Juga dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

Proses *cleaning* mempunyai fungsi yaitu membuang duplikasi data, memeriksa data yang tidak konsisten serta memperbaiki kesalahan data tipografi sehingga data tersebut siap untuk dilakukan proses selanjutnya yaitu pemilahan data.

3. Data Selection

Pada tahap ini, data yang sesuai untuk dilakukan analisis dipilih dari dataset. Data meteorologi yang digunakan pada penelitian ini yaitu temperature rata-rata harian, temperature minimum harian, temperature maksimum harian, kelembaban, kecepatan angin, arah angin dan lama penyinaran matahari. Parameter-parameter cuaca tersebut yang selanjutnya disebut atribut akan disimpan ke dalam *dataset* baru menggunakan *Microsoft Excel* dengan format xls berjumlah 3528 dataset dengan 8 variabel yang terdiri atas 7 variabel numerik dan 1 variabel textual. Hasil pengolahan pada tahap ini yaitu data memiliki 9 atribut (Tabel 1) dan terdapat 0 *missing value.*

Tabel 1. Atribut Dataset Meteorologi

Atribut	Tipe	Keterangan
Year	*Numerical*	*Year considered*
Month	*Numerical*	*Month considered*
Temperatur rata-rata	*Numerical*	Temperatur rata-rata harian
Temperatur Minimum	*Numerical*	Temperatur minimum harian
Temperatur Maksimum	*Numerical*	Temperatur maksimum harian

Kelembaban	*Numerical*	Kelembaban rata-rata dalam satu hari
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

4. Data Transformation

Data transformasi disebut juga dengan penggabungan data. Pada tahap ini data yang telah dipilih dirubah kedalam bentuk data yang sesuai untuk data mining

5. Validasi

Tahap validasi dilakukan dengan membagi data menjadi 2 yaitu data t*raining* (data latih) sebanyak 70% dari jumlah total dataset dan data *test* (data uji) sebanyak 30% dari jumlah total data set. Data *training* digunakan untuk *processing* algoritma *Classification Tree* sedangkan data test digunakan untuk proses validasi. Pendekatan yang digunakan adalah Dalam pendekatan *cross validation,* setiap record digunakan beberapa kali dalam jumlah yang sama untuk *training* dan tepat sekali untuk *testing*. Metode ini mempartisi data ke dalam dua sub set data yang berukuran sama. Pilih salah satu sebagai *data training* dan satu lagi untuk *testing,* kemudian dilakukan pertukaran fungsi dari sub set sedemikian sehingga sub set yang sebelumnya sebagai *training set* menjadi *test set* demikian sebelumnya. Pendekatan ini dinamakan *two-fold-cross-validation*. Total error diperoleh dengan menjumlahkan error-error untuk kedua proses tersebut.

6. Evaluasi/Pengujian

Cross Validation merupakan metode umum yang digunakan untuk mengevaluasi kinerja classifier. Cross Validation adalah bentuk sederhana dari teknik statistic. Jumlah fold standar untuk memprediksi tingkat error dari data adalah dengan menggunakan 10-fold cross validation (Witten, Frank and Hall, 2011).

Dalam pendekatan cross validation, setiap record digunakan beberapa kali dalam jumlah yang sama untuk training dan tepat sekali untuk testing. Metode ini mempartisi data ke dalam dua sub set data yang berukuran sama. Pilih salah satu sebagai data training dan satu lagi untuk testing, kemudian dilakukan pertukaran fungsi dari sub set sedemikian sehingga sub set yang sebelumnya sebagai training set menjadi test set demikian sebelumnya. Pendekatan ini dinamakan two-fold-cross-validation. Total error diperoleh dengan menjumlahkan error-error untuk kedua proses tersebut.

	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*

Pada gambar 1 ditunjukkan contoh nilai penghitungan dengan nilai fold yang digunakan adalah 4-fold *cross validation*. Berikut diberikan langkah-langkah pengujian data dengan 4-fold cross validation.

