PENERAPAN TEKNIK VOTE MENGGUNAKAN C4.5 NAIVE BAYES DAN K-NEAREST NEIGHBOR PADA DATA GANGGUAN AUTISME

Mohamad Arifandi Pratama, Agung Nilogiri, Habibatul Azizal Al Faruq

Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember

Jl. Karimata No.49 Jember Kode Pos 68121

Email : arifandy0805@gmail.com

ABSTRAK

Autisme merupakan gangguan pada perkembangan otak yang memengaruhi kemampuan penderita dalam berkomunikasi dan berinteraksi dengan orang lain. Di samping itu, autisme juga menyebabkan gangguan perilaku dan membatasi minat penderitanya. Dikarenakan untuk mengidentifikasi autisme dibutuhkan dokter spesialis yang jumlahnya tidak terlalu banyak dan waktu yang cenderung lama, dengan mengklasifikasikan gejala-gejala penderita gangguan autisme maka akan semakin cepat untuk untuk mengetahui gangguan yang dialami. Pada penelitian yang dilakukan oleh (Zhang, dkk, 2014) pada data Breast-cancer dengan algoritma C4.5 didapatkan akurasi 75,5%. Penelitian pada dataset gangguan autisme pada anak pernah dilalukan oleh (Sugara, dkk., 2018) pada penelitiannya yang menggunakan algorima C4.5 didapatkan akurasi sebesar 72%. Berdasarkan latar belakang tersebut penelitian dilakukan untuk meningkatkat akurasi pada data gangguan autisme. Untuk itu akan digunakan teknik *voting* pada algoritma C4.5, atribut yang digunakan yaitu GJ01 hingga GJ24. Metode yang akan digunakan pada *vote* yaitu C4.5, *K-nearest Neighbor* dan *naive bayes*. *Ensemble method* merupakan teknik untuk meningkatkan akurasi. Salah satu contoh dari *ensemble method* adalah *voting* atau bisa disebut *majority vote*. lalu digunakan ensemble method majority vote didapatkan akurasi 88,89%, dimana akurasi tersebut mendapatkan peningkatan akurasi sebesar 13,39% .

Kata Kunci : *Gangguan Autisme, Algoritma C4.5, vote, K-nearest neighbor, Naive bayes*

ABSTRACT

Autism is a disorder in brain development that affects the sufferer's ability to communicate and interact with others. In addition, autism also causes behavioral disorders and limits the interests of sufferers. Due to identifying autism requires specialist doctors who are not too many in number and tend to take a long time, by classifying the symptoms of people with autism disorders it will be faster to find out the disorders experienced. In a study conducted by (Zhang, et al, 2014) on breast-cancer data with the C4.5 algorithm, an accuracy of 75.5% was obtained.on autism disorder datasets in children has been carried out by (Sugara, et al., 2018) in his research using the C4.5 algorithm, an accuracy of 72% was obtained. Based on this background, research was conducted to increase the accuracy of data on autism disorders. For this reason, a voting technique will be used on the C4.5 algorithm, the attributes used are GJ01 to GJ24. The methods that will be used in voting are C4.5, K-nearest Neighbor and naive bayes.Ensemble method is a technique to improve accuracy. One example of the ensemble method is voting or can be called majority vote. Then the majority vote ensemble method was used to get an accuracy of 88.89%, where the accuracy got an increase in accuracy of 13.39%.

Keywords: Autism Disorder, C4.5 Algorithm, vote, K-nearest neighbor, Naive bayes

PENDAHULUAN

Autisme merupakan gangguan yang sejak dulu menjadi salah satu misteri di kedokteran dengan perkembangan yang sangat komplek. Gangguan autisme sudah ada sejak lama, namun banyak yang tidak mengetahui sebagai gangguan autis. Pada cerita zaman dulu sering dianggap hal yang tidak normal pada anak, gejala autis sudah menunjukkan perilaku yang tidak normal serperti pada umumnya. Menolak ketika digendong, sering menangis saat malam hari dan tidur pada siang hari. Seringkali membuat orang tuanya bingung dengan bahasa yang tidak mereka mengerti. Mereka bisa mencakar, mengigit ataupun menyerang saat dalam kodisi marah. Terkadang tertawa seperti ada yang mengajaknya bercanda dan orang tua ada yang menganggapnya tertukar dengan anak peri, sehingga tidak bisa menyesuaikan perilaku dengan kehidupan manusia normal pada umumnya.

