REPOSITOR, Vol. 2, No. 8, Agustus 2020, Pp. 1007-1016
ISSN : 2714-7975
E-ISSN : 2716-1382

Implementasi Algoritma C5.0 Untuk Menganalisa Gejala Prioritas Pada Anak yang Mengalami Bullying

Nabillah Annisa Rahmayanti[*1], Yufis Azhar[2], Gita Indah Marthasari[3]
[1,2,3]Universitas Muhammadiyah Malang/Teknik Informatika
nabillah_437043@webmail.umm.ac.id[*1], yufis@umm.ac.id[2] , gita@umm.ac.id[3]

Abstrak

Bullying sering terjadi pada anak-anak khususnya remaja dan meresahkan para orang tua. Maraknya kasus bullying di negeri ini bahkan sampai menyebabkan korban jiwa. Hal ini dapat dicegah dengan cara mengetahui gejala-gejala seorang anak yang mengalami bullying. Kondisi seorang anak yang tidak dapat mengungkapkan keluh kesahnya, tentu membuat orang tua dan juga guru di sekolah sukar dalam mengerti apa yang sedang menimpanya. Hal tersebut bisa saja dikarenakan anak sedang mengalami tindakan bullying oleh teman-temannya. Oleh karena itu peneliti memiliki tujuan untuk menghasilkan fitur yang telah terseleksi dengan menggunakan algoritma C5.0. Sehingga dengan menggunakan fitur yang telah terseleksi dapat meringankan pekerjaan dalam mengisi kuisioner dan juga mempersingkat waktu dalam menentukan seorang anak apakah terkena bullying atau tidak berdasarkan gejala yang ada di setiap pertanyaan pada kuisioner. Untuk menunjang data dalam penelitian ini, peneliti menggunakan kuisioner untuk mendapatkan jawaban dari pertanyaan yang berisi tentang gejala anak yang menjadi korban bullying. Jawaban dari responden akan diolah menjadi kumpulan data yang nantinya akan dibagi menjadi data latih dan data uji untuk selanjutnya diteliti dengan menggunakan Algoritma C5.0. Metode evaluasi yang digunakan pada penelitian ini yaitu 10 fold cross validation dan untuk menilai akurasi menggunakan confusion matrix. Penelitian ini juga melaukan perbandingan dengan beberapa algoritma klasifikasi lainnya yaitu Naive Bayes dan KNN yang bertujuan untuk melhat seberapa akurat algoritma C5.0 dalam melakukan seleksi fitur. Hasil pengujian menunjukkan bahwa algoritma C5.0 mampu melakukan seleksi fitur dan juga memiliki tingkat akurasi yang lebih baik jika dibandingkan dengan algoritma Naive Bayes dan KNN dengan hasil akurasi sebelum menggunakan seleksi fitur sebesar 92,77% dan setelah menggunakan seleksi fitur sebesar 93,33%.

Kata Kunci: *Bullying, Algoritma C5.0, Pohon Keputusan, Seleksi Fitur*

Abstract

Bullying often occurs in children, especially teenagers and unsettles parents. The rise of cases of bullying in this country even caused casualties. This can be prevented by knowing the symptoms of a child who has bullying. The condition of a child who cannot express his complaints, certainly makes parents and teachers at school difficult to understand what is happening to them. This could be because the child is experiencing bullying by his friends. Therefore, researchers have a goal to produce selected features using the C5.0 algorithm. So using the selected features can ease the work in filling out questionnaires and also shorten the time in determining whether a child is exposed to bullying or not based on the symptoms in each question in the questionnaire. To support the data in this study, the researcher used a questionnaire to get answers to questions that contained the symptoms of children who were victims of bullying. The answer from the respondent will be processed into a data collection which will later be divided into training data and test data for further research using the C5.0 Algorithm. The evaluation method used in this study is 10 fold cross validation and to assess accuracy using confusion matrix. This study also carried out a comparison with several other classification algorithms, namely Naive Bayes and KNN which aimed to see how accurate the C5.0 algorithm was in feature selection. The test results show that the C5.0 algorithm is capable of feature selection and also has a better accuracy compared to the Naive Bayes and KNN algorithms with accuracy results before using feature selection of 92.77% and after using feature selection of 93.33%.