Parameter yang digunakan untuk evaluasi komparasi algoritma adalah *Confusion Matrix* (akurasi, presisi dan *recall*). Evaluasi dengan *confusion matrix* menghasilkan akurasi dan laju *error*. Analisis akurasi dalam prediksi dikotomi yaitu prediksi dalam bentuk dua kategori hujan atau tidak dapat dilakukan untuk mengetahui tingkat ketepatan algoritma klasifikasi yang digunakan. Nilai akurasi diperoleh dari suatu matriks kontingensi yaitu suatu matriks bujur sangkar yang disebut "*error matrix*" atau "*confusion matrix*". Akurasi adalah persentase total data yang diprediksi secara benar. Laju *error* adalah persentase dari total data yang diprediksi secara salah, sebagai berikut:

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

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

Setelah dibuat *confusion matrix,* selanjutnya dihitung nilai *precision*, r*ecall* dan *accuracy.* Untuk mengukur *accuracy, precision* dan *recall* biasanya digunakan *confusion matrix.*

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

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

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

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

7. Interpretasi

Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya. Tahap interpretasi digunakan untuk mendapatkan pola informasi yang dihasilkan dari proses *data mining.* Informasi yang dihasilkan pada software *Orange* akan menampilkan dan memberikan informasi mengenai kinerja masing-masing algoritma dalam metode Klasifikasi.

HASIL DAN PEMBAHASAN

Pada **tahap Data Collection**, data yang digunakan adalah data observasi meteorologi permukaan (*synoptic*) rata-rata harian selama 10 tahun (2006-2015) sebanyak 3528 dataset. Lokasi penelitian adalah Stasiun Meteorologi Kemayoran Jakarta (96745) yang beralamat di Jalan Angkasa I No. 2, Kemayoran, Jakarta. Unsur unsur cuaca yang diamati adalah

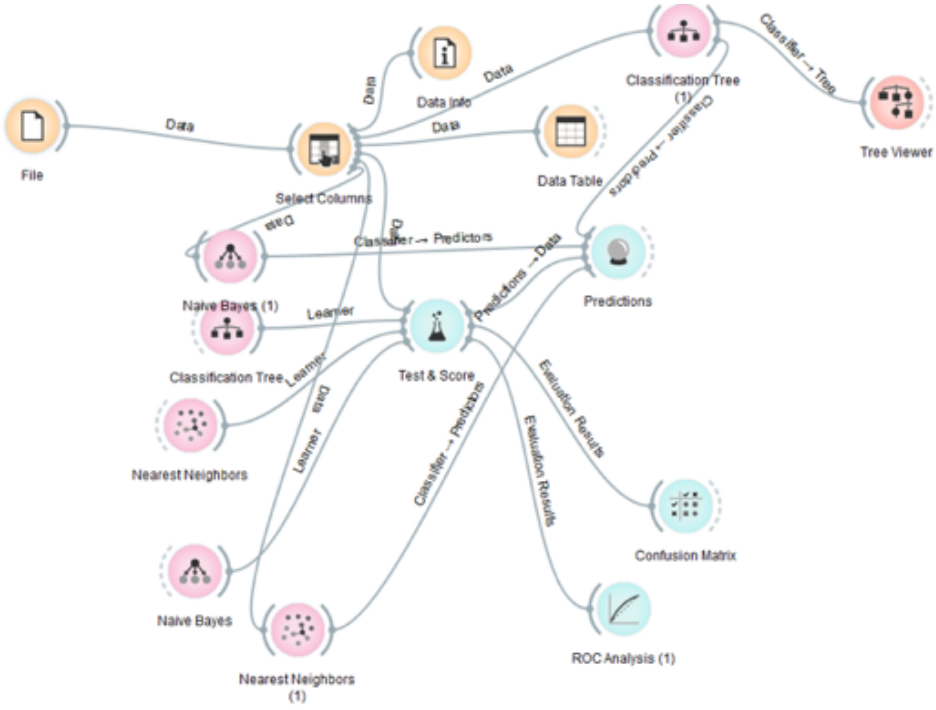
Temperatur, Tekanan, *Visibility* (Jarak Pandang), keadaan cuaca, arah angin, kecepatan angin, titik embun, jenis awan, jumlah awan, radiasi matahari, lamanya penyinaran matahari dan lain-lain.

