Institut Riset dan Publikasi Indonesia (IRPI)
SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat
Homepage: https://journal.irpi.or.id/index.php/sentimas
Prosiding Seminar Nasional/ ISSN(E): 2830-3083
25 Agustus 2022/ Pages: 187-197

Comparison of Data Mining Methods for Predition of Rainfall with C4.5, Naïve Bayes, and KNN Algorithm

Perbandingan Metode Data Mining untuk Prediksi Curah Hujan dengan Algoritma C4.5, Naïve Bayes, dan KNN

Alfian Al Arif[1*], Muhammad Firdaus[2], Rahmaddeni[3], Yustis Maruhawa[4]

[1,2,3,4]STMIK AMIK Riau, Jl. Purwodadi Panam, Pekanbaru, Indonesia

E-Mail: [1]2110031802139@sar.ac.id, [2] 2110031802146@sar.ac.id,
[4]rahmaddeni@sar.ac.id, [3]2110031802137@sar.ac.id

Corresponding Author: Rahmaddeni

Abstract

Rain is one thing that must be observed because it is classified as rainfall. The Meteorology, Climatology and Geophysics Agency (BMKG) is one of the government agencies in charge of conveying weather information. For rainfall there are standards that will be achieved by BMKG, namely temperature, humidity, and wind speed. This rainfall dataset was taken from the BMKG database in Jatiwangi, Majalengka from 01/2/2008 to 26/12/2018 which was taken from www.bmkg.go.id. To estimate the rainfall, a data mining method with a classification function is used. The Discovery Knowledge of Databases (KDD) process usually begins with the data selection stage, pre-processing (cleaning data), changing data, data mining and evaluation. The method used for this classification data mining model consists of three algorithms, namely C4.5 or Decision tree, k-nearest neighbor (kNN,) and Naïve Bayes. The software used to process the data is Rapid Miner. The final result of the three algorithms used is that the C4.5 algorithm is the best algorithm for estimating rainfall with accuracy (88.03%) and error (11.97%).

Keyword: C4.5, Data Mining, KNN, Naïve Bayes, Rainfal.

Abstrak

Hujan adalah salah satu hal yang harus diamati karena tergolong curah hujan. Badan Metereologi Klimatologi dan Geofisika (BMKG) adalah salah satu lembaga pemerintahan yang bertugas menyampaikan informasi cuaca. Untuk curah hujan terdapat standar yang akan dicapai oleh BMKG yaitu suhu, kelembaban dan kecepatan angin. Dataset curah hujan ini diambil dari database BMKG Jatiwangi, Majalengka mulai tanggal 01/2/2008 sampai tanggal 26/12/2018 yang diambil dari www.bmkg.go.id. Untuk memperkirakan curah hujan tersebut dipakai metode *data mining* dengan fungsi klasifikasi. Proses *Discovery Knowledge of Databases* (KDD) biasanya diawali dari langkah seleksi data*, pre-processing (cleaning data),* mengubah data*, data mining* dan evaluasi*.* Pada penelitian ini digunakan 3 (tiga) metode algoritma yaitu C4.5 atau *Decision tree*, *k-nearest neigbor* (kNN,) dan *Naïve Bayes*. *Software* yang dipakai untuk memproses data adalah *Rapid Miner*. Kesimpulan dari ketiga algoritma yang dipakai didapatkan algoritma C4.5 adalah algoritma terbaik untuk memperkirakan curah hujan dengan nilai *accuracy (*88,03%) dan *error* (11,97%).

Kata Kunci: Curah Hujan, C4.5, *Data Mining*, KNN, *Naïve Bayes*.

1. PENDAHULUAN

Curah hujan adalah salah satu jenis cuaca yang diprediksi oleh Badan Metereoligi Klimatologi dan Geofisika (BMKG). Dataset curah hujan ini diambil dari database BMKG Jatiwangi, Majalengka mulai tanggal 01/2/2008 sampai tanggal 26/12/2018. Dari data curah hujan yang terdapat di *database* tentu saja ada sejumlah variabel-variabel prediktor yang bisa dipakai untuk memperkirakan variabel target terdiri dari suhu rata-rata, kelembaban rata-rata, lama penyinaran, dan kecepatan angin rata-rata. Salah satu

cara untuk memperkirakan adalah memakai *data mining*, yang mana akan dipakai pola yang terdapat dalam *database* curah hujan untuk mengklasifikasi curah hujan yang terjadi. Cara-cara yang akan dipakai tergolong ke dalam fungsi klasifikasi dengan sejumlah algoritma.

Data mining adalah menggabungkan teknik analisis pola-pola yang penting. Atau bisa juga didefenisikan menjadi proses pemilahan data, observasi dan versi dari beberapa data untuk mendapatkan pola yang umumnya tidak disadari keberadaannya

Metode data mining cara yang diterapkan tetapi perlu disesuaikan dengan tujuan dari pengunaannya. Contoh metode dats mining adalah *K Nearest Neighbor, Support Vector Machine, Naïve Bayes,* C4.5 dan lainnya. Metode *C45, KNN dan Naïve Bayes Classifier* ialah algoritma yang dipakai untuk perbandingan nilai akurasi dan nilai *error.* Akurasi adalah ketepatan nilai ukur yang betul atas total jumlah *sample* dipertimbangkan. *Error* adalah skala nilai percobaan yang salah terhadap jumlah *sample* yang dipertimbangkan.