Keywords: *Bullying, C5.0 Algorithm, Decision Tree, Feature Selection*

1. Pendahuluan

Perilaku bullying adalah perilaku yang membuat seseorang merasa tidak nyaman, depresi, cedera fisik atau psikologis [1]. Penindasan yang dilakukan oleh sekelompok anak biasanya cenderung lebih bersifat kekerasan fisik. Tetapi tidak menutup kemungkinan bahwa bullying yang dialami oleh seorang anak mungkin tidak terlihat dari fisik tetapi sebenarnya anak tersebut mengalami tekanan mental. Jika seorang anak mengalami tekanan mental, juga dimungkinkan untuk menyebabkan hilangnya nyawa karena anak itu dapat melakukan bunuh diri karena dia tidak tahan dengan perlakuan buruk orang lain.Penindasan sering terjadi di kalangan remaja dalam rentang usia 9-15 tahun, yang merupakan saat ketika anak-anak mencari identitas mereka [2]. Tidak heran perilaku anak lebih ekspresif atau malah sebaliknya. Dengan sikap seorang anak yang sulit berbicara dengan orang lain tentang hal-hal yang tidak menyenangkan adalah masalah utama.

Oleh karena itu, dengan sikap anak yang sulit mengungkapkan hal-hal buruk yang terjadi padanya di sekolah, maka cara untuk mendeteksi apakah seorang anak adalah korban bullying atau tidak adalah dengan melakukan tes psikologi dengan kuesioner.Kuesioner yang digunakan dalam penelitian ini meliputi 54 pertanyaan terkait dengan gejala anak-anak yang mengalami bullying. 54 pertanyaan dalam kuesioner yang diberikan tentu akan membutuhkan waktu lama untuk diisi dan juga cukup melelahkan untuk membaca setiap pertanyaan. Sehingga perlu untuk menganalisa gejala prioritas dengan menerapkan algoritma C5.0 yang dalam prosesnya menggunakan seleksi fitur untuk mendapatkan gejala prioritas.Gejala prioritas yang telah dihasilkan dari proses klasifikasi dengan algoritma C5.0 kemudian akan diuji ulang menggunakan algoritma klasifikasi lain yaitu Naive Bayes dan KNN sebagai pembanding. perbandingan akan terlihat dari hasil akurasi yang diperoleh dari masing-masing algoritma.

1.1 Penelitian Yang Berhubungan

Beberapa penelitian yang sama mengenai seleksi fitur dan juga penerapan algoritma C5.0 pernah dilakukan. Dalam beberapa penelitian tersebut mengusulkan untuk menggunakan algoritma yang memungkinkan untuk menghasilkan performa yang lebih baik dan juga di dapatkan kesimpulan bahwa seleksi fitur dapat meningkatkan hasil klasifikasi dengan signifikan.

Penelitian [3] menerapkan seleksi fitur untuk meningkatkan hasil diagnosis pada kanker payudara. Seleksi fitur diterapkan pada beberapa algoritma klasifikasi yaitu SMO, MLP, C4.5, dan Naive Bayes. Peneliti pada penelitian ini menunjukkan masing-masing algoritma klasifikasi memiliki peforma yang berbeda terhadap masing-masing metode seleksi fitur dan ia memberikan kesimpulan bahwa metode seleksi dapat meningkatkan hasil diagnosis klasifikasi kanker payudara secara signifikan dengan jumlah fitur yang lebih kecil.

Penelitian [4] menerapkan seleksi fitur information gain untuk memprediksi performa akademik siswa. Seleksi fitur pada penelitian ini diterapkan pada beberapa algoritma machine learning di antaranya decision tree J48, random forest, neural network, SVM, Naive Bayes. Dari penelitian ini dihasilkan kesimpulan bahwa dengan seleksi fitur information gain dapat mempengaruhi tingkat akurasi dari setiap algoritma klasifikasi machine learning yang di ujikan.

Penelitian [5] menerapkan algoritma decision tree C5.0 untuk peramalan forex. Pada penelitian ini peneliti menjelaskan bahwa algoritma decision tree C5.0 melakukan pemilihan atribut berdasarkan nilai dari ifnormation gain tertinggi pada pembentukan tree yang akan memberikan hasil prediksi. Penelitian ini juga menyarankan untuk pengembangan selanjutnya dapat melakukan variasi pada penerapan metode C5.0 agar menghasilkan prediksi yang optimal.

1.2 Konsep Dasar Teori

1.2.1 *Entropy*

Entropy menerapkan konsep probabilitas dalam menentukan seberapa besar entropy tersebut dalam suatu kejadian. Entropy digunakan untuk menentukan cabang pada pohon keputusan. Dengan perhitungan yang sama , dilakukan pada atribut berdasarkan pengelompokan jumlah kasus. Persamaan dalam menentukan entropy dapat dilihat pada Persamaan 1.

$$Entropy = \sum_{1-i}^{n} - pi * log_2\ pi \quad (1)$$

1.2.2 Information Gain

Information Gain merupakan salah satu metode seleksi fitur yang banyak digunakan untuk menentukan batas dari seberapa besar pengaruh suatu atribut tersebut [6]. Atribut dengan kriteria pembobotan yang sudah terpenuhi nantinya akan digunakan pada proses klasifikasi dengan algoritma C5.0. Atribut dengan information gain tertinggi akan menjadi parent dan yang memiliki information gain terendah akan menjadi bagian leaf [7]. Information gain dengan nilai tertinggi akan menjadi akar atau parent pada suatu pohon keputusan. Untuk menentukan information gain dapat dilihat pada Persamaan 2.

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} (\frac{|S_i|}{|S|} * Entropy(S_i) \quad (2)$$

1.2.3 Metode Klasifikasi

a. Algoritma C5.0

C5.0 adalah versi komersial dari C4.5 yang secara luas digunakan di banyak pemaketan data mining seperti Clementine and RuleQuest. Tidak seperti C4.5, penggunaan algoritma yang tepat untuk C5.0 belum terungkap. Hasil menunjukkan bahwa C5.0 meningkatkan pada penggunaaan memori sekitar 90%, lebih cepat daripada C4.5 [3] .

Penggunaan memori lebih efisien pada C5.0 dibanding dengan penggunaan C4.5. C5.0 mendapat pohon keputusan yang lebih ringkas dibandingkan dengan C4.5. Aturan C5 memiliki tingkat kesalahan yang lebih rendah pada kasus yang tidak terlihat. Jadi membandingkan dengan C4.5 tentunya keakuratan hasil yang baik terdapat pada algoritma C5.0. Algoritma C5.0 secara otomatis memungkinkan penghapusan atribut yang tidak membantu atau tidak relevan [8].

Algoritma C5.0 digunakan pada penelitian ini untuk menghasilkan seleksi fitur dengan pengujian menggunakan 54 atribut. Selanjutnya dilakukan pengujian kembali setelah dilakukan seleksi fitur dengan menggunakan 19 atribut. Algoritma C5.0 juga mampu memberikan hasil akurasi yang tinggi, hal ini dikarenakan C5.0 menghapus atribut-atribut yang menurutnya tidak relevan.

b. *Naive Bayes*

Algoritma Naive Bayes merupakan salah satu dari algoritma klasifikasi. Naive bayes mengklasifikasikan dengan menggunakan metode probabilitass dan statistik yang ditemukan oleh seorang ilmuwan yang bernama Thomas Bayes. Klasifikasi pada Naive Bayes dapat diasumsikan yaitu dengan ada atau tidaknya suatu ciri tertentu dari sebuah kelas yang tidak memiliki keterhubungan dengan ciri dari kelas lainnya [9]. Untuk menentukan perhitungan algoritma naive bayes dapat dilihat pada Persamaan 3.

$$P(C|X) = \frac{P(x|c)P(c)}{P(x)} \quad (3)$$

Pada penelitian ini, algoritma Naive Bayes diperlukan untuk membandingkan hasil akurasi dari klasifikasinya yang di dapatkannya dengan beberapa algoritma lain. Yaitu dengan algoritma C5.0 dan KNN. Hal ini karena algoritma Naive Bayes merupakan salah satu dari algoritma klasifikasi juga.

c. *KNN (K-Nearest Neighbors)*

K-NN termasuk dari kelompok instance-based learning. Algoritma KNN juga merupakan salah satu teknik dari lazy learning. Proses K-NN dalam mengklasifikasi dilakukan dengan mencari kelompok k objek dalam data latih yang paling dekat atau yang palig mirip dengan objek pada data baru atau data uji [10]. Pada penelitian ini, algoritma KNN digunakan sebagai pembanding hasil akurasi dari klasifikasi yang di dapatkannya dengan beberapa algoritma lain yaitu Naive Bayes dan C5.0. hal ini karena KNN merupakan salah satu metode klasifikasi juga. Untuk mendeifinisikan jarak antara objek x dan y, maka digunakan rumus jarak *Euclidian* yang dapat dilihat pada Persamaan 4.

$$dxy = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (4)$$

1.2.4 Evaluasi Performa

a. Akurasi

Dalam penelitian ini, kinerja setiap algoritma klasifikasi sebelum dan sesudah pemilihan fitur akan diukur berdasarkan akurasi. Dengan persamaan berikut. Jika hasil akurasi setiap algoritma telah diperoleh, maka itu dibandingkan untuk melihat akurasi mana yang lebih tinggi di antara ketiga algoritma. Seleksi fitur dapat mempengaruhi akurasi menjadi lebih baik [11]. Algoritma C5.0 lebih baik untuk menghasilkan seleksi fitur dan akurasi atau bahkan algoritma perbandingan yang lebih baik. Persamaan untuk menentukan entropi dapat dilihat pada Persamaan 5.

$$Akurasi = \frac{TP + TN}{TP + FP + FN + TN} \; X \; 100\% \tag{5}$$

2. Metode Penelitian

2.1 Kuisioner

Pada penelitian ini untuk mendapatkan data beberapa responden, yaitu menggunakan kuisioner seperti pada Tabel 1, yang berisi 54 pertanyaan atau atribut yang berhubungan dengan gejala-gejala anak yang menjadi korban bullying. Tentunya isi dari pertanyaan tersebut mencakup aspek penilaian dari segi lingkungan sekolah, keluarga dan juga kepribadian anak yang berdampak bagi anak menjadi korban daripada bullying [11].

Tabel 1. Potongan Data Kuisioner

No	Pernyataan
1	Saya adalah anak yang lemah di sekolah, karena saya lemah saya selalu diganggu oleh temen- temen sekelas.
2	saya sering sekali merasa beda dari temen- temen yang lain. Dan ini membuat saya rendah diri jika bergaul dengan teman –teman.
3	Saya merasa sendirian dan tidak memilki teman di sekolah
4	Saya sering sekali di ejek oleh temen kelas. Hal itu membuat saya terhina
5	Saya merasa hinaan dan perilaku buruk yang teman saya berikan terhadap saya di sekolah sangat berpengaruh buruk terhadap kepribadian saya

2.2 Data Set

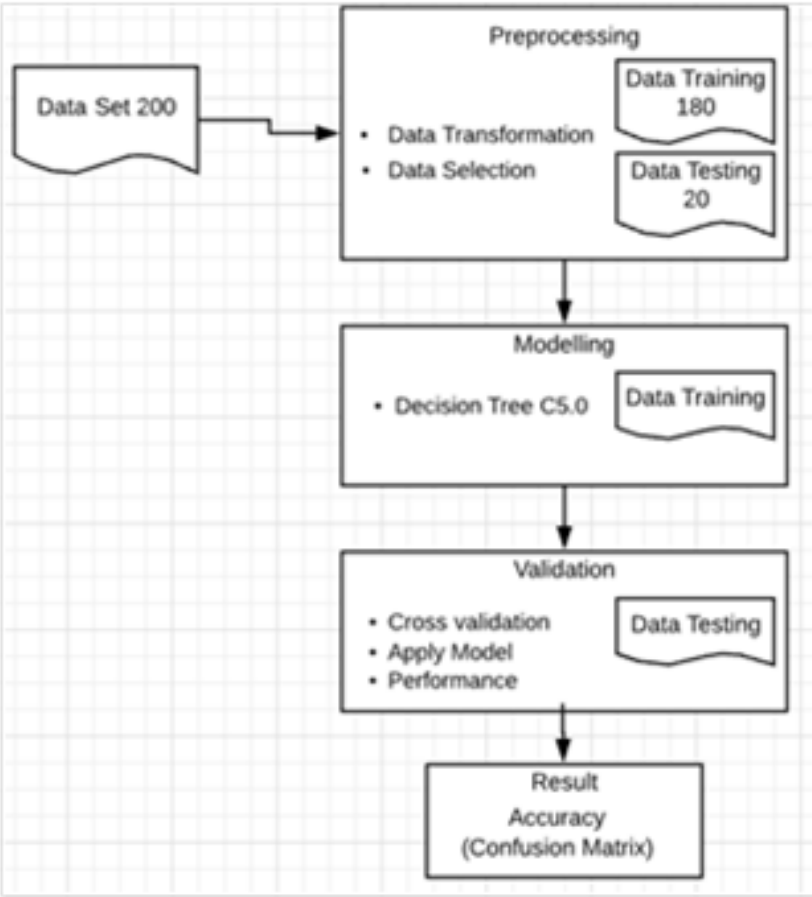
Data yang digunakan dalam penelitian ini diambil dari kuesioner yang diberikan kepada siswa SMP Negeri 5 Kota Pasuruan. Data yang digunakan adalah jawaban dari kuesioner yang diberikan. Dataset ini berisi 200 data dengan 54 atribut. Data dengan kelas positif (+) memiliki total 100 data sedangkan data dengan kelas negatif (-) memiliki total 100 data. Setiap kelas memiliki persentase 50%. Data yang dihasilkan dapat dilihat pada Tabel 2.

Tabel 2. Potongan Data Set Hasil Kuisioner

Nama	Kelas	Pernyataan					Bullying
		P1	P2	P3	P4	P5	
Hanifah	7E	1	1	1	1	1	+
M.Nasir	8C	4	3	3	3	3	-

2.3 Alur Penelitian

Penelitian ini berjalan sesuai dengan jalur yang harus dilalui sebelum beralih ke tahap implementasi. bagan alur penelitian dapat dilihat pada Gambar 1.



Gambar 1. Skema Alur Penelitian

a. Transformasi Data

Data yang telah dihasilkan dari akuisisi kuesioner berikutnya akan melalui fase transformasi sesuai dengan tipe data yang dapat dibaca oleh program pengolahan data. Seperti pada Tabel 3 di bawah ini, yang merupakan contoh data kuesioner yang memiliki parameter seperti sangat setuju, setuju, kurang setuju dan tidak setuju.

Tabel 3. Hasil Kuisioner Sebelum Dilakukan Transformasi Data

Pernyataan	Sangat Setuju	Setuju	Kurang Setuju	Tidak Setuju
P1				✓
P2			✓	
P3		✓		

Parameter penilaian kuesioner diubah menjadi angka, parameter sangat setuju memiliki nilai 4, parameter setuju memiliki nilai 3, parameter kurang setuju memiliki nilai 2, sedangkan parameter tidak setuju memiliki nilai 1 Seperti yang dapat dilihat pada Tabel 4 di bawah ini.

Tabel 4. Hasil Kuisioner Setelah Dilakukan Transformasi Data

Subjek	Pernyataan				
	P1	P2	P3	P4	P5
1	1	1	1	1	1
2	1	1	1	1	1
15	1	1	1	2	1
16	1	1	1	3	1
17	4	3	3	3	3

b. Seleksi Data

Data yang telah dikumpulkan ke dalam kumpulan data berikutnya akan dipilih kembali menjadi data pelatihan dan data uji dengan skala perbandingan 90% data pelatihan dan 10% data uji. Hasil distribusi data pelatihan dapat dilihat pada Tabel 5.

Tabel 5. Potongan Data Latih

Subject	Statement					Bullying
	P1	P2	P3	P4	P5	
21	1	1	1	1	2	+
22	1	2	1	1	1	+
38	1	1	1	1	1	-
39	2	2	1	2	2	-

Jika jumlah tersebut dihitung, data pelatihan adalah 180 data dan data uji adalah 20 data. data pelatihan dalam gambar hanya sebagian dari seluruh lath data yang berjumlah 180 data. Hasil distribusi data pengujian dapat dilihat pada Tabel 6.

Tabel 6. Potongan Data Uji

Subjek	Pernyataan					Bullying
	P1	P2	P3	P4	P5	
1	1	1	1	1	1	+
2	1	1	1	1	1	+
3	1	1	1	1	1	+
4	1	2	1	2	1	-

c. Pemodelan

Metode yang akan digunakan dalam pemodelan ini adalah Algoritma C5.0. Dalam pemodelan ini, algoritma C5.0 akan mencari fitur yang dipilih yang akan menjadi gejala prioritas dan juga dalam hal kinerjanya, yaitu kinerja vektor (akurasi) dan kebingungan matriks. Data yang digunakan telah melalui proses pemilihan data dan dibagi menjadi 180 data pelatihan dan 20 data uji. Implementasi algoritma C5.0 dalam penelitian ini menggunakan bahasa pemrograman R dengan pengujian sepuluh kali menggunakan validasi silang untuk menemukan pemilihan fitur yang memiliki tingkat akurasi tinggi dari sepuluh tes. Kode sumber untuk menghasilkan model dan matriks kebingungan untuk menghitung akurasi dapat dilihat pada Gambar 2.

```
crx <- read.table( file.choose(), header=FALSE, sep="," )
write.table( crx, "pengujianR.csv", quote=FALSE, sep="," )
head( crx, 6 )
crx <- crx[ sample( nrow( crx ) ), ]
X <- crx[1:54]
y <- crx[,55]
trainX <- X[21:200,]
trainy <- y[21:200]
testX <- X[1:20,]
testy <- y[1:20]
#install.packages("C50")
library(C50)
model <- C50::C5.0( trainX, trainy )
summary( model )
plot(model)

8:20  (Top Level)  R Script
```

Gambar 2. Source Code Pemodelan dengan Algoritma C5.0

d. Validasi

Sebelum pemodelan menggunakan algoritma C5.0, data uji dibagi menjadi 10 bagian menggunakan metode cross-validation. Sebelum pemodelan menggunakan algoritma C5.0, data uji dibagi menjadi 10 bagian menggunakan metode cross-validation. Sehingga pengujian dilakukan 10 kali dengan data uji yang berbeda. Setelah pengujian menggunakan validasi silang, algoritma C5.0 kemudian dimodelkan untuk menghasilkan akurasi dan pemilihan fitur. Skema untuk mendistribusikan data uji dari seluruh kumpulan data dapat dilihat pada Gambar 3.

Pengujian 1	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6	Data 7	Data 8	Data 9	Data 10
	Data 11	Data 12	Data 13	Data 14	Data 15	Data 16	Data 17	Data 18	Data 19	Data 20
Pengujian 2	Data 21	Data 22	Data 23	Data 24	Data 25	Data 26	Data 27	Data 28	Data 29	Data 30
	Data 31	Data 32	Data 33	Data 34	Data 35	Data 36	Data 37	Data 38	Data 39	Data 40
Pengujian 3	Data 41	Data 42	Data 43	Data 44	Data 45	Data 46	Data 47	Data 48	Data 49	Data 50
	Data 51	Data 52	Data 53	Data 54	Data 55	Data 56	Data 57	Data 58	Data 59	Data 60
Pengujian 4	Data 61	Data 62	Data 63	Data 64	Data 65	Data 66	Data 67	Data 68	Data 69	Data 70
	Data 71	Data 72	Data 73	Data 74	Data 75	Data 76	Data 77	Data 78	Data 79	Data 80
Pengujian 5	Data 81	Data 82	Data 83	Data 84	Data 85	Data 86	Data 87	Data 88	Data 89	Data 90
	Data 91	Data 92	Data 93	Data 94	Data 95	Data 96	Data 97	Data 98	Data 99	Data 100
Pengujian 6	Data 101	Data 102	Data 103	Data 104	Data 105	Data 106	Data 107	Data 108	Data 109	Data 110
	Data 111	Data 112	Data 113	Data 114	Data 115	Data 116	Data 117	Data 118	Data 119	Data 120
Pengujian 7	Data 121	Data 122	Data 123	Data 124	Data 125	Data 126	Data 127	Data 128	Data 129	Data 130
	Data 131	Data 132	Data 133	Data 134	Data 135	Data 136	Data 137	Data 138	Data 139	Data 140
Pengujian 8	Data 141	Data 142	Data 143	Data 144	Data 145	Data 146	Data 147	Data 148	Data 149	Data 150
	Data 151	Data 152	Data 153	Data 154	Data 155	Data 156	Data 157	Data 158	Data 159	Data 160
Pengujian 9	Data 161	Data 162	Data 163	Data 164	Data 165	Data 166	Data 167	Data 168	Data 169	Data 170
	Data 171	Data 172	Data 173	Data 174	Data 175	Data 176	Data 177	Data 178	Data 179	Data 180
Pengujian 10	Data 181	Data 182	Data 183	Data 184	Data 185	Data 186	Data 187	Data 188	Data 189	Data 190
	Data 191	Data 192	Data 193	Data 194	Data 195	Data 196	Data 197	Data 198	Data 199	Data 200

Gambar 3. Skema Pembagian Data Uji Dengan Cross Validation

e. Hasil

Penelitian ini bertujuan untuk menghasilkan seleksi fitur di mana pemilihan fitur menyajikan gejala prioritas dari keseluruhan gejala. Gejala yang dimaksudkan adalah pertanyaan yang tersedia dalam kuesioner. Hasil pengujian menggunakan metode *cross-validation* menggunakan algoritma C5.0 tidak hanya fitur yang telah dipilih tetapi juga menghasilkan *confusion matrix*. Dengan *confusion matrix* ini, nilai akurasi dapat dihitung untuk menentukan model mana yang memiliki akurasi tertinggi dari sepuluh tes yang telah dilakukan.

Pengujian juga dilakukan dengan menggunakan algoritma lain seperti Naive Bayes dan KNN dengan bantuan WEKA dalam pengujian sebelum menggunakan seleksi fitur. Model yang dipilih tentu saja akan menjadi model referensi yang berisi fitur yang terseleksi untuk diuji ulang menggunakan algoritma C5.0 dan juga dengan algoritma klasifikasi lainnya seperti Naive Bayes dan KNN menggunakan WEKA. Selanjutnya, keakuratan setiap algoritma akan dibandingkan dan menentukan algoritma yang memiliki akurasi tertinggi.

3. Hasil Penelitian dan Pembahasan

3.1 Implementasi Dengan Menggunakan Algoritma C5.0

Tabel 7 berikut adalah hasil perbandingan akurasi dari sepuluh tes menggunakan validasi silang pada algoritma C5.0. akurasi tertinggi diperoleh pada tes pertama dengan nomor urut data uji 1 hingga 20. Tabel 7 menjelaskan hasil penerapan algoritma C5.0 untuk pengujian sebelum menggunakan fitur yang dipilih.

Tabel 7. Hasil Akurasi dengan Algoritma C5.0

Pengujian	Data Uji	Algoritma C5.0
		Akurasi
1	1-20	92,77%
2	21-40	89,44%
3	41-60	91,11%
4	61-80	92,22%
5	81-100	91,11%
6	100-120	90%
7	121-140	91,11%
8	141-160	92.22%
9	161-180	85,55%
10	181-200	90,55%

3.2 Implementasi Sebelum Menggunakan Seleksi Fitur

Setelah hasil akurasi masing-masing metode diperoleh, kita bisa melihat perbandingannya. Akurasi tertinggi ditemukan dalam pengujian dengan algoritma C5.0. Tabel 8 mengilustrasikan hasil implementasi dengan beberapa algoritma sebelum menggunakan fitur yang dipilih dengan uji data uji 1 hingga 20 seperti pada hasil tes sebelumnya bahwa tes pertama dengan skema menghasilkan akurasi tertinggi.

Tabel 8. Perbandingan Akurasi Sebelum Menggunakan Fitur Seleksi

Algoritma	Akurasi
C5.0	92,77%
Naive Bayes	70%
KNN	55%

Dalam tes ini tidak hanya menghasilkan matriks kebingungan tetapi juga menghasilkan fitur yang telah dipilih dengan urutan masing-masing atribut sesuai dengan pohon keputusan yang dihasilkan dari hasil klasifikasinya. Tabel di bawah mengilustrasikan hasil implementasi sebelum menggunakan fitur pemilihan yang menggunakan semua atribut sebagai data uji dan data pelatihan. Ada contoh dari 19 fitur yang dipilih seperti yang dapat dilihat pada Tabel 9.

Tabel 9. Potongan Hasil Seleksi Fitur

Atribut/Fitur	Pernyataan
P24	Orang tua saya memberikan hadiah jika saya mendapatkan suatu prestasi
P38	Saya pernah di ejek oleh kakak kelas. Itu sangat menyakitkan buat saya
P30	Orangtua saya membeda-bedakan saya dengan saudara saya
P4	Saya sering sekali di ejek oleh teman kelas. Hal itu membuat saya terhina
P53	Saya tidak suka konseling kelompok karena khawatir teman-teman menceritakan masalah kepada orang lain

3.3 Implementasi Dengan Menggunakan Fitur Seleksi

Tes yang dilakukan saat ini menggunakan fitur yang telah dipilih sebagai data pelatihan dan data uji. Tidak lagi menggunakan atribut keseluruhan yang diuji sebelum menggunakan fitur yang dipilih. Fitur yang digunakan dalam tes ini berjumlah 19 fitur dengan skema tes yang sama yaitu data uji dari urutan 1 hingga 20 dan data pelatihan dari urutan ke 21-200. Tabel 10 mengilustrasikan hasil implementasi beberapa algoritma menggunakan fitur yang dipilih.

Tabel 10. Perbandingan Akurasi Sebelum dan Sesudah Menggunakan Fitur Seleksi

Algoritma	Akurasi	
	Sebelum	Sesudah
C5.0	92,77%	93,33%
Naive Bayes	70%	65%
KNN	55%	55%

4. Kesimpulan

Dalam penelitian ini, algoritma C5.0 berhasil fitur seleksi dengan memproduksi 19 fitur yang dipilih dari semua 54 fitur. Fitur yang dipilih memiliki urutan yang cocok dengan model pohon keputusan yang dihasilkan, dengan fitur pertama yang merupakan induk atau root dan fitur terakhir adalah fitur ke-19 adalah daun atau simpul terakhir yang tidak memiliki cabang lagi. Urutan pertanyaan didasarkan pada perolehan informasi yang dapat diperoleh selama proses seleksi dalam algoritma C5.0. Akurasi yang diperoleh dengan menggunakan algoritma klasifikasi lain seperti Naive Bayes dan KNN berfungsi sebagai perbandingan dengan algoritma C5.0, hasil akurasi yang dihasilkan oleh dua algoritma komparatif juga lebih rendah dari hasil akurasi

algoritma C5.0. Akurasi yang diperoleh dengan pengujian menggunakan algoritma C5.0 sebelum menggunakan fitur seleksi sebesar 92,77% dan setelah menggunakan fitur pilihan 93,33%.

Penggunaan lebih dari satu set data memungkinkan algoritma C5.0 untuk menghasilkan akurasi yang lebih baik dalam memilih setiap fitur dan juga menerapkan algoritma C5.0 ke bahasa pemrograman yang memiliki tampilan desktop yang lebih baik, seperti PHP, yang sangat membantu dalam hal penampilan dibandingkan dengan penelitian ini. hanya menggunakan bahasa pemrograman R dan juga WEKA. Mengubah studi kasus dengan memilih siswa sekolah menengah dan responden siswa memungkinkan untuk mendapatkan hasil akurasi yang berbeda. Ini karena perbedaan usia dapat mempengaruhi kemampuan emosional seseorang untuk menghasilkan hasil klasifikasi yang lebih bervariasi.

Notasi

1. Persamaan Entropy

S : Himpunan Kasus
A : Fitur
n : Jumlah partisi S
pi : Proporsi dari Si terhadap S

2. Persamaan Information Gain

S : Himpunan Kasus
A : Fitur
N : Jumlah partisi atribut A
|Si| : Jumlah kasus pada partisi ke-i
|S| : Jumlah kasus dalam S

3. Persamaan Naive Bayes Classifier

X : Data dengan kelas yang belum diketahui
C : Hipotesis data X merupakan suatu kelas spesifik
P(C|X) : Probabilitas hipotests C berdasar kondisi X
P(C) : Probabilitas hipotesis C
P(X|C) : Probabilitas X berdasar kondisi pada hipotesis H
P(X) : Probabilitas dari X

4. Persamaan K-Nearest Neighbors

dxy : Jarak Sampel
Xi : Data sampel pengetahuan
Yi : Data input variabel ke-i
n : Jumlah sampel

5. Persamaan Akurasi

TP : True Positive (hasil benar)
FP : False Positive (hasil tak terduga)
FN : False Negative (hasil yang hilang)
TN : True Negative (tidak ada hasil yang benar)

Referensi

[1] P. Ppds, P. F. K. Unair, S. P. Departemen, S. M. F. Psikiatri, F. K. Unair, and P. Ii, “Dokter, Peserta PPDS I Psikiatri FK UNAIR/RSUD Dr. Soetomo, Peneliti I ** Psikiater, Konsultan, Staf Pengajar Departemen/SMF Psikiatri FK UNAIR/RSUD Dr. Soetomo, Peneliti II *** Dokter, Staf Pengajar Ilmu Kesehatan Masyarakat FK UNAIR Surabaya, Konsultan ,” pp. 1–11.

[2] B. Okeke-oti, “Page 1 1,” *System*, pp. 3–4, 2010.

[3] E. S. Wahyuni, F. T. Industri, P. Studi, T. Elektro, U. I. Indonesia, and N. Bayes, “Penerapan metode seleksi fitur untuk meningkatkan hasil diagnosis kanker payudara,” vol. 7, no. 1, pp. 283–294, 2016.

[4] B. N. Sari, “Implementasi Teknik Seleksi Fitur Information Gain Pada Algoritma Klasifikasi Machine Learning Untuk Prediksi Performa Akademik Siswa,” pp. 6–7, 2016.

[5] K. P. Wirdhaningsih, D. E. Ratnawati, U. B. Malang, D. Mining, and D. Tree, “Penerapan Algoritma Decision Tree C5.0 Untuk Peramalan Forex,” pp. 1–6, 2012.

[6] H. Deng and G. Runger, “Feature Selection via Regularized Trees,” pp. 10–15, 2012.
[7] B. Azhagusundari and A. S. Thanamani, “Feature Selection based on Information Gain,” no. 2, pp. 18–21, 2013.
[8] R. Pandya and J. Pandya, “C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning,” *Int. J. Comput. Appl.*, vol. 117, no. 16, pp. 18–21, 2015.
[9] T. Informatika, U. Malikussaleh, and A. Utara, “Penerapan Algoritma Naive Bayes,” vol. 8, no. 1, pp. 884–898, 2014.
[10] L. Data *et al.*, “Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa,” vol. 13, no. 2, pp. 195–202, 2016.
[11] A. L. B. Masalah, “No Title,” pp. 1–7, 2011.