Pada **tahap *Data Cleaning***, data meteorologi yang digunakan pada penelitian ini memiliki 9 atribut (Tabel 1) dan terdapat 0 *missing value.*

Tahap *Data Selection* dilakukan proses analisis data dengan memilih parameter cuaca yang relevan terhadap penelitian ini yaitu temperature rata-rata harian, temperature minimum harian, temperatur maksimum harian, kelembaban, kecepatan angin, arah angin dan lama penyinaran matahari. Parameter-parameter cuaca tersebut yang selanjutnya disebut atribut akan disimpan ke dalam *dataset* baru menggunakan *Microsoft Excel* dengan format xls berjumlah 3528 dataset dengan 8 variabel yang terdiri atas 7 variabel numerik dan 1 variabel textual.

Tahap *Data Transformation* dilakukan dengan merubah formal xls menjadi format csv sesuai dengan tipe data input pada *software Orange.*

Tahap Validasi, proses perancangan model pada *software orange* adalah sebagai berikut:



Gambar 2. Proses Klasifikasi Pada Software Orange Vers. 3.3.1.2

Data observasi cuaca permukaan (*synoptic*) di proses menggunakan algoritma klasifikasi yaitu *Classification Tree.* Model keluaran masing-masing metode tersebut diuji dengan sebagian data masukan untuk melihat kehandalan model.

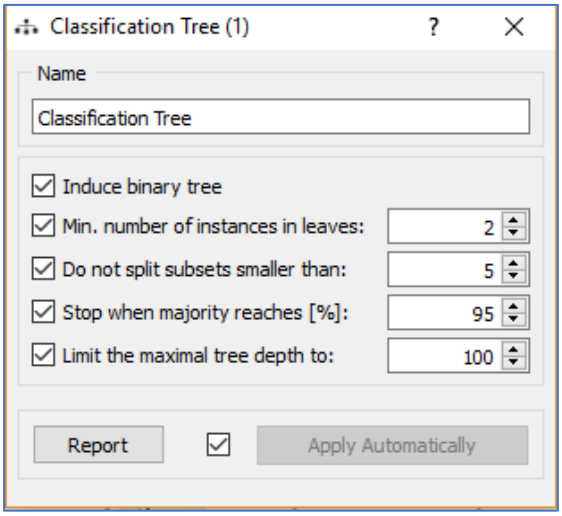
Tahap Evaluasi/Pengujian dilakukan dengan membagi data menjadi 2 (dua) yaitu data latih (*training*) dan data uji (data *test*). Data latih sebanyak 70% digunakan sebagai proses *mining* dan mendapatkan nilai probabilitas sedangkan data uji sebanyak 30% digunakan untuk menguji nilai probabilitas yang telah terbentuk. Uji dengan *confusion matrix* dilakukan untuk memperoleh nilai *precision*, *recall* dan *accuracy* dari hasil pengujian. Hasil pengujian adalah untuk mengukur tingkat akurasi dan *Area Under Curve* (AUC) dari penentuan dengan metode *10-fold Cross Validation*. Berikut hasil pengujian dari masing-masing algoritma:

Tabel 2. *Confusion Matrix* pada *Classification Tree*

***Accuracy*: 74.7%**		
	True Normal (Hujan)	***True Anomaly (Tidak Hujan)***
Pred. Normal (Hujan)	286	532
Pred Anomaly (Tidak Hujan)	140	137
Proportion of Predicted	67.1%	79.5%

Dari table 2 kemudian dihitung nilai: *Precision* = 0.74; *Recall* = 0.74; *Accuracy* = 0.74 dan *AUC* = 0.73

Berdasarkan hasil pada Tabel 2, dapat dilihat bahwa tingkat akurasi dengan menggunakan algoritma *Classification Tree* adalah sebesar 74.7% dengan jumlah prediksi benar adalah 818 dataset dari jumlah total data yang diuji yaitu 1095 dataset.