Beberapa penelitian tentang implementasi data *mining* untuk mencari informasi yang diperoleh pada *database* curah hujan diantaranya adalah menurut [1] yang membandingkan klasifikasi curah hujan memakai metode SVM dan NBC dimana diperoleh akurasi SVM 79,45% lebih baik dibandingkan NBC 65,75%. Menurut [2] melakukan analisa perbandingan data *mining* untuk memperkirakan nilai dan ketepatan kelulusan mahasiswa/i memakai algoritma Naïve Bayes, C.45, KNN, dan SVM. Hasil yang terbaik diperoleh algoritma Naïve bayes.

2. METODOLOGI PENELITIAN

a. Sumber Data

Dataset yang dipakai merupakan data public yang diperoleh dari BMKG Stasiun Meteorologi. Jatiwangi, Majalengka dan bisa dicari di website www.bmkg.go.id.

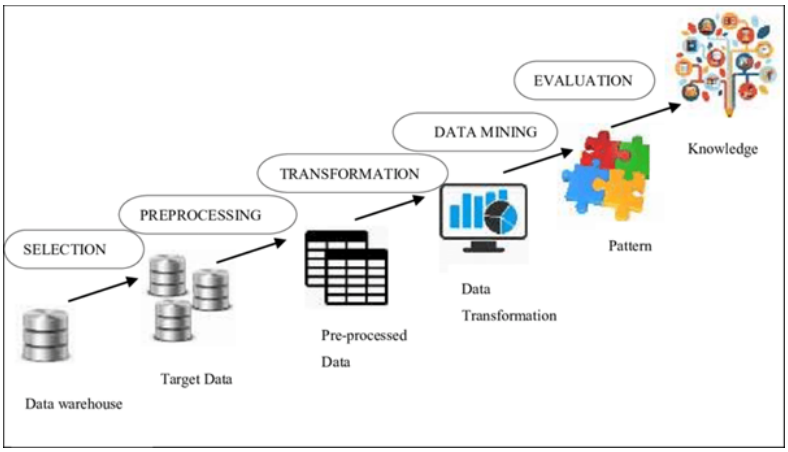
b. Variabel Penelitian

Ada variabel yang dipakai pada penelitian ini yaitu :

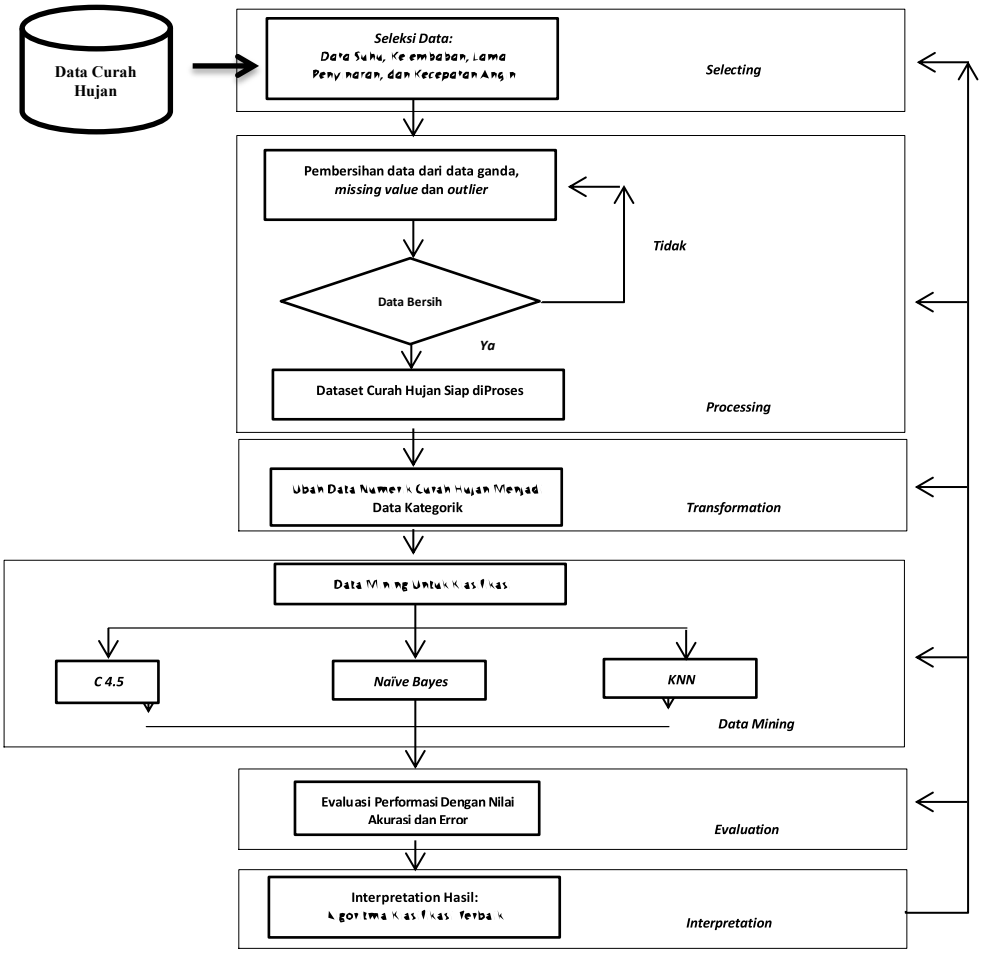
1. Variabel dependen (*Y*) adalah variabel ini dipengaruhi oleh variabel lain. Didalam penelitian ini adalah status curah hujan yang dibagi dua kategori yaitu hujan dan tidak hujan.
2. Variabel independen (*X*) adalah variabel ini mempengaruhi variabel lain. Didalam penelitian ini memakai empat variabel yaitu rata-rata temperatur, rata-rata kelembapan, rata-rata penyinaran matahari, dan rata-rata kecepatan angin.