Machine learning adalah ilmu pengetahuan yang berperan besar dalam bidangnya. Tanpa disadari seluruh pengguna *machine learning* menggunakan produk yang dihasilkan oleh penerapan *machine learning*. Pada era perkembangan teknologi

machine learning sudah sering digunakan untuk membantu dalam mendiagnosa suatu penyakit pada gangguan autisme dengan metode klasifikasi. *Decision tree* merupakan metode yang sering digunakan dalam klasifikasi. Salah satunya algoritma *decision tree C4.5*. *Ensemble method* adalah menggabungkan beberapa klasifikasi tree untuk menghasilkan kinerja prediksi yang lebih baik daripada klasifikasi tree tunggal.sehingga meningkatkan akurasi model. Salah satunya adalah *Majority Voting* yang merupakan peningkatan dari algoritma C4.5 oleh karena itu pemilihan metode yang digunakan membuktikan kelebihan pada *Majority Voting* mendapatkan akurasi yang lebih baik dibandingkan dengan metode klasifikasi biasa, dengan demikian pada penelitian ini akan dilakukan klasifikasi algoritma C4.5 dengan metode *Majority Voting.*

Rumusan Masalah

Berdasarkan latar belakang yang sudah dijelaskan di atas, rumusan masalah dalam penelitian ini adalah sebagai berikut :

1. Berapa tingkat presisi *Majority Voting* algoritma C4.5 pada deteksi dini gangguan autisme.
2. Berapa tingkat akurasi algoritma C4.5 sebelum dan sesudah menggunakan metode teknik *Majority Voting.*

TINJAUAN PUSTAKA

Autisme berasal dari kata “auto” yang berarti sendiri. Karena istilah ini penderita autis umumnya lebih memilih untuk menghindar dari segala bentuk interaksi secara langsung yang membuat mereka seolah hidup sendirian. Gangguan autisme bukanlah gangguan motorik biasa. Pengaruh terhadap pola pikir dan tindakan anak sangat besar, pada masa depan anak sangat berpengaruh dan tidak menutup kemungkinan menjadi abnormal selamanya jika dibiarkan tanpa mendapatkan terapi khusus (Sunu, 2012).

Autisme merupakan gangguan yang menjadi misteri di salah satu kedokteran dengan perkembangan yang sangat komplek. Autisme gangguan yang sudah ada sejak lama, namun banyak yang tidak mengetahui sebagai autis. Seringkali membuat orang tuanya bingung dengan bahasa yang tidak mereka mengerti. Mereka bisa mencakar, mengigit

ataupun menyerang saat dalam kodisi marah. Terkadang tertawa sendiri seolah-olah ada yang mengajaknya bercanda dan orang tua ada yang menganggapnya tertukar dengan anak peri, sehingga tidak bisa menyesuaikan dengan kehidupan manusia normal (Budhiman, 2002).

Machine learning menyelesaikan masalah dengan menggunakan data dari pembelajaran yang dilakukan manusia. Penerapan *machine learning* yaitu klasifikasi dan prediksi, klasifikasi adalah metode yang mengurutkan data dengan mengklasifikasikan atribut satu sama lain. Sedangkan prediksi digunakan untuk *output* dari *input* data yang telah diperoleh melalui data *training*.

Kemampuan yang belajar dengan mesin yang telah diprogram dari awal dengan rangkaian perintah tertentu. Mesin *learning* memiliki tugas yang dinamis dengan kapibilitas yang tinggi. Adapun aspek tugas *machine learning* dalam mesin pembelajaran .

a. *Machine reasoning* atau penalaran mesin yang merujuk kepada kemampuan suatu sistem untuk mengambil kesimpulan dengan metode logis berdasarkan data yang disediakan kepadanya.
b. *Language processing* atau pemrosesan bahasa**,** yang berarti kemampuan sistem untuk mencerna dan menginterpretasi bahasa manusia.

Klasifikasi adalah suatu pengelompokan data dimana data yang digunakan tersebut mempunyai kelas label atau target. Sehingga algoritma-algoritma untuk menyelesaikan masalah klasifikasi dikategorisasikan ke dalam *supervised learning* atau pembelajaran yang diawasi. Teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan.Teknik ini menggunakan *supervised induction*, yang memanfaatkan kumpulan pengujian dari record yang terklasifikasi untuk menentukan kelas-kelas tambahan. Salah satu contoh yang mudah dan populer adalah dengan *Decision tree* yaitu salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi.

Decision tree merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon (*tree*) di mana setiap *node* merepresentasikan atribut,

cabangnya merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas. *Node* yang paling atas dari *decision tree* disebut sebagai *root* (Gorunescu, 2011).

Ada banyak algoritma pada klasifikasi *decision tree* ini. Suatu algoritma biasanya dikembangkan untuk meningkatkan kinerja algoritma yang sudah ada. Penentuan algoritma yang terbaik dalam *decision tree* tentunya tidak bisa ditentukan secara mutlak tetapi sangat tergantung dengan karakteristik training set-nya. Beberapa algoritma *decision tree* yang cukup populer antara lain : ID3, C4.5, dan CART. Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut :

a. Pilih atribut sebagai akar.
b. Buat cabang untuk tiap-tiap nilai.
c. Bagi kasus dalam cabang.
d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Naive Bayes merupakan sebuah pengklasifikasian probalistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma menggunakan teorema bayes dan mengansumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. *Naive Bayes* juga didefinisikan sebagai pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuan inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya. *Naive Bayes* didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai *output*. Dengan kata lain, diberikan nilai *output*, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. *Naive Bayes* sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang komples dari pada yang diharapkan (Saleh, 2015).

Algoritma *K-nearest neighbor* (KNN) merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data terdekat dari objek tersebut, KNN juga merupakan algoritma supervised learning dimana hasil klasifikasi data baru berdasarkan pada kategori sebagian besar jarak terdekat menuju K. Algoritma ini

menghitung berdasarkan jarak minimum dari data baru ke K terdekat yang sudah ditetapkan. Jarak antara data baru dengan data *learning* dihitung dengan cara mengukur jarak antara titik yang mewakili data baru dengan titik yang mewakili data *learning* dengan rumus *Euclidean Distance* (Fitroh, 2015).

Aturan keputusan *Majority Voting* yang memilih salah satu dari banyak alternatif, berdasarkan kelas prediksi dengan suara terbanyak. Pemungutan suara mayoritas tidak memerlukan parameter apa pun untuk mengklasifikasi klasifikasi individual yang telah dilatih. Dalam hal pemungutan suara terbobot, bobot pemungutan suara harus bervariasi di antara kelas *output* yang berbeda di setiap *classifikasi*. Masalah pembobotan dapat dilihat sebagai masalah optimisasi.

Cross-validation menggeneralisasi pendekatan ini dengan mensegmentasi data ke dalam k partisi berukuran sama. Selama proses, salah satu dari partisi dipilih untuk *training*, sedangkan sisanya untuk testing. Prosedur ini diulangi k kali sedemikian sehingga setiap partisi digunakan untuk *testing* tepat satu kali. Total *error* ditentukan dengan menjumlahkan *error* untuk semua k proses tersebut (Tan, dkk. 2005).

Confusion Matrix adalah pengukuran terhadap kinerja suatu sistem klasifikasi merupakan hal yang penting. Kinerja sistem klasifikasi menggambarkan seberapa baik sistem dalam mengklasifikasikan data. *Confusion matrix* merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada dasarnya *confusion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya (Rosandy, 2016).

Tabel 1 Confusion Matrix

		True Value	
		Positif	Negatif
Predictio n	Positif	TP	FP

	Negatif	FN	TN

Dimana:

- TP adalah *True Positive*, yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem.
- TN adalah *True Negative*, yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem.
- FN adalah *False Negative*, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.
- FP adalah *False Positive*, yaitu jumlah data positif namun terklasifikasi salah oleh sistem.

$$precision = \frac{TP}{TP + FP}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

METODOLOGI PENELITIAN

Penerapan teknik *Majority Voting* untuk mengetahui akurasi terbaik dari hasil klasifikasi menggunakan teknik pada dataset gangguan autisme pada anak. Metodologi yang digunakan pada penelitian ini terdiri dari beberapa tahap yaitu pendahuluan, pengumpulan data, analisis data, perancangan sistem dan penarikan kesimpulan.

Data yang digunakan pada penelitian ini adalah data gejala gangguan autisme pada anak dari penelitian oleh (Sugara, Widyatmoko, Prakoso & Saputro, 2017). Data yang digunakan terdiri dari 24 parameter dan juga 3 output yang dihasilkan. Data yang akan digunakan berjumlah 50 *record*. Data diambil pada suatu lembaga autisme di bekasi pada periode 2018 oleh (Sugara et al., 2018). Data yang nantinya akan digunakan

sebagai contoh perhitungan adalah 20 *record* 15 data untuk *training* dan 5 data untuk *testing*.

Analisis Data

Dalam penelitian ini, analisis yang digunakan adalah deskriptif untuk data gejala pada gangguan autism pada anak menggunakan software Microsoft excel 2010. Kemudian digunakan teknik *Majority Voting* pada algoritma C4.5, Naive Bayes, *K-nearest Neighbor* (KNN).

Proses Algoritma C4.5

a. Menyiapkan *Data Training*

Data training yang akan digunakan pada contoh perhitungan C4.5 ini menggunakan data gejala gangguan autisme. Yang mana pemilihan data dilakukan secara acak (random).

b. Perhitungan C4.5

Pertama tentukan atribut akar berdasarkan pada nilai Gain tertinggi dari atribut-atribut yang ada. Sedangkan untuk mendapatkan nilai Gain, harus menentukan terlebih dahulu nilai Entropy. Rumus dasar dari Entropy dan Gain sebagai berikut :

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

Keterangan:

S : Himpunan Kasus

n : Jumlah partisi S

pi: Proporsi dari Si terhadap S

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

Keterangan:

S : Himpunan kasus

A : Atribut

n : Jumlah partisi atribut A

|Si| : Jumlah kasus pada partisi ke i

|S| : Jumlah kasus dalam S

Proses *K-nearest neighbor* (KNN)

a. Pada proses *K-nearest neighbor* akan dilakukan pengolahan data yang akan dijadikan *numeric.*
b. Menentukan Nilai K . Nilai K = 3
c. Menghitung jarak Euclid data baru terhadap data yang ada. Tabel 3.11 menunjukkan perhitungan jarak Euclid. Rumus Euclidean distance sebagai berikut :

$$D(x, y) = \sqrt{\sum_{k=1}^{m}(x_{ik} - y_{ik})^2}$$

Dimana,

x_{ik} = nilai x pada training data

y_{ik} = nilai x pada testing data

m = batas jumlah banyaknya data

Nilai x_i merupakan nilai yang ada pada data *training*, sedangkan nilai y_i

d. Mengurutkan jarak Euclid terkecil.

Proses Naive Bayes

a. Menyiapkan data *training* yang akan dilakukan klasifikasi.
b. Melakukan perhitungan nilai jumlah sub atribut pada masing-masing *output* yang ada pada masing-masing atribut.
c. Setelah dihitung lalu siapkan data *testing* yang akan digunakan untuk menentukan klasifikasi yang akan digunakan untuk penilaian pada data *testing*.

$$P(\mathrm{H}|\mathrm{X}) = \frac{P\,(X|H).p(H)}{p(H)}$$

Dimana :

X : data dengan classyang belum diketahui

H : hipotesis data menggunakan suatu class spesifik

P(H|X) : probabilitas hipotesis H berdasar kondisi X (parteriori probabilitas)

P(H) : probabilitas hipotesis H (prior probabilitas)

P(X|H) : probabilitas X bedasarkan kondisi pada hipotesis H

P(X) : probabilitas H

d. Setelah itu lakukan perhitungan masing-masing pada masing-masing output dimana ada 3 yang akan dirubah menjadi c1,c2,c3. Dimana c1 adalah gangguan komunikasi, c2 adalah gangguan perilaku, dan c3 adalah gangguan interaksi sosial. Hasil yang diperoleh dari perhitungan.

Evaluasi

Evaluasi akan dilakukan dengan membandingkan hasil klasifikasi pada 3 metode yang telah ada. Suara terbanyak akan dijadikan hasil klasifikasi akhir.

IMPLEMENTASI DAN PENGUJIAN

Gambaran Dataset

Data penelitian yang akan digunakan yaitu data dari penelitian yang dilakukan oleh (Sugara, dkk, 2018). Data ini berupa data gangguan autisme pada anak yang terdiri dari 24 atribut yaitu GJ01 hingga GJ24 dan 3 output yaitu gangguan interaksi sosial, gangguan komunikasi, dan gangguan perilaku. Data ini diambil pada periode 2018 dengan jumlah data awal 50 *record*. Lalu karena dataset tidak *balance* maka data awal berubah menjadi 36 data dengan 12 data masing-masing *output*. Berikut adalah keterangan lebih lengkapnya :

Tabel 2 Keterangan Atribut

No	Nama Atribut	Keterangan
1	GJ01	Tidak memiliki kontak mata
2	GJ02	Suka diam/menyendiri
3	GJ03	Tidak suka dipeluk
4	GJ04	Tidak dapat merespon jika dipanggil orang
5	GJ05	Suka melakukan kegiatan/gerakan secara berulang-ulang
6	GJ06	Suka terpaku terhadap benda-benda tertentu
7	GJ07	Suka menyukai hal yang aneh seperti mencium-cium benda
8	GJ08	Suka mengungkapkan emosi (sedih, senang, marah dll) dengan sendirinya tanpa sebab
9	GJ09	Tidak bisa diam
10	GJ10	Tidak dapat berbicara
11	GJ11	Bisa berbicara namun tidak jelas
12	GJ12	Sering berbicara berlebihan
13	GJ13	Suka mengucapkan bahasa/kata-kata yang aneh secara berulang-ulang
14	GJ14	Tidak dapat menunjuk sesuatu dengan jari sendiri
15	GJ15	Tidak dapat menunjukkan keinginan dengan kata-kata
16	GJ16	Suka menarik-narik orang lain jika menginginkan sesuatu
17	GJ17	Tidak ada usaha dalam berkomunikasi
18	GJ18	Menghindar jika didekati
19	GJ19	Tidak dapat berinteraksi dengan lingkungan sekitar
20	GJ20	Tidak tertarik dengan orang lain
21	GJ21	Tidak peduli dengan sekitarnya
22	GJ22	Tidak suka dengan keramaian
23	GJ23	Tidak suka bermain dengan teman sebayanya

24	GJ24	Tidak dapat bersosialisasi dengan orang lain

Skenario Uji

Pengujian yang akan dilakukan menggunakan *tool Rapid miner* studio versi 9.5 sebagai pengaplikasiannya. Sebelum digunakan pada *tool Rapid miner* data terlebih dahulu dibagi menjadi data *training* dan data *testing*. Dimana dengan jumlah data yang berbeda sesuai dengan k-fold *cross validation*. Data akan dibagi menjadi k sama dengan 2, 3, 4, dan 6 seperti yang diperlihatkan sebagai berikut :

Tabel 3 Skenario K-Fold

skenario	K-fold percobaan yang ke-	Range Data Training	%	Range Data Testing	%
1	2 – fold percobaan ke - 1	1 – 18	50%	19 – 36	50%
2	2 – fold percobaan ke – 2	19 – 36	50%	1 – 18	50%
3	3 – fold percobaan ke - 1	1 – 24	66%	25 - 36	34%
4	3 – fold percobaan ke – 2	1 – 12 & 25 – 36	66%	13 – 24	34%
5	3 – fold percobaan ke – 3	13 – 36	66%	1 – 12	34%
6	4 – fold percobaan ke – 1	1 – 27	75%	28 -36	25%
7	4 – fold percobaan ke – 2	1 – 18 & 28 – 36	75%	19 – 27	25%
8	4 – fold percobaan ke – 3	1 – 9 & 19 – 36	75%	10 – 18	25%
9	4 – fold	10 – 36	75%	1 – 9	25%

	percobaan ke – 4				
10	6 – fold percobaan ke – 1	1 – 30	83%	31 - 36	17%
11	6 – fold percobaan ke – 2	1 – 24 & 31 – 36	83%	25 – 30	17%
12	6 – fold percobaan ke – 3	1 – 18 & 25 – 36	83%	19 – 24	17%
13	6 – fold percobaan ke – 4	1 – 12 & 19 – 36	83%	13 – 18	17%
14	6 – fold percobaan ke – 5	1 – 6 & 13 – 36	83%	7 – 12	17%
15	6 – fold percobaan ke – 6	7 – 36	83%	1 – 6	17%

Dapat dilihat pada tabel 4.2 akan ada 15 pengujian yang akan dilakukan dengan data *training* dan data *testing* yang berbeda-beda.

Hasil Pengujian Metode C4.5

Pada hasil pengujian ini yang akan perlihatkan adalah hasil terbaik dari *cross validation* sebelum menggunakan teknik *Ensemble*, didapatkan akurasi terbaik pada k-fold 6 percobaan ke 3 sebesar 83,33%, dengan presisi dimana nilai positif gangguan interaksi sosial memiliki hasil 100%, nilai positif pada gangguan komunikasi sebesar 100%, dan pada nilai positif gangguan memiliki perilaku presisi 66,67%.

accuracy: 83.33%

	true Gangguan Interaksi Sosial	true Gangguan Komunikasi	true Gangguan Perilaku	class precision
pred. Gangguan Interaksi Sos...	1	0	0	100.00%
pred. Gangguan Komunikasi	0	2	0	100.00%
pred. Gangguan Perilaku	1	0	2	66.67%
class recall	50.00%	100.00%	100.00%	

Gambar 1 Hasil Akurasi Dan Presisi Percobaan Ke-3

Hasil Pengujian Metode Naive Bayes

Pada hasil pengujian ini yang akan perlihatkan adalah hasil terbaik dari *cross validation* sebelum menggunakan teknik *Ensemble*, didapatkan akurasi terbaik pada k-fold 6 percobaan ke 4 sebesar 83,33%, dengan presisi dimana nilai positif gangguan interaksi sosial memiliki hasil 100%, nilai positif pada gangguan komunikasi sebesar 100%, dan pada nilai positif gangguan memiliki perilaku presisi 50%.

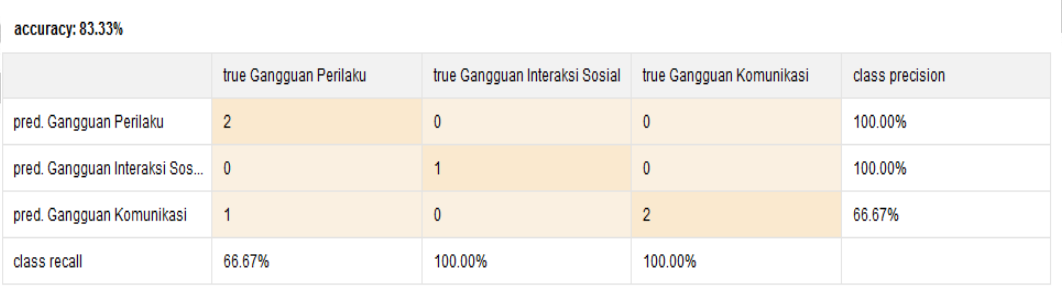
accuracy: 83.33%

	true Gangguan Interaksi Sosial	true Gangguan Komunikasi	true Gangguan Perilaku	class precision
pred. Gangguan Interaksi Sos...	2	0	0	100.00%
pred. Gangguan Komunikasi	0	2	0	100.00%
pred. Gangguan Perilaku	1	0	1	50.00%
class recall	66.67%	100.00%	100.00%	

Gambar 2 Hasil Akurasi Dan Presisi Percobaan Ke-4

Hasil Pengujian Metode K-nearest Neighbor

Pada hasil pengujian ini yang akan perlihatkan adalah hasil terbaik dari *cross validation* sebelum menggunakan teknik *Ensemble*, didapatkan akurasi terbaik pada k-fold 6 percobaan ke 6 sebesar 83,33%, dengan presisi dimana nilai gangguan interaksi sosial memiliki hasil 100%, nilai positif pada gangguan komunikasi sebesar 66,67%, dan pada nilai positif gangguan memiliki perilaku presisi 100%.



accuracy: 83.33%

	true Gangguan Perilaku	true Gangguan Interaksi Sosial	true Gangguan Komunikasi	class precision
pred. Gangguan Perilaku	2	0	0	100.00%
pred. Gangguan Interaksi Sos...	0	1	0	100.00%
pred. Gangguan Komunikasi	1	0	2	66.67%
class recall	66.67%	100.00%	100.00%	

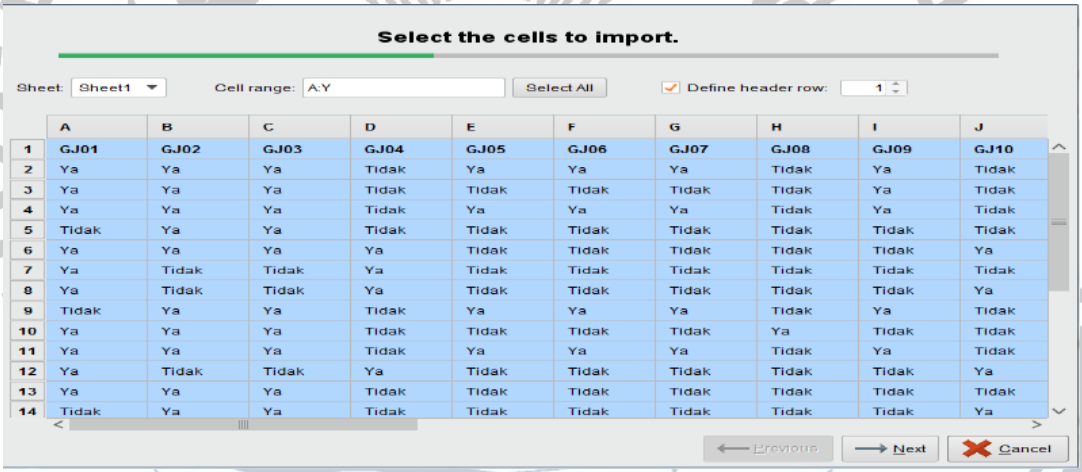
Gambar 3 Hasil Akurasi Dan Presisi Percobaan Ke-6

Hasil Pengujian Teknik *Majority Voting*

Pada hasil pengujian ini yang akan perlihatkan adalah hasil terbaik dari masing-masing k-fold *cross validation*. Dikarenakan nilai *output* ada 3 yaitu gangguan komunikasi, gangguan perilaku, dan gangguan interaksi sosial maka akan dilakukan pengujian pada *Rapid miner*, kemudian hasilnya akan dihitung dengan *confusion matrix* menggunakan 2 output dengan 3 *confusion matrix*.

Hasil Uji 2-Fold

Berdasarkan skenario uji pada pada pengujian dengan k = 2 yaitu membagi data latih (*training