Gambar 3. Pengaturan *Classification Tree* pada *Software Orange*

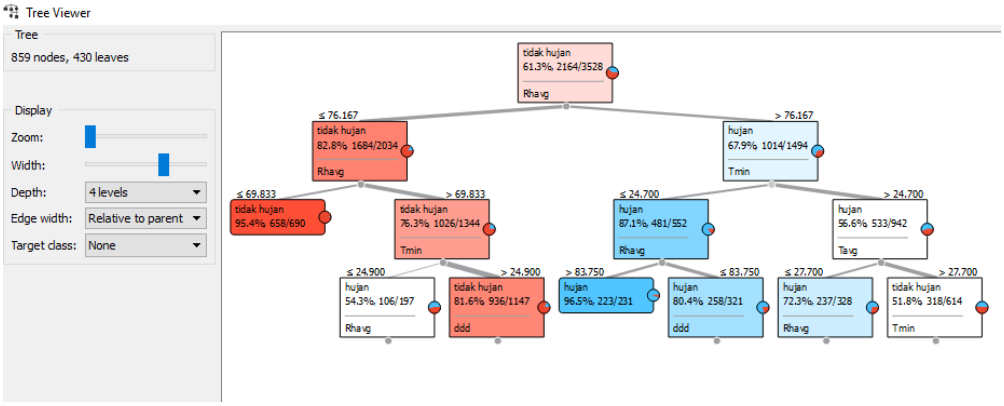
Tahap Interpretasi pada algoritma *Classification Tree* merupakan salah satu metode Klasifikasi yang paling popular karena mudah untuk diinterpretasi. *Classification Tree* adalah model prediksi menggunakan struktur pohon atau disebut juga struktur berhirarki.

Model *Classification Tree* memiliki kemampuan untuk mengubah data menjadi pohon keputusan dan aturan-aturan keputusan sehingga dapat memproses pengambilan keputusan dengan atribut yang kompleks menjadi lebih mudah. Pengaturan *widget Classification Tree* pada Orange Software dapat dilihat pada gambar 3.

Min. Number of Instances in Leaves yaitu dimana algoritma tidak akan menghasilkan node selain jumlah yang telah ditentukan. *Do Not Split Subsets Smaller Than* berfungsi untuk mencegah algoritma memecah simpul dengan jumlah lebih kecil dari *instance* yang ditentukan. *Limit the Maximal Tree Depth* berfungsi untuk membatasi jumlah kedalaman simpul hingga mencapai jumlah kedalaman simpul yang ditentukan. *Stop When Majority Reaches (%)* menghentikan simpul setelah *threshold* tertentu dicapai. Masalah dalam Klasifikasi dapat diproses dengan mengajukan sejumlah atribut dari test record.

Setiap kali node diperoleh maka sejumlah atribut diproses kembali sampai diperoleh sebuah kesimpulan mengenai label kelas dari *record*. Rangkaian proses tersebut dapat diasosiasikan ke dalam bentuk pohon

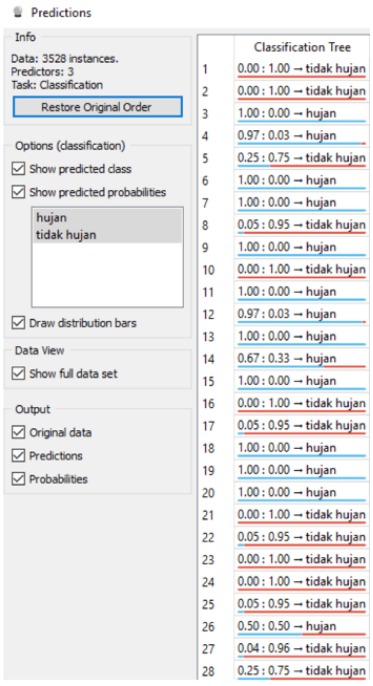
keputusan yang merupakan struktur hirarki yang terdiri dari node dan simpul. Pohon keputusan hasil pengolahan algoritma *Classification Tree* dapat dilihat pada gambar 4.



Gambar 4. *Classification Tree* untuk prediksi hujan

Hasil pengolahan data pada *Classification Tree* menghasilkan 859 nodes dan 430 leaves dengan *depth* sebanyak 5 level. *Root node* yaitu atribut *Receiver Humidity* rata-rata *(RHavg)* digunakan untuk memisahkan kondisi cuaca hujan atau tidak hujan, dari jumlah dataset sebanyak 3528 terdapat 2164 dataset atau 61.3% peluang cuaca tidak hujan. Selanjutnya, dari *root node* dipecah menjadi 2 *internal node* yaitu atribut RHavg dan *temperature minimum* (Tmin). Bila nilai RH ≤ 76.167% maka kondisi cuaca berpeluang tidak hujan dan bila nilai RH > 76.167% maka kondisi cuaca berpeluang hujan. Dari *internal node* Tmin dipecah lagi menjadi 2 *internal node* yaitu RHavg dan *temperature* rata-rata (Tavg), bila Tmin ≤ 24.7^0 C maka kondisi cuaca berpeluang hujan dan bila Tmin > 24.7^0 C maka kondisi cuaca berpeluang kecil hujan.