c. Tahapan Analisis Data

Pada Gambar 6 merupakan deskripsi untuk setiap langkah-langkah yang akan dilalui. Proses *knowledge discovery in database* (KDD) diawali dengan menentukan tujuan dan diakhiri dengan evaluasi [3]. Urutan dari KDD bisa diperhatikan di Gambar 1 bawah ini :



Gambar 1 Langkah-langkah KDD



Gambar 2. Tahapan penelitian berdasarkan KDD

Dibawa ini adalah penjelasan dari tiap tahap penelitian pada gambar 1 dan gambar 2.

1. Seleksi Data
   Tahap ini dilakukan pemilahan data rata-rata, kelembaban rata-rata, lama penyinaran, dan kecepatan angin rata-rata cuaca yang terbagi dari variabel *predictor* dan target variabel.

2. *Preprocessing*
   Data yang diambil merupakan perkiraan curah hujan. Dari data yang diperoleh, dilakukan *cleaning* agar terdeteksi data ganda, data hilang dan data *outlier.*

3. *Transformation*
   Setelah data *clear* dari data ganda dan lainnya, lalu dilakukan perubahan data sesuai dengan jenis data pada dan akan dibagi menjadi data yang bersifat kategori yaitu kategori variabel Target dan variabel prediktor. Bisa diperhatikan pada Tabel 2 dibawah ini.

Tabel 1. Kategori variabel target dan variabel prediktor

Variabel Target	Kategori
Curah Hujan	Tidak Hujan Hujan
Variabel Prediktor	
Suhu Rata-rata	0 - 100 ^{0}C
Kelembaban Rata-rata	1 – 100 %
Lama Penyinaran	0 – 12 jam
Kecepatan Angin	1 – 10 Knot

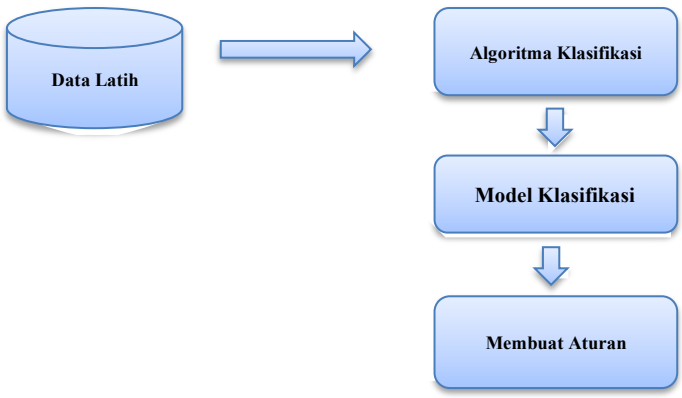
4. *Data Mining*
   Tahap ini melakukan proses untuk penentuan teknik data *mining* yang selaras dan menggunakan metode Naïve Bayes, C4.5, dan kNN dikarena klasifikasinya ialah supervised learning atau data yang telah ada label.

5. Evaluasi
   Proses dilakukan untuk menguji hasil perkiraan yang didapat dari ketiga metode dan memilih metode yang membuat nilai mengarah klasifikasi data awal.

d. Metode

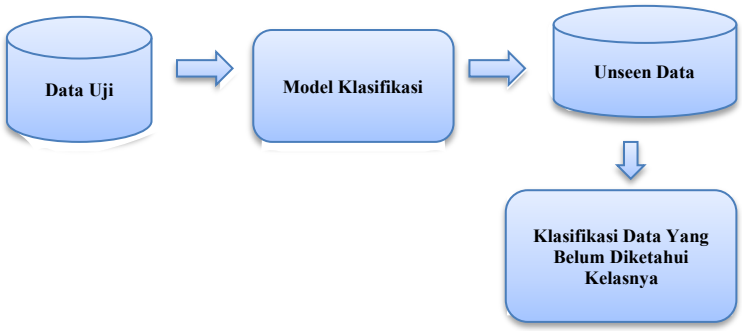
Algoritma data *mining* bisa dipisah menjadi tiga [1], adalah *supervised, unsupervised* dan *semi-supervised*. Pada *supervised learning* adalah menggunakan data yang telah mempunyai label. Pada *supervised learning* menggunakan data belum ada label atau kelasnya belum diketahui, teknik ini dipergunakan sebagai pengelompokkan berdasarkan kesamaannya. Sedangkan *semi- supervised learning* adalah beberapa data telah mempunyai label dan juga ada beberapa data yang belum mempunyai label. Proses pengujian terbagi menjadi dua tahapan [4]:

1. Tahap membuat model
   Tahap awal pengujian dibuat menurut data yang sudah diketahui kelasnya. Bagian ini disebut sebagai proses induksi yang digambarkan pada Gambar 3.



Gambar 3. Tahap membuat model

2. Tahap memakai model klasifikasi
   Langkah kedua belum diketahui kelasnya, bisa diperhatikan pada Gambar 4.



Gambar 4. Tahap memakai model

e. Algoritma Klasifikasi

Untuk menentukan algoritma apa yang akan dipakai atau yang cocok bisa dilihat dari beberapa factor antara lain kegunaan, tujuan atau dari hasil akhir. Algoritma klasifikasi sangatlah banyak, berikut metode yang akan dipakai dalam penelitian ini:

1. Algoritma C4.5

Algoritma C4.5 bisa dikatakan adalah pohon keputusan. Pohon keputusan berguna untuk mengekspolarasi data, menemukan tersembunyi antara sejumlah variable input dan variable target yang akan diuji. Algoritma *decision tree* masuk kedalam *supervised learning*. Sejumlah data latih

bisa disediakan untuk membangun metode dengan nilai-nilai variabel target. Persamaan (1) adalah rumus *information gain* [5]

$$Gain\ (S,) = (S) - \sum_{|S|}\ (S) \quad (1)$$

Keterangan :
A : atribut
$|S_v|$: nilai v untuk jumlah sampel
|S| : jumlah sampel untuk data seluruhnya

Entropi adalah keberagaman suatu data. Pada persamaan (2) merupakan rumus entropi :

$$Entropy\ (S) = -\sum p log_2 p \quad (2)$$

Dimana :
p_i = porsi atau rasio antara jumlah sampel kelas i dengan jumlah semua sampel pada himpunan data

2. *Naïve Bayes*

Klasifikasi Bayes adalah metode klasifikasi data mining menggunakan perhitungan probabilitas dan statistik, klasifikasi ini mengestimasi kesempatan keanggotaan kelas seperti probabilitas suatu tupel adalah milik kelas tertentu.

Rumus Naive Bayes [5] dapat dilihat dari persamaan (3) berikut :

$$P(Y \mid X) = P(Y) \prod P(X \mid Y) \quad (3)$$

keterangan:
P(X | Y) : probabilitas data dengan vektor X pada kelas Y
P(Y) : probabilitas awal kelas Y dan P(Xi | Y) adalah probabiltas independen kelas Y pada semua fitur dalam vektor X

3. KNN

Metode kNN adalah strategi untuk mencari kasus dengan menghitung kolelasi antara kasus baru dengan kasus lama. K-nearest neighbor (kNN) tegolong kelompok *instance-based learning*. Rumus metode kNN dapat dilihat dari persamaan (4)

$$d_{Euclidian\,(x,y)} = \sqrt{\sum_i (x_i - \ _i)^2} \quad (4)$$

f. Evaluation Tools

Tabel 2 dibawah merupakan *confusion matrix* untuk menguraikan ukuran performansi klasifikasi, Ukuran performansi tergolong kedalam tahapan penilaian.

Tabel 2. *Confusion Matrix*

Aktual	Prediksi	
	Positif	Negatif
Positif	*Benar* Positif	*Salah Negatif*
Negatif	*Salah* Positif	*Benar Negatif*

Tahapan performansi untuk teknik klasifikasi adalah *Area Under Receiver Operating Characteristics* (ROC), *Area Under* c*urve* (AUC), akurasi dan *error.* AUC (*area under curve*) memperkirakan untuk mengukur perbedaan performansi. Untuk klasifikasi *data mining*, nilai AUC dalam klasifikasi data *mining* dibagi menjadi lima kelompok [6]

a. 0.90 – 1.00 = Klasifikasi Sangat Baik (*Excellent Classification*)
b. 0.80 – 0.90 = Klasifikasi Baik (*Good Classification)*
c. 0.70 – 0.80 = Klasifikasi Cukup (*Fair Classification)*
d. 0.60 – 0.70 = Klasifikasi Buruk (*Poor Classification)*

e. 0.50 - 0.60 = Klasifikasi Salah (*failure)*

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FN + FP} \quad (5)$$

$$Error = \frac{\text{FN} + \text{FP}}{TP + TN + FN + FP} \quad (6)$$

g. Rapid Miner

Rapid Miner adalah alat data mining yang dikembangkan menggunakan Java sebagai bahasa pemrograman. Bahasa Java dianggap sebagai salah satu bahasa yang memiliki berbagai keunggulan, seperti kesederhanaan, keamanan, kekuatan, dampak, kemampuan berorientasi objek tingkat tinggi, dan banyak keunggulan lainnya [7]. Rapid Miner dikembangkan untuk tugas penambangan data umum. Sebagian besar versi sebelumnya adalah *open source* (lebih rendah dari 5), tetapi versi keenam diadopsi oleh beberapa opsi lisensi (*Starter, Personal, Professional, Enterprise*).

h. Penelitian Terdahulu

Sejumlah penelitian menggunakan data *mining* untuk mendapatkan penjelasan yang dimiliki oleh *database* curah hujan diantaranya adalah menurut [1] yang membandingkan hasil klasifikasi curah hujan memakai metode SVM dan NBC dimana diperoleh akurasi SVM 79,45% lebih baik dibandingkan NBC 65,75%. Dalam jurnal [8] melakukan analisa perbandingan data mining untuk estimasi nilai dan ketepatan kelulusan mahasiswa/i dengan algoritma Naïve Bayes, C.45, KNN dan SVM. Diperoleh Naïve bayes memberikan hasil yang terbaik.

Menurut [8] melakukan perbandingan algoritma SVM dan NBC dalam analisa sentimen pilkada pada twitter dengan hasil akurasi 81,7 pemanggilan 81,7 dan akurasi 80% adalah hasil NBC, Sedangkan akurasi 80,7 pemanggilan 80,7 dan akurasi 84% adalah hasil SVM. Maka dapat ditentukan bahwa algoritma NBC lebih tinggi dalam *accuracy and recall* sedangkan dalam *precision* yang lebih tinggi adalah algoritma SVM.

Menurut jurnal [9] Perbandingan Support Vector Machine dan XGBSVM Dalam Menganalisis Opini Publik Vaksinasi Covid-19, dapat kita ketahui SVM mendapatkan akurasi tertinggi yakni 83% dengan spliting data 90:10, kemudian XGBSVM menghasilkan akurasi 79% dengan spliting data 90:10.

3. ANALISA DAN PERANCANGAN

a. Seleksi Data

Dataset yang digunakan didapat dari database BMKG Jatiwangi, Majalengka mulai tanggal 01/2/2008 sampai tanggal 26/12/2018. Data nilai berasal dari alamat *website* www.bmkg.go.id.

b. Analisa Data

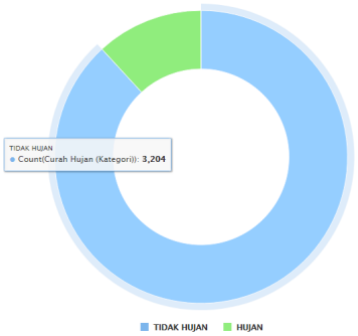
Untuk mengetahui curah hujan setiap harinya dibuat perhitungan kembali dari data suhu, kelembaban, lama penyinaran dan kecepatan angin sebab diperoleh beberapa duplikat data nilai maksimum dan minimum karena faktor cuaca yang seringkali berubah-ubah. Contoh dari akumulasi data dapat dilihat pada Tabel 3 di bawah ini:

Tabel 3. Data sebenarnya Curah hujan

Row No.	Curah Hujan...	Suhu Rata-r...	Kelembaban...	Lama Penyi...	Kecepatan ...
1	TIDAK HUJAN	26.100	82	2.300	3
2	TIDAK HUJAN	26.300	90	1	2
3	HUJAN	26	86	0.300	2
4	TIDAK HUJAN	26.900	84	0.900	2
5	TIDAK HUJAN	26.900	86	5	2
6	TIDAK HUJAN	26.600	87	1.800	2
7	TIDAK HUJAN	25.300	90	4.300	2
8	TIDAK HUJAN	24.400	93	0	3
9	TIDAK HUJAN	25.200	90	0.300	2
10	TIDAK HUJAN	24.400	95	0	2
11	TIDAK HUJAN	25.500	85	0.100	3
12	TIDAK HUJAN	25.600	88	1	2
13	TIDAK HUJAN	25.500	87	0	2
14	TIDAK HUJAN	24.200	92	0	2
15	TIDAK HUJAN	24.800	88	0	2

ExampleSet (3,634 examples, 1 special attribute, 4 regular attributes)

Berdasarkan data cuaca pada BMKG Jatiwangi, Majalengka tanggal 01/2/2008 hingga 26/12/2018 menghasilkan klasifikasi "Tidak Hujan" atau "Hujan" dapat dilihat pada Gambar 5 berikut.



Gambar 5. Perbandingan klasifikasi kelas "Tidak Hujan" atau "Hujan"

c. Transformasi data

Dataset yang ingin diuji di*import* pada tabel data awal yang akan dilakukan perubahan dibeberapa jenis data yang bersifat angka adalah curah hujan. Contoh data yang sudah dilakukan perubahan pada Tabel 4 dibawah ini.

Tabel 4. Data yang sudah diubah

Row No.	Curah Hujan...	Curah Hujan...	Suhu Rata-r...	Kelembaban...	Lama Penyi...	Kecepatan ...
1	1	0	26.100	82	2.300	3
2	1	0	26.300	90	1	2
3	0	1	26	86	0.300	2
4	1	0	26.900	84	0.900	2
5	1	0	26.900	86	5	2
6	1	0	26.600	87	1.800	2
7	1	0	25.300	90	4.300	2
8	1	0	24.400	93	0	3
9	1	0	25.200	90	0.300	2
10	1	0	24.400	95	0	2
11	1	0	25.500	85	0.100	3
12	1	0	25.600	88	1	2
13	1	0	25.500	87	0	2
14	1	0	24.200	92	0	2
15	1	0	24.800	88	0	2

ExampleSet (3,634 examples, 0 special attributes, 6 regular attributes)

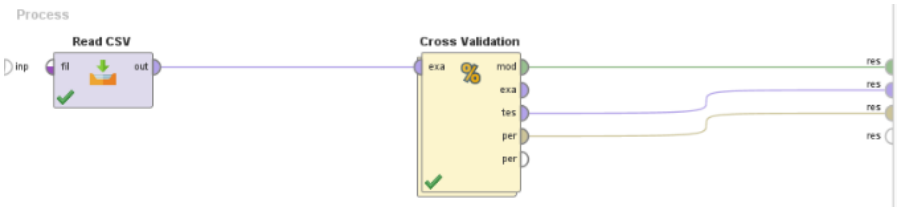
d. *Data Mining*

Data yang sudah ada dilanjutkan ke tahap pemodelan untuk klasifikasi memakai metode Naïve Bayes, kNN dan C4.5. pengerjaan dataset memakai *software Rapid miner*. Banyaknya data yang akan diolah berjumlah 3.634 dataset. Cara proses data pada *Rapid miner* untuk metode Naïve Bayes, kNN dan C4.5 menggunakan *file excel* dari data yang sudah diubah, perhatikan Gambar 8 dibawah ini. Lalu dilakukan *cross validation* untuk mengevaluasi data yang sudah ada. *K-Fold Cross Validation* adalah teknik validasi yang dipakai pada proses klasifikasi.

K-Fold Cross validation merupakan salah satu metode algoritma dengan mengelompokkan data dan membagi sampel data secara acak sebanyak nilai K k-fold. Lalu salah satu kelompok k-fold tersebut akan dijadikan sebagai data uji sedangkan sisa kelompok yang lain akan dijadikan sebagai data latih. Nilai k diambil 10 *fold* maka dari 3.634 data akan membentuk 10 *subset* data dengan ukuran sama yaitu diperkirakan 363,4 atau 364 data. Dari masing-masing 10 *subset* tersebut, 3.270 data membentuk data latih dan 364 data membentuk data uji.

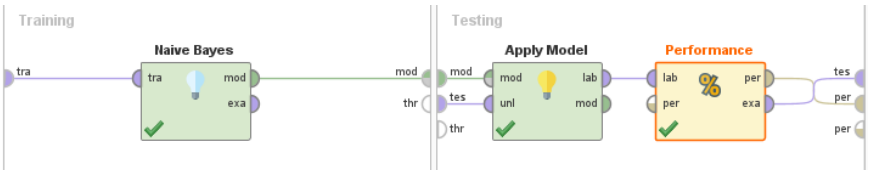
e. Hasil Perhitungan Klasifikasi dengan Rapid Miner

Proses pengambilan data dilakukan implementasi metode data *mining* selaras dengan yang dipakai pada penelitian ini adalah metode Naïve Bayes, kNN, dan C4.5 contoh gambar 6 dibawah ini.



Gambar 6. Proses pengumpulan data untuk metode Naïve bayes, kNN dan C4.5

Metode Naïve Bayes dipakai untuk data latih dan untuk teknik kategori bisa diperhatikan Gambar 7 dibawah ini.



Gambar 7. Model klasifikasi dengan algoritma naïve bayes

Hasil dari *confusion matrix* bias di perhatikan pada tabel 5 *confusion matrix* berikut.

Tabel 5. *Confution Matrix* Metode Naïve Bayes

accuracy: 81.15% +/- 1.18% (micro average: 81.15%)

	true TIDAK HUJAN	true HUJAN	class precision
pred. TIDAK HUJAN	2812	293	90.56%
pred. HUJAN	392	137	25.90%
class recall	87.77%	31.86%	

Keterangan dari tabel 5:

a. Jumlah data awal TIDAK HUJAN dan diperkirakan TIDAK HUJAN adalah 2812.
b. Jumlah data awal HUJAN dan diperkirakan HUJAN adalah 137.
c. Jumlah data awal TIDAK HUJAN dan diperkirakan HUJAN adalah 392.
d. Jumlah data awal HUJAN dan diperkirakan TIDAK HUJAN adalah 2812.

Evaluasi untuk model ini memakai nilai akurasi *error*.

Akurasi dari model yaitu :

$$Akurasi = \frac{2812 + 137}{2812 + 137 + 392 + 293} = \mathbf{81{,}15\%}$$

$$Error = \frac{392 + 293}{2812 + 137 + 392 + 293} = \mathbf{18{,}85\%}$$

Banyaknya dataset yang diperkirakan dengan valid oleh metode C4.5 tampilkan dalam tabel 8 *confusion matrix* ini

Tabel 6. *Confusion Matrix* Metode C4.5

accuracy: 88.03% +/- 0.30% (micro average: 88.03%)

	true TIDAK HUJAN	true HUJAN	class precision
pred. TIDAK HUJAN	3199	430	88.15%
pred. HUJAN	5	0	0.00%
class recall	99.84%	0.00%	

Penjelasan tabel 6:

a. Jumlah data awal TIDAK HUJAN dan diperkirakan TIDAK HUJAN adalah 3199.
b. Jumlah data awal HUJAN dan diperkirakan HUJAN adalah 0.

c. Jumlah data awal TIDAK HUJAN dan diperkirakan HUJAN adalah 5.
d. Jumlah data awal HUJAN dan diperkirakan TIDAK HUJAN adalah 430.

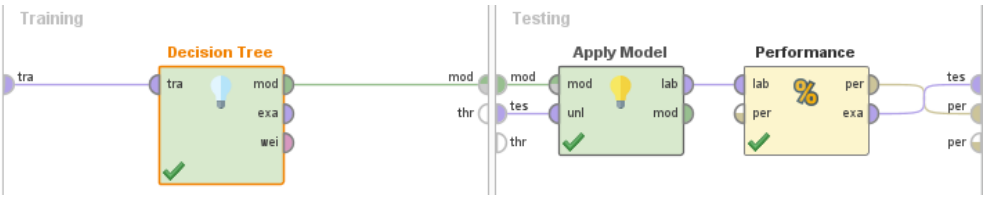
Evaluasi untuk model ini memakai nilai akurasi *error*.

Akurasi dari model yaitu :

$$Akurasi = \frac{3199 + 0}{3199 + 0 + 5 + 430} = \mathbf{88{,}03\%}$$

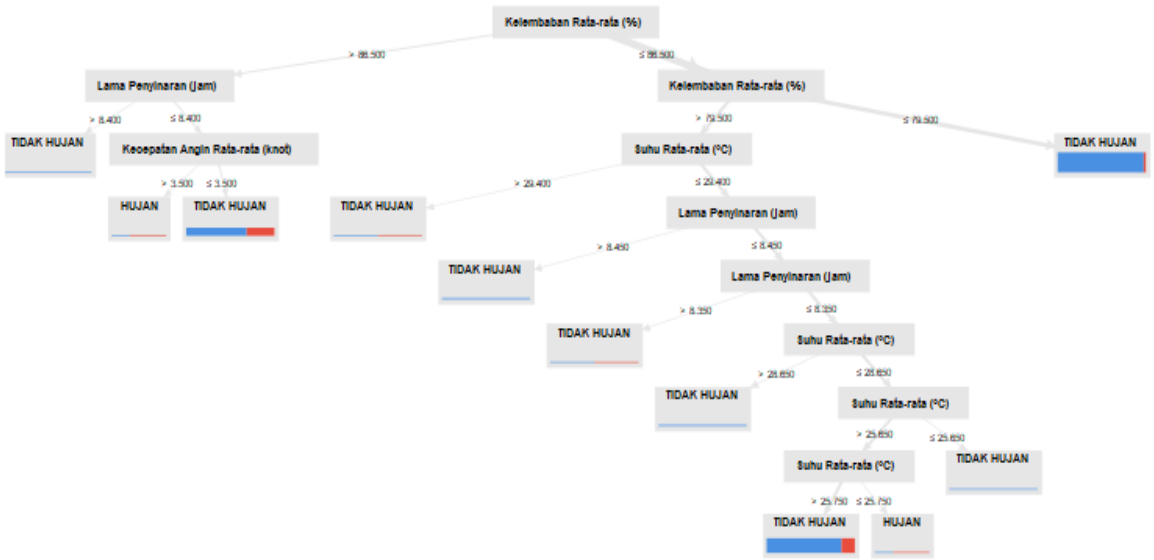
$$Error = \frac{5 + 430}{3199 + 0 + 5 + 430} = \mathbf{11{,}97\%}$$

Pada data latih dilakukan metode kNN untuk teknik kategori ditampilkan pada Gambar 8.



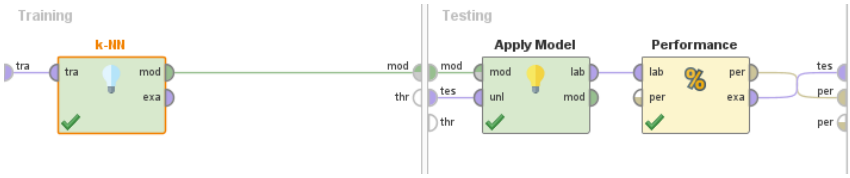
Gambar 8. Model klasifikasi dengan algoritma C4.5

Pada gambar 9 dibawah ini adalah pohon keputusan (*decision tree)* yang diperoleh dari algoritma C4.5. dari tiga variable yang dipakai terdapat empat variabel yang membangun pohon yaitu suhu, kelembaban, lama penyinaran, kecepatan angin. Indeks prestasi yang membentuk pohon lama penyinaran, kelembaban, suhu. Yang merupakan simpul akar ialah kelembaban karena mempunyai *gain* teratas. Bila kelembaban tinggi maka bisa diperkirakan curah hujan lebih banyak yang "HUJAN" sedangkan bila kelembaban kecil diperkirakan curah hujan lebih banyak "TIDAK HUJAN", tetapi bila nilai kelembaban sedang, maka lihat suhu, lama penyinaran, dan kecepatan angin. Jika besar maka diperkirakan curah hujan lebih banyak "TIDAK HUJAN" dan jika kecil maka diperkirakan curah hujan lebih banyak "HUJAN".



Gambar 9. *Decision tree* C4.5

Pada data latih dilakukan metode kNN untuk teknik kategori seperti pada Gambar 10 dibawah ini.



Gambar 10. Model klasifikasi dengan algoritma kNN

Banyaknya dataset yang diperkirakan dengan benar oleh metode kNN ditampilkan dalam *confusion matrix*. Perhatikan Tabel 7 dibawah.

Tabel 7. *Confusion Matrix* Metode kNN dengan k = 5

accuracy: 85.61% +/- 0.84% (micro average: 85.61%)

	true TIDAK HUJAN	true HUJAN	class precision
pred. TIDAK HUJAN	3059	378	89.00%
pred. HUJAN	145	52	26.40%
class recall	95.47%	12.09%	

Penjelasan dari tabel 7:

a. Jumlah data awal TIDAK HUJAN dan diperkirakan TIDAK HUJAN adalah 3059.
b. Jumlah data awal HUJAN dan diperkirakan HUJAN adalah 52.
c. Jumlah data awal TIDAK HUJAN dan diperkirakan HUJAN adalah 145.
d. Jumlah data awal HUJAN dan diperkirakan TIDAK HUJAN adalah 378.

Evaluasi untuk model ini memakai nilai akurasi *error*.

Akurasi dari model yaitu :

$$Akurasi = \frac{3059 + 52}{3059 + 52 + 145 + 378} = \mathbf{85{,}61\%}$$

$$Error = \frac{145 + 378}{3059 + 52 + 145 + 378} = \mathbf{14{,}39\ \%}$$

Hasil dari performa pada setiap model adalah *accuracy* dan *error* lalu akan dibandingkan untuk melihat metode mana yang lebih bagus untuk memperkirakan curah hujan yang terjadi. Lihat tabel 8 perbandingan setiap tabel berikut :

Tabel 8. Komparasi Nilai Performansi Setiap Algoritma

Algoritma	*Accuracy*	*Error*
C4.5	88,03 %	11,97%
kNN, k = 5	85,61%	14,39%
Naïve Bayes	81,15%	18,85%

4. KESIMPULAN

Hasil dari perbandingan dapat dilihat bahwa algoritma C4.5 memuat nilai yang teratas untuk semua kategori performa dibandingkan dengan metode yang lain. Untuk nilai akurasi yang tinggi dan untuk *error* adalah nilai terbawah. Kesimpulan dari penelitian ini adalah algoritma yang sudah diuji dapat dipakai untuk memperkirakan curah hujan, dapat dilihat dari nilai *accuracy* dan *error* dari semua metode ditemui dalam kategori "baik", "sedang" dan "cukup". Hasil dari pertimbangan didapat jawaban bahwa algoritma C4.5 yang sangat cocok untuk memperkirakan tingkat curah hujan yang diingikan karena memilliki nilai ketepatan yang paling baik dan *error* terendah dibandingkan dengan algoritma kNN dan Naïve Bayes. Untuk menyempurnakan nilai ketepatan dari metode dapat menambah variabel-variabel lain yang dapat mempengaruhi curah hujan. Bisa dicobakan dengan metode lain diluar dari metode yang sudah dipakai dalam penelitian ini.

REFERENSI

[1] M. L. Laia and Y. Setyawan, "Perbandingan hasil klasifikasi curah hujan menggunakan metode SVM dan NBC," *J. Stat. Ind. dan Komputasi*, vol. 5, no. 02, pp. 51–61, 2020.
[2] S. Widaningsih, "Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan

Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4, 5, Naïve Bayes, Knn Dan Svm," *J. Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019.

[3] S. M. Gorade, A. Deo, and P. Purohit, "A study of some data mining classification techniques," *Int. Res. J. Eng. Technol.*, vol. 4, 2017.

[4] A. B. Annasaheb and V. K. Verma, "Data mining classification techniques: A recent survey," *Int. J. Emerg. Technol. Eng. Res.*, vol. 4, no. 8, pp. 51–54, 2016.

[5] M. Suyatno, M. Jumintono, D. I. Pambudi, and A. Mardati, "Design of Values Education in School For Adolescents," in *2nd International Conference on Innovative Research Across Disciplines (ICIRAD 2017)*, 2017, pp. 6–9.

[6] F. Gorunescu, *Data Mining: Concepts, models and techniques*, vol. 12. Springer Science & Business Media, 2011.

[7] M. Al-Batah, B. Zaqaibeh, S. A. Alomari, and M. S. Alzboon, "Gene Microarray Cancer Classification using Correlation Based Feature Selection Algorithm and Rules Classifiers.," *Int. J. Online Biomed. Eng.*, vol. 15, no. 8, 2019.

[8] E. S. R. B. Situmorang, M. K. Anam, R. Rahmaddeni, and A. N. Ulfah, "Perbandingan Algoritma Svm Dan Nbc Dalam Analisa Sentimen Pilkada Pada Twitter," *CSRID (Computer Sci. Res. Its Dev. Journal)*, vol. 13, no. 3, pp. 169–179, 2021.

[9] R. Rahmaddeni, M. K. Anam, Y. Irawan, S. Susanti, and M. Jamaris, "Comparison of Support Vector Machine and XGBSVM in Analyzing Public Opinion on Covid-19 Vaccination," *Ilk. J. Ilm.*, vol. 14, no. 1, 2022.