E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 314 – 324

OPTIMASI KLASIFIKASI CURAH HUJAN MENGGUNAKAN SUPPORT VECTOR MACHINE (SVM) DAN RECURSIVE FEATURE ELIMINATION (RFE)

Arief Riski Indra Pratama[1)], Siti Amalia Latipah[2)], Betha Nurina Sari[3)]
[1, 2, 3)]Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang
Jl. HS.Ronggo Waluyo, Puseurjaya, Kec. Telukjambe Tim., Kabupaten Karawang, Jawa Barat, Indonesia, 41361
e-mail: arief.riski18090@student.unsika.ac.id[1)], siti.amalia18017@student.unsika.ac.id[2)], betha.nurina@staff.unsika.ac.id[3)]

ABSTRAK

Indonesia merupakan negara tropis yang mempunyai curah hujan tinggi. Curah hujan yang tinggi dapat mengakibatkan bencana alam berupa banjir. Untuk menangani permasalahan tersebut, perlu dilakukan prediksi cuaca yang akurat. Penelitian ini bertujuan untuk menyelesaikan masalah tersebut dengan mengklasifikasikan curah hujan dengan kategori hujan sedang dan hujan lebat menggunakan metode data mining CRISP-DM. Algoritma yang digunakan untuk klasifikasi curah hujan adalah SVM (Support Vector Machine) dengan metode optimasi seleksi fitur menggunakan RFE (Recursive Feature Elimination). Dengan penggunaan RFE sebagai metode seleksi fitur, atribut terbaik yang dihasilkan berjumlah 3 dari total 10 atribut, yaitu Tavg (temperatur rata-rata), ss (lamanya penyinaran matahari) dan Tn (temperatur minimum). Hasil evaluasi Confusion Matrix menggunakan SVM sebelum menerapkan RFE memiliki akurasi paling besar 77% yang ada pada skenario 3 (70:30), dan setelah menerapkan RFE akurasi paling besar meningkat 2% menjadi 79% yang terdapat pada skenario 2 (80:20) dan 3 (70:30). Hal ini menunjukkan penggunaan RFE pada SVM lebih unggul dan dapat meningkatkan akurasi klasifikasi curah hujan.
***Kata Kunci**: Crisp-DM, Curah Hujan, RFE, SVM*

ABSTRACT

Indonesia is a tropical country that has high rainfall. Heavy rainfall can result in natural disasters such as floods. To overcome these problems, it is necessary to make accurate weather predictions. This study aims to solve this problem by classifying rainfall into the category of moderate rain and heavy rain using the CRISP-DM data mining method. The algorithm used for rainfall classification is SVM (Support Vector Machine) with the RFE (Recursive Feature Elimination) feature selection optimization method. By using RFE as a feature selection method, 3 out of a total of 10 attributes obtained the best attributes, namely Tavg (average temperature), ss (length of sunlight) and Tn (minimum temperature). The results of the Confusion Matrix evaluation using SVM before applying the RFE had the greatest accuracy of 77% in scenario 3 (70:30), and after the application of the RFE the highest accuracy increased 2% to 79% which was found in scenario 2 (80:20).) and 3 (70:30). This shows that the use of RFE in SVM is superior and can increase the accuracy of rainfall.
***Keywords**: Crisp-DM, Rainfall Intensity, RFE, SVM*

I. PENDAHULUAN

Indonesia merupakan negara dengan iklim hutan hujan tropis dengan ciri-ciri suhu udara yang hampir seragam, serta memiliki curah hujan yang tinggi yang terbagi rata di sepanjang tahun. Salah satu daerah di Indonesia yang mempunyai tingkat curah hujan paling tinggi adalah Bogor atau dikenal dengan sebutan Kota Hujan. Hal ini disebabkan oleh intensitas curah dan hari hujannya mencapai 4000 hingga 4500 mm[1]. Intensitas curah hujan yang tinggi bisa menimbulkan bencana seperti banjir dan longsor.Perlu dilakukan peramalan atau prediksi guna memperkirakan seberapa besar curah hujan yang akan datang serta dapat mempersiapkan rencana untuk menanggulangi bencana tersebut.

Prediksi curah hujan merupakan salah satu tantangan besar dalam bidang meteorologi yang telah banyak dijadikan subjek penelitian. Pendekatan dalam prediksi curah hujan bisa dilakukan dengan metode empiris maupun dinamis. Prediksi curah hujan dengan jangka pendek dapat dilakukan melalui penerapan metode dinamis yang merupakan suatu pendekatan analitis yang didasari oleh prinsip-prinsip dinamika fluida, sedangkan metode empiris merupakan pendekatan yang dilakukan secara statistik dan matematis menggunakan data historis, seringnya digunakan untuk prediksi cuaca jangka panjang[2].

Data historis yang diolah menjadi sebuah pengetahuan dapat bermanfaat bagi banyak orang.Mengubah datanya menjadi pengetahuan, manusia dapat melakukan prediksi maupun estimasi tentang kejadian apa yang terjadi ke depan. Maka dari itu, perlu adanya proses yang menggunakan data mining untuk mengolah pengetahuan atau menemukan pola dari suatu data yang besar yang dalam hal ini adalah prediksi curah hujan[3].

Data mining merupakan teknik penggabungan metode metode analisis data secara statistik menggunakan algoritma-algoritma guna memproses data berukuran besar[4]. Data Mining mempunyai 5 peran utama menurut[5],

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 314 – 324

yaitu estimasi, klasifikasi, prediksi, clustering, dan asosiasi. Klasifikasi merupakan teknik yang paling umum dalam data mining. Tujuan klasifikasi yaitu guna menganalisis data historis yang disimpan dalam dataset dan secara otomatis menghasilkan suatu model yang bisa memprediksi keadaan di masa mendatang[6]. Klasifikasi memiliki beberapa algoritma di antaranya adalah Naïve Bayes Classifier (NBC), Support Vector Machine (SVM), Artificial Neural Network, Classification Tree, K-Nearest Neighbor, Analisis Diskriminan, dan lain-lain[7].

Beberapa penelitian data mining telah dilakukan untuk prediksi curah hujan diantaranya adalah dengan menggunakan Iterative Dichotomiser 3, diperoleh hasil dari 179 data (10% dari dataset) yang diperoleh secara acak, diketahui bahwa terdapat 132 data yang tepat prediksi dengan data cuaca sesungguhnya. Maka bisa disimpulkan akurasinya sebesar 73,74%[5]. Penelitian juga dilakukan oleh[3] dalam menerapkan data mining untuk memprediksi curah hujan dengan menggunakan Algoritma CART. Hasil dari model evaluasi confusion metriksnya menunjukkan akurasi sebesar 89.4%. Penelitian lainnya yang ditulis[8], menerapkan data mining dengan algoritma KNN sebagai cara untuk memprediksi potensi hujan harian. Hasil nya menunjukkan bahwa prediksi penentuan cuaca harian dengan algoritma K-Nearest Neighbor mendapatkan nilai RMSE 9.899 +/- 0.000. Untuk itu penelitian ini akan menggunakan algoritma SVM sebagai cara untuk memprediksi curah hujan.

Algoritma SVM (Support Vector Machine) merupakan salah satu metode yang dilandasi oleh teori pembelajaran statistik dan memberi hasil yang memuaskan dibanding dengan metode lain[9]. Kelebihan dari algoritma ini adalah kemampuannya untuk mengimplementasikan pemisahan yang linear dalam input data non linear berdimensi tinggi, hal tersebut diperoleh dengan menjalankan fungsi kernel yang diperlukan. Efektivitas SVM sangat dipengaruhi dengan jenis fungsi kernel yang ditentukan dan diterapkan berdasarkan karakteristik data[10]. Kekurangan dari SVM yaitu tidak dapat menghasilkan prediksi yang akurat ketika mempunyai banyak fitur yang tidak relevan, tidak semua fitur digunakan dalam proses pembuatan model, hal ini bisa diatasi dengan metode seleksi fitur.

Seleksi fitur bekerja dengan cara mengurangi jumlah fitur kemudian memilih fitur yang benar-benar memberikan manfaat, jumlah fitur yang berkurang dapat mengatasi masalah yang berupa overfitting[11]. Dalam dataset prediksi curah hujan yang diperoleh dari BMKG terdapat 10 fitur, seleksi fitur dapat digunakan untuk mengekstrak atribut penting yang secara efektif berdampak pada akurasi klasifikasi. Atribut yang tidak relevan atau berlebihan akan dihilangkan sehingga meningkatkan kinerja model SVM.

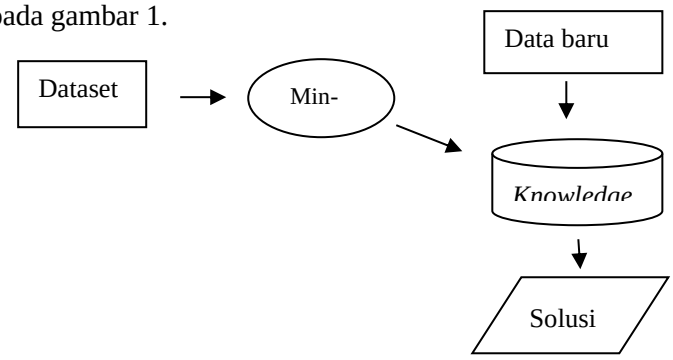
Terdapat banyak metode dalam seleksi fitur, salah satunya adalah RFE (Recursive Feature Elimination). RFE dikenal dengan metode teknik wrapper-based yang mengecek korelasi tidak menggunakan score. Proses ini mengurutkan fitur berdasarkan tingkatan pentingnya terhadap proses prediksi. Pada setiap iterasi, rangking fitur yang penting diukur dan fitur yang kurang relevan dihilangkan. Ranking tersebut dapat dihitung menggunakan metode Support Vector Machine (SVM) kernel linear[12].

Berdasarkan uraian di atas peneliti akan mencoba untuk menggunakan algoritma SVM dengan menggunakan metode RFE sebagai seleksi fitur untuk melihat hasil akurasi prediksi curah hujan, dengan tujuan dari penggunaan RFE ini dapat meningkatkan akurasi algoritma SVM dalam mencari akurasi prediksi curah hujan. Hasil dari penelitian ini akan dibuat perbandingan akurasi antara algoritma SVM setelah dioptimasi RFE dengan algoritma SVM sebelum optimasi RFE.

II. METODE PENELITIAN.

A. Data Mining

Data mining merupakan disiplin ilmu yang mempunyai padanan dengan knowledge discovery atau penemuan pengetahuan dan pattern recognition atau pengenalan pola pengetahuan yang masih tersembunyi di dalam bongkahan data. Dimana keduanya memiliki tujuan yang sama yaitu untuk menemukan, menggali pengetahuan dari informasi atau data yang kita miliki yang dikenal di zaman sekarang dengan data mining[4]. Alur kerja *data mining* secara umum dapat dilihat pada gambar 1.

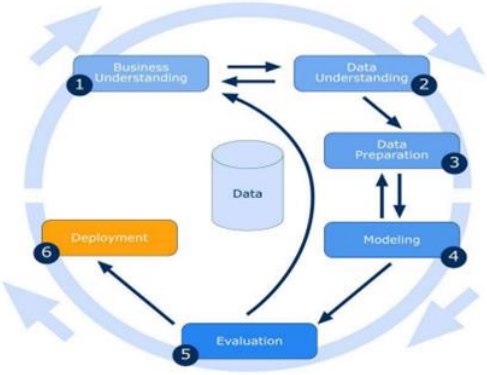


E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 314 – 324

Gambar 1. Alur data mining

Dalam penelitian ini, metodologi *data mining* yang dipakai adalah CRISP-DM sebagai solusi permasalahan umum dalam bisnis dan penelitian. Metodologi CRISP-DM mempunyai enam tahapan yaitu Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, dan Deployment. Proses tahapan ini dapat dilihat pada gambar 2.



Gambar 2. Metodologi CRISP-DM

B. *Support Vector Machine*

Support Vector Machine (SVM) pada mulanya dipelopri oleh Vapnik di tahun 1992 bersama temannya Bernhard Boser dan Isabelle Guyon. SVM adalah algoritma yang bekerja memanfaatkan pemetaan non linier guna mengganti data training asli ke dimensi yang tinggi. Pada hal ini dimensi yang baru hendak mencari hyperplane untuk membagi secara linier dan dengan pemetaan non linier yang tepat ke dimensi yang lebih tinggi, data dari dua kelas selalu dapat dibagi dengan hyperplane tersebut. SVM menemukan ini memanfaatkan support vector dan margin[13].

Dalam penggunaan teknik SVM, cara ini mencoba untuk menemukan fungsi pemisah (classifier) yang optimal untuk memisahkan dua kelas yang berbeda. Teknik ini berusaha menemukan fungsi pembagi (hyperplane) terbaik antara fungsi yang tidak dibatasi jumlahnya untuk membagi dua jenis objek. *Hyperplane* yang baik adalah hyperplane yang letaknya berada di tengah antara dua set objek dari dua kelas. Pendefinisian persamaan pada *hyperplane* pemisah dapat dituliskan pada persamaan 1.

$$W.X + b = 0 \quad (1)$$

Dimana W adalah bobot suatu vektor, yaitu $W = \{w_1, w_2, w_3, \ldots, w_n\}$; dimana n merupakan jumlah atribut dan b merupakan nilai skalar atau sering disebut bias. Jika berdasarkan dua nilai masukan dari atribut A_1 dan A_2 dengan tupel latih $X = (x_1, x_2)$, dimana nilai x_1 dan x_2 adalah nilai atribut A_1 dan A_2, dan jika atribut b dianggap bobot tambahan w_0, maka *hyperplane* pemisah menjadi seperti yang ada pada persamaan 2.

$$w_0 + w_1x_1 + w_2x_2 = 0 \quad (2)$$

Dengan demikian, nilai titik yang ada diatas *hyperplane* pemisah akan memenuhi persamaan 3

$$w_0 + w_1x_1 + w_2x_2 > 0 \quad (3)$$

Sebaliknya, titik yang berada dibawah *hyperplane* pemisah akan memenuhi persamaan 3

$$w_0 + w_1x_1 + w_2x_2 < 0 \quad (4)$$

Dilihat dari persamaan 1,2,3 dan 4, bobot dapat disesuaikan menjadi *hyperplane* yang mendefinisikan sisi margin seperti persamaan 5

$$h_1 : w_0 + w_1x_1 + w_2x_2 \geq 0, \text{ untuk } y_i = +1$$
$$h_2 : w_0 + w_1x_1 + w_2x_2 \leq 0, \text{ untuk } y_i = -1 \quad (5)$$

Untuk mendapatkan nilai *Maximum Margin Hyperplane* SVM bisa dengan menggunakan trik matematika yaitu *Lagrangian formulation* lalu solusinya dapat diselesaikan dengan kondisi *Karush-Kuhn-Tucker*. Berdasarkan *Langrangian formulation, Maximum Margin Hyperplane* dapat ditulis kembali menjadi suatu batas keputusan (*decision boundary*) seperti persamaan 6.

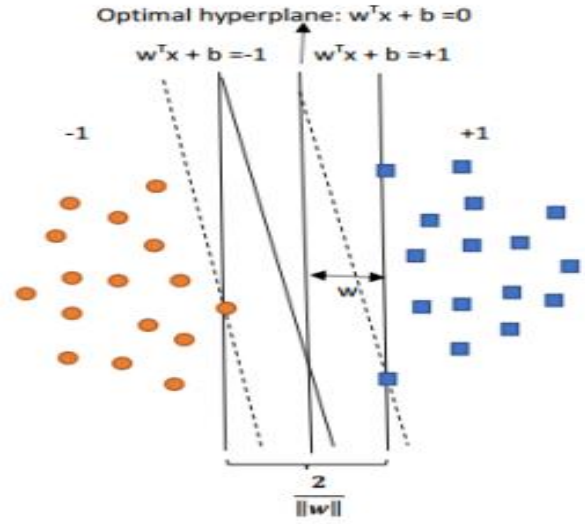
$$d(x^T) = \left(\sum_{i=1}^{l} y_i\, a_i\, X_i\, X^T + b_0\right) \quad (6)$$

Dimana y_i merupakan label dari kelas *support vector* X_i, X^T adalah tupel tes. Notasi a_i dan b_0 merupakan parameter numerik yang otomatis ditentukan oleh optimalisasi algortima SVM dan 1 adalah jumlah *support vector*.

Gambar 3, menunjukkan bagaimana SVM mendapatkan hyperplane terbaik sebanding dengan mengoptimalkan margin atau jarak dua set dari kelas yang berlainan.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 314 – 324



Gambar 3.SVM mencoba menemukan *hyperplane* terbaik yang memisahkan *class* -1 dan *class* +1

C. Recursive Feature elimination (RFE)

Recursive feature elimination (RFE) merupakan metode seleksi fitur yang bekerja dengan cara mengurangi fitur yang tidak saling terkait atau lemah dan dilakukan terus menerus secara rekursif hingga fitur terbaik didapatkan untuk digunakan dalam membangun model. Fitur yang tersisa tersebut akan digunakan untuk membangun model untuk dihitung nilai akurasinya. Fitur diperingkatkan secara relatif sesuai urutan eliminasi. Tingkatan fitur yang dipilih akan membuat model cocok dengan semua atribut.

Atribut yang memiliki nilai p-value tertinggi akan dipilih dan jika p-value lebih besar dari tingkatan signifikansi maka akan ditolak. Sekali lagi, model ini dibangun dengan nilai atribut atau fitur yang tersisa. Penghapusan fitur atau atribut ini akan dilakukan secara berulang kali sampai mendapatkan akurasi model yang baik. Pemeringkatan dilakukan berdasarkan nilai koefisien dari atribut. Semakin tinggi nilai koefisien makan semakin baik peringkatnya dan semakin besar kemungkinan untuk dipilih. RFE mampu untuk mengombinasikan atribut yang berkontribusi dalam prediksi variabel atau kelas target. Dalam penelitian ini nilia setiap fitur atau atribut dihitung menggunakan RFE kemudian diurutkan berdasarkan tingkat kepentingannya. Hal ini dilakukan untuk mengeksplorasi dan ingin mengetahui fitur mana saja yang paling menonjol dan mendominasi[16].

III. HASIL DAN PEMBAHASAN

A. Business Understanding

Pada tahap ini dilakukan pemahaman dari data curah hujan yang berkaitan dengan pola intensitas curah hujan untuk mengetahui kapan daerah tersebut berpotensi banjir. Serta untuk mengetahui atribut apa saja yang dapat memengaruhi intensitas curah hujan.

B. Data Understanding

Penelitian ini mengambil sebagian data yang ada di website BMKG Stasiun Meteorologi Citeko, Kabupaten Bogor. Data yang diambil merupakan data curah hujan dari periode 2017-2020 berjumlah total 1461 *record* yang bisa dilihat pada tabel 1. Parameter awal yang terdapat pada dataset ada 11 untuk diteliti, yaitu Tanggal, Arah angin saat kecepatan maksimum (ddd_x), Arah angin terbanyak (ddd_car), Kecepatan angin maksimum (ff_x), Kecepatan angin rata-rata (ff_avg), Kelembaban rata-rata (RH_avg), Lamanya penyinaran matahari (ss), Temperatur maksimum (Tx), Temperatur minimum (Tn), Temperatur rata-rata (Tavg) dan 1 parameter target yaitu Curah hujan (RR). Validasi untuk mengklasifikasi curah hujan yang berpotensi banjir dengan mengkategorikan parameter Curah hujan (RR) menjadi 2 kategori, yaitu hujan sedang dan hujan lebat berdasarkan nilai ambang (threshold) dari BMKG sesuai tabel 2

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 314 – 324

TABEL I
DATA SET BMKG

Tanggal	Tn	Tx	Tavg	RH_av	ddd_car	ss	ff_x	ddd_x	ff_avg	RR
01-01-2017	19.0	27.2	22.2	78.0	N	3.4	2.0	340.0	1.0	14.7
02-01-2017	19.0	26.1	21.6	82.0	N	6.3	3.0	270.0	1.0	2.1
03-01-2017	20.0	26.9	23.4	76.0	N	4.8	3.0	300.0	1.0	2.3
04-01-2017	18.0	24.6	20.9	87.0	N	2.2	1.0	120.0	1.0	14.0
					…					
30-12-2020	19.2	23.0	20.9	86.0	C	0.0	2.0	300.0	0.0	17.0
31-12-2020	19.2	23.3	20.8	88.0	C	0.0	2.0	290.0	0.0	7.8

TABEL II
KATEGORI HUJAN BMKG

No.	Kategori Hujan	Intensitas Curah Hujan (mm/hari)
1.	Berawan	0
2.	Hujan ringan	0.5 – 20
3.	Hujan sedang	20 -50
4.	Hujan lebat	50-100
5.	Hujan sangat lebat	100-150
6.	Hujan ekstrem	>150

C. Data Preparation

Pada data preparation ini cleaning dataset dilakukan dengan menghapus kolom tanggal dan ddd_car karena tidak digubakan dalam dataset. Selanjutnya dilakukan penanganan berupa Missing Value dengan melakukan imputation menggunakan nilai rata-rata, terdapat hanya 169 record label yang dikotomi dengan intensitas curah hujan sedang dan lebat dari total 1461 record. Dalam 169 record tersebut. Ditemukan bahwa terdapat imbalance dataset, yaitu perbedaan jumlah data yang terlalu jauh pada suatu label. Jumlah label yang masuk ke dalam kategori hujan sedang berjumlah 166 record dan 3 record untuk kategori hujan lebat. Karena itu perlu dilakukannya penanganan *imbalance dataset* dengan *resampling*. *Resampling* merupakan teknik dimana mencoba melakukan peneyimbangan data asli melalu proses algoritma *sampling* dengan menyesuaikan jumlah sampel dalam kelas berbeda.Pendekatan *resampling* terbagi menjadi 3, yaitu *random over-sampling*, *random under-sampling dan random over-under sampling*. *Random oversampling* melakukan duplikasi sampel pada kelas minoritas, *random undersampling* melakukan duplikasi sampel pada kelas mayoritas dan *random over-undersampling* ini melakukan duplikasi sampel kecil pada kelas mayoritas dan minoritas.

Sebagai contoh misalnya, jika kita memiliki dataset dengan distribusi kelas 1:100, pertama-tama kita mungkin menerapkan oversampling untuk meningkatkan rasio menjadi 1:10 dengan menduplikasi contoh dari kelas minoritas, kemudian menerapkan undersampling untuk lebih meningkatkan rasio ke 1:2 dengan menghapus contoh dari kelas mayoritas. Hasil *random over-undersampling* bisa dilihat pada tabel 3.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 314 – 324

TABEL III
HASIL RANDOM OVER AND UNDERSAMPLING

Tn	Tx	Tavg	RH_avg	ss	ff_dx	ddd_x	ff_avg	kate-gori
19.9	23.6	20.9	92.0	6.3	2.0	320.0	0.0	Lebat
17.0	26,0	21.0	83.0	5.1	3.0	280.0	1.0	Lebat
19.3	24.2	21.0	92.0	0.3	4.0	320.0	0.0	Lebat
19.9	23,6	20.9	92.0	6.3	2.0	320.0	0.0	Lebat
17.0	26.0	21.0	83.0	5.1	3.0	280.0	1.0	Lebat
				…….				
18.0	24.0	21.3	91.0	3.0	4.0	160.0	0.0	Sedang
20.0	24.2	21.6	97.0	2.8	2.0	310.0	0.0	Sedang
18.7	27.6	22.2	81.0	2.6	2.0	80.0	0.0	Sedang
19.0	25.6	21.0	88.0	0.0	3.0	160.0	1.0	Sedang
18.0	21.4	19.2	96.0	0.0	2.0	350.0	0.0	Sedang

Setelah menangani *imbalance dataset* tahap selanjutnya adalah melakukan normalisasi data. Hal ini perlu dilakukan agar rentan nilai antar atribut tidak berbeda jauh. Normalisasi menggunakan *min-max sclaer*, dimana rumus ini mempunya rumus dalam bentuk persamaan 7.

$$X\ scaled = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (7)$$

Dimana X merupakan nilai data perkolom, nilai Xmin adalah nilai minimum dari seluruh nilai X dan nilai Xmax merupakan nilai maksimum dari seluruh nilai X. Hasil normalisasi menggunakan *min-max scaler* data bisa dilihat pada tabel 4. Selanjutnya adalah melakukan transformasi data, dengan merubah kategori hujan sedang dan hujan lebat menjadi angka 0 dan 1.

TABEL IV
DATA SETELAH NORMALISASI

Tn	Tx	Tavg	RH_avg	ss	ff_dx	ddd_x	ff_avg
0.975	0.395062	0.348837	0.708333	0.954545	0.333333	0.888889	0.0
0.250	0.691358	0.372093	0.333333	0.772727	0.500000	0.777778	0.5
0.825	0.469136	0.372093	0.708333	0.045455	0.666667	0.888889	0.0
0.825	0.469136	0.372093	0.708333	0.045455	0.666667	0.888889	0.0
0.975	0.395062	0.348837	0.708333	0.954545	0.333333	0.888889	0.0

D. Modelling

Setelah selesai melakukan data preparation tahap selanjutnya adalah melakukan percobaan dengan menggunakan algoritma SVM dengan fungsi kernel linear tanpa melakukan seleksi fitur RFE. Pemodelan dibentuk dengan

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 314 – 324

membagi data training dan data testing menjadi sebanyak 4 skenario yaitu (90%:10%), (80%:20%), (70%:30%) dan (60%:40%). Dilakukannya pembagian ini untuk menguji masing-masing metode terhadap beberapa skenario dengan tujuan untuk mengetahui bagaimana tiap-tiap metode menghadapi perubahan dataset. Hasil perhitungan dari klasifikasi menghasilkan nilai akurasi, nilai presisi dan nilai recall. Skenario pembagian data training dan data testing dapat dlihat pada tabel 5.

TABEL V
PEMBAGIAN DATA TRAINING DAN TESTING

Skenario	*Data Training*	*Data Testing*
1	90%	10%
2	80%	20%
3	70%	30%
4	60%	40%

Langkah selanjutnya adalah melakukan pengujian terhadap tiap skenario sesuai dengan tabel 5 dengan menggunakan SVM kernel linear tanpa RFE. Skenario 1 (90:10) : hasil dari pengujian SVM pada model klasifikasi curah hujan tanpa RFE bisa dilihat pada gambar 4.

	precision	recall	f1-score	support
0	0.70	0.78	0.74	9
1	0.78	0.70	0.74	10
accuracy			0.74	19
macro avg	0.74	0.74	0.74	19
weighted avg	0.74	0.74	0.74	19

Gambar 4.Hasil pengujian SVM kernel linear tanpa RFE dengan skenario 1

Skenario 2 (80:20) : hasil dari pengujian SVM pada model klasifikasi curah hujan tanpa RFE bisa dilihat pada gambar 5.

	precision	recall	f1-score	support
0	0.70	0.78	0.74	9
1	0.78	0.70	0.74	10
accuracy			0.74	19
macro avg	0.74	0.74	0.74	19
weighted avg	0.74	0.74	0.74	19

Gambar 5.Hasil pengujian SVM kernel linear tanpa RFE dengan skenario 2

Skenario 3 (70:30) : hasil dari pengujian SVM pada model klasifikasi curah hujan tanpa RFE bisa dilihat pada gambar 6.

	precision	recall	f1-score	support
0	0.78	0.81	0.79	31
1	0.75	0.72	0.73	25
accuracy			0.77	56
macro avg	0.77	0.76	0.76	56
weighted avg	0.77	0.77	0.77	56

Gambar 6.Hasil pengujian SVM kernel linear tanpa RFE dengan skenario 3

Skenario 4 (60:40) : hasil dari pengujian SVM pada model klasifikasi curah hujan tanpa RFE bisa dilihat pada gambar 7.

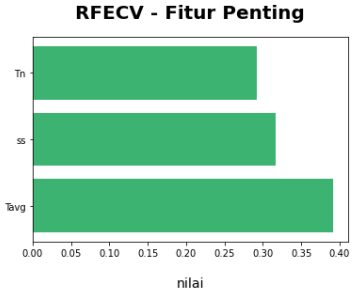
	precision	recall	f1-score	support
0	0.59	0.82	0.69	40
1	0.63	0.34	0.44	35
accuracy			0.60	75
macro avg	0.61	0.58	0.57	75
weighted avg	0.61	0.60	0.57	75

Gambar 7.Hasil pengujian SVM kernel linear tanpa RFE dengan skenario 4.

Langkah selanjutnya adalah menggunakan metode RFE sebagai seleksi fitur untuk selanjutnya diolah menggunakan algoritma SVM. Penggunaan seleksi fitur RFE menghasilkan 3 fitur yang penting, yaitu Tavg (temperatur rata-rata), ss (lamanya penyinaran matahari dan Tn (temperatur minimum).Hasil seleksi fitur ini bisa dilihat pada gambar 8.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 314 – 324



Gambar8. Fitur yang sudah di seleksi menggunakan RFE

Proses selanjutnya adalah melakukan percobaan kembali menggunakan algoritma SVM dengan atribut yang sudah diseleksi menggunakan RFE. Pemodelan kembali ini dilakukan dengan membagi data training dan data testing menjadi 4 skenario yaitu (90%:10%), (80%:20%), (70%:30%) dan (60%:40%). Skenario pembagian data training dan data testingdapat dlihat pada tabel 5.

Tahap selanjutnya adalah melakukan pengujian terhadap tiap skenario sesuai dengan tabel 5 dengan SVM kernel linear setelah menggunakan RFE. Skenario 1 (90:10) : hasil dari pengujian SVM pada model klasifikasi curah hujan setelah RFE bisa dilihat pada gambar 9.

	precision	recall	f1-score	support
0	0.70	0.78	0.74	9
1	0.78	0.70	0.74	10
accuracy			0.74	19
macro avg	0.74	0.74	0.74	19
weighted avg	0.74	0.74	0.74	19

Gambar 9.Hasil pengujian SVM kernel linear dan RFE dengan skenario 1

Skenario 2 (80:20) : hasil dari pengujian SVM pada model klasifikasi curah hujan setelah RFE bisa dilihat pada gambar 10.

	precision	recall	f1-score	support
0	0.73	0.89	0.80	18
1	0.88	0.70	0.78	20
accuracy			0.79	38
macro avg	0.80	0.79	0.79	38
weighted avg	0.81	0.79	0.79	38

Gambar 10.Hasil pengujian SVM kernel linear dan RFE dengan skenario 2

Skenario 3 (70:30) : hasil dari pengujian SVM pada model klasifikasi curah hujan setelah RFE bisa dilihat pada gambar 11.

	precision	recall	f1-score	support
0	0.79	0.84	0.81	31
1	0.78	0.72	0.75	25
accuracy			0.79	56
macro avg	0.79	0.78	0.78	56
weighted avg	0.79	0.79	0.78	56

Gambar 11.Hasil pengujian SVM kernel linear dan RFE dengan skenario 3

Skenario 4 (60:40) : hasil dari pengujian SVM pada model klasifikasi curah hujan setelah RFE bisa dilihat pada gambar 12.

	precision	recall	f1-score	support
0	0.58	0.80	0.67	40
1	0.60	0.34	0.44	35
accuracy			0.59	75
macro avg	0.59	0.57	0.56	75
weighted avg	0.59	0.59	0.56	75

Gambar 12.Hasil pengujian SVM kernel linear dan RFE dengan skenario 4

E. Evaluation

Hasil yang didapatkan dari seluruh percobaan dengan model SVM akan dibandingkan untuk mengetahui model mana yang terbaik dalam proses klasifikasi. Hasil perbandingan evaluasi dengan 4 skenario bisa dilihat pada tabel 6.

E-ISSN : 2540 - 8984

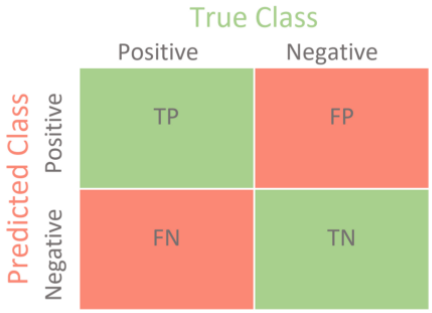
JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 314 – 324

TABEL VI
PERBANGDINGAN EVALUASI MODEL

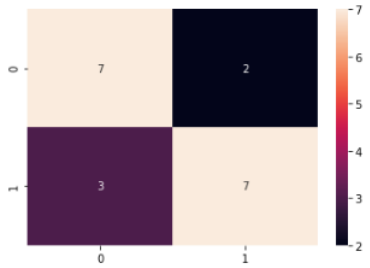
Skenario	Akurasi %		Presisi %			
	Non-RFE	RFE	Non-RFE		RFE	
			0	1	0	1
1	74	74	70	78	70	78
2	74	79	70	78	73	88
3	77	79	78	75	79	78
4	60	59	59	63	58	60

	Recall%				F-Score%			
	Non-RFE		RFE		Non-RFE		RFE	
	0	1	0	1	0	1	0	1
1	78	70	78	70	74	74	74	74
2	78	70	89	70	74	74	80	78
3	81	72	84	72	79	73	81	75
4	82	34	80	34	69	44	67	44

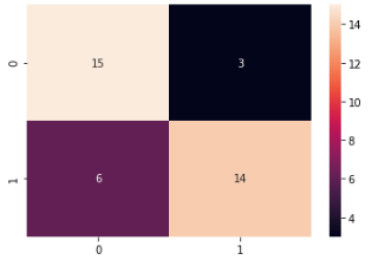
Berdasarkan nilai tabel 6, dapat dilihat bahwa hasil akurasi algoritma SVM dengan menerapkan RFE mendapat nilai akurasi terbesar pada skenario 2(80:20) dan skenario 3(70:30). Hasil skenario 2 dan 3 setelah menggunakan RFE cukup jauh lebih besar dibanding dengan skenario 4 setelah menggunakan RFE disebabkan perbedaan hasil pembagian rasio dataset, hal ini dapat dilihat juga pada skenario 4 sebelum menggunakan RFE juga mendapatkan akurasi yang kecil diantara skenario lainnya. Hal ini menunjukkan semakin besar *data testing* dapat mengurangi akurasi. Namun hasil akurasi saja belum cukup untuk menentukan model yang terbaik dalam penelitian ini. Dapat juga dilihat hasil prediksi berupa confusion matrix yang memiliki susunan seperti gambar 13 dan hasil confusion matrixnya pada gambar 14 sampai dengan gambar 21.



Gambar 13.Bentuk susunan confusion matrix



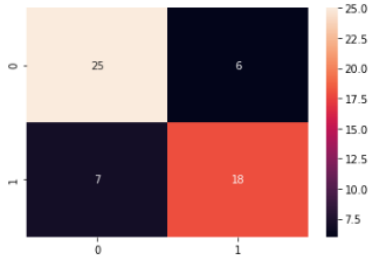
Gambar 14.Confusion matrix SVM tanpa RFE skenario 1.



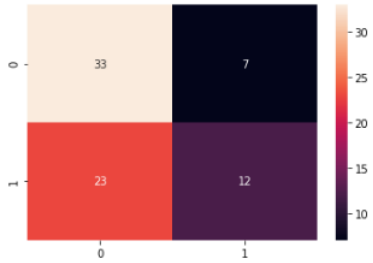
Gambar 15.Confusion matrix SVM tanpa RFE skenario 2.

E-ISSN : 2540 - 8984

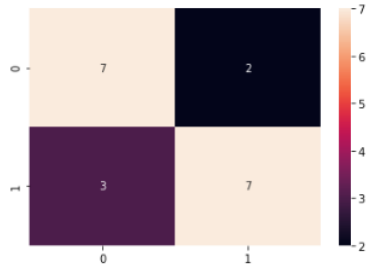
JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 314 – 324



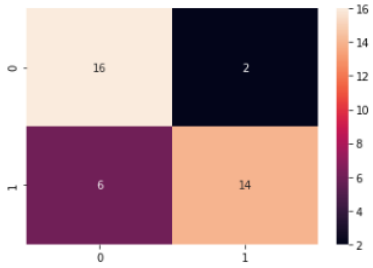
Gambar 16.Confusion matrix SVM tanpa RFE skenario 3.



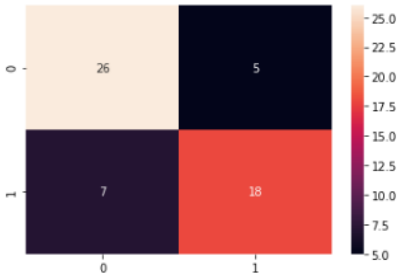
Gambar 17.Confusion matrix SVM tanpa RFE skenario 4.



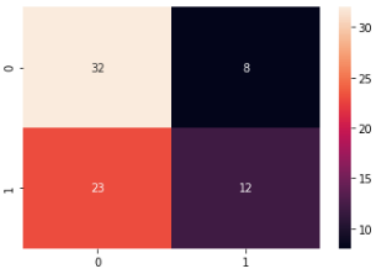
Gambar 18.Confusion matrix SVM Setelah RFE skenario 1.



Gambar 19.Confusion matrix SVM Setelah RFE skenario 2.



Gambar 20.Confusion matrix SVM Setelah RFE skenario 3.



Gambar 21.Confusion matrix SVM Setelah RFE skenario 4.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 07, Nomor 02, Juni 2022 : 314 – 324

F. Deployment

Setelah selesai tahap evaluasi dimana menilai dengan detail hasil dari model yang telah dibuat, maka selanjutnya dilakukan pengimplementasian dari keseluruhan model yang sudah dibuat. Selain itu dapat dilakukan penyesuaian model sehingga hasil suatu model dapat sesuai dengan target awal pada tahap CRISP-DM ini.

IV. KESIMPULAN

Hasil dari penelitian ini telah mendapatkan beberapa kesimpulan, yaitu penerapan algoritma SVM dengan kernel linear dan bantuan seleksi fitur RFE serta pemodelan dengan 4 skenario untuk klasifikasi curah hujan telah dilakukan untuk memperoleh model terbaik. Hal ini bisa dilihat pada skenario 2(20:80) dan 3(30:70) setelah dilakukan RFE mendapatkan akurasi paling besar yaitu 79%. Sedangkan akurasi terbesar jika tidak menggunakan RFE hanya 77% yang bisa dilihat pada skenario 3(30:70), hal ini menunjukkan bahwa optimasi SVM dengan RFE dalam klasifikasi curah hujan dapat meningkatkan akurasi sebesar 2%. Untuk penelitian selanjutnya agar dapat meningkatkan hasil akurasi klasifikasi curah hujan dengan menggunakan metode seleksi fitur lain seperti *Wrapper based methods* dan *Embedded based methods*.

DAFTAR PUSTAKA

[1] R. Hidayat and A. W. Farihah, "Identifikasi perubahan suhu udara dan curah hujan di Bogor," *J. Pengelolaan Sumberd. Alam dan Lingkung. (Journal Nat. Resour. Environ. Manag.*, vol. 10, no. 4, pp. 616–626, 2020, doi: 10.29244/jpsl.10.4.616-626.
[2] R. Prasetya, "Penerapan Teknik Data Mining Dengan Algoritma," vol. 2, no. 2, 2020.
[3] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," vol. 5, no. 2, 2021.
[4] S. Susanto, *Pengantar Data Mining*. Yogyakarta: Andi Publisher, 2010.
[5] B. P. T.P and R. D. Indah Sari, "Penerapan Data Mining Untuk Prakiraan Cuaca Di Kota Malang Menggunakan Algoritma Iterative Dichotomiser Tree (Id3)," *Jouticla*, vol. 2, no. 2, pp. 101–108, 2017, doi: 10.30736/jti.v2i2.68.
[6] A. V. D. Sano, "CARA KERJA DATA MINING – SERI DATA MINING FOR BUSINESS INTELLIGENCE (3)," 2019. https://binus.ac.id/malang/2019/01/cara-kerja-data-mining-seri-data-mining-for-business-intelligence-3/ (accessed Nov. 24, 2021).
[7] M. L. Laia and Y. Setyawan, "Perbandingan Hasil Klasifikasi Curah Hujan Menggunakan Metode SVM dan NBC," vol. 05, no. 2, pp. 51–61, 2020.
[8] H. Rofiq, K. C. Pelangi, and Y. Lasena, "PENERAPAN DATA MINING UNTUK MENENTUKAN POTENSI HUJAN HARIAN DENGAN MENGGUNAKAN ALGORITMA K NEAREST NEIGHBOR (KNN)," *J. Manaj. Inform. dan Sist. Inf.*, vol. 3, no. 1, pp. 8–15, 2020, [Online]. Available: http://mahasiswa.dinus.ac.id/docs/skripsi/jurnal/19417.pdf.
[9] P. Eko, *Data mining : konsep dan aplikasi menggunakan MATLAB*, 1st ed. CV Andi Offset, 2012.
[10] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.
[11] A. D. Achmad, "Metode Moving Average Dan Metode Support Vector Machine Dan Untuk Prediksi Variabel Meteorologi," vol. 4, no. 1, pp. 45–50, 2017.
[12] F. Zhang, H. L. Kaufman, Y. Deng, and R. Drabier, "Recursive SVM biomarker selection for early detection of breast cancer in peripheral blood," *BMC Med. Genomics*, vol. 6, no. S1, p. S4, Jan. 2013, doi: 10.1186/1755-8794-6-S1-S4.
[13] P. P. Widodo, *Penerapan data mining dengan Matlab*. Bandung: Rekayasa Sains, 2013.
[14] B. Santosa, *Data mining : Teknik pemanfaatan data untuk keperluan bisnis*. Yogyakarta: Garah Ilmu, 2007.
[15] A. S. Nugroho, "Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika," 2003.
[16] T. E. Mathew, "To Predict Continuous or Discrete Target Values From," vol. 10, no. 3, pp. 55–63, 2019.