Pada proses terakhir, *Internal node* RH menghasilkan *leaf node* hujan bila nilai RH > 83.75% maka peluang terjadinya hujan cukup besar yaitu 96.5% dan bila nilai RH ≤ 69.83% maka peluang tidak hujan cukup besar yaitu 95.4%.



Gambar 5. Prosentase prediksi hujan pada tiap dataset

Pada Gambar 5, juga dapat dilihat prosentase prediksi hujan pada tiap data set, dimana total terdapat 3528 dataset. *Classification Tree* berhasil mengklasifikasikan parameter apa saja yang paling berpengaruh terhadap prediksi curah hujan yaitu berdasarkan tingkat pengaruhnya secara berturut-turut yaitu RHavg, Tmin, RH, ddd (arah angin), LPM (lamanya penyinaran matahari) dan Tmax. Selain itu dari pohon klasifikasi yang terbentuk juga dapat diketahui peluang intensitas hujan yang akan terjadi, hal ini tampak dari warna yang lebih kuat pada masing-masing *node*.

KESIMPULAN

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

1. Bahwa berdasarkan hasil dari parameter uji *Confusion Matrix*, algoritma *Classification Tree* dapat diaplikasikan untuk prediksi hujan dengan kategori yang cukup baik yaitu *fair classification*.
2. Interpretasi pengetahuan yang dapat diaplikasikan untuk prediksi cuaca jangka pendek khususnya untuk algoritma *Classification Tree* adalah bahwa algoritma tersebut berhasil mengklasifikasikan data uji sebanyak 1095 dataset menjadi 287 *nodes* dan 144 *leaves*.
3. Parameter cuaca yang paling signifikan terhadap pembentukan hujan adalah kelembaban (RHavg), temperature minimum (Tmin), temperature maksimum (Tmax), temperature rata-rata (Tavg) dan arah angin (ddd).

SARAN

Penelitian ini perlu dikembangkan kembali dengan:

1. Penelitian ini sebaiknya dilanjutkan dengan menggunakan parameter uji yang lain seperti RMSE (*Root Mean Squared Error)* atau dengan menambah jumlah parameter uji.
2. Metode empiris dengan model data mining ini sebaiknya diuji tidak hanya pada satu titik stasiun cuaca saja tetapi menggunakan sebaran data observasi beberapa stasiun cuaca (spasial) sehingga dapat merepresentasikan wilayah tertentu dengan lebih baik.

DAFTAR PUSTAKA

Bhatkande, S. S. and Hubballi, R. G. (2016) 'Weather Prediction Based on Decision Tree Algorithm Using Data Mining Techniques', 5(5), pp. 483–487. doi: 10.17148/IJARCCE.2016.55114.

Chauhan, D. and Thakur, J. (2014) 'Data Mining Techniques for Weather Prediction: A Review', *International Journal on Recent and Innovation Trends in Computing and Communication*, 2(8), pp. 2184–2189. Available at: http://ijritcc.org/IJRITCC Vol_2 Issue_8/Data Mining Techniques for Weather Prediction A Review.pdf.

E.Manjula, S. D. (2016) 'Analysis of Data Mining Techniques for Agriculture Data', *International Journal of Computer Science and Engineering Communications*, 4(2), pp. 1311–1313. doi: 10.7910/DVN/MYBLHC.

Gaikwad, G. and Nikam, V. B. (2013) 'Different Rainfall Prediction Models And General Data Mining Rainfall Prediction Model', *International Journal of Engineering Research and Technology*, 2(7), pp. 115–123.

Prasad, R. S. and Nejres, S. M. (2015) 'International Journal of Advanced Research in Use of Data Mining Techniques for Weather Data in Basra City', 5(12), pp. 135–139.

Tan, P. and Steinbach, M. (2006) 'Introduction to Data Mining Instructor ' s Solution

Manual'.

Tobergte, D. R. and Curtis, S. (2013) 'Metode Classification', *Journal of Chemical Information and Modeling*, 53(9), pp. 1689–1699. doi: 10.1017/CBO9781107415324.004.

Witten, I. H., Frank, E. and Hall, M. a (2011) *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)*, *Complementary literature None*. doi: 0120884070, 9780120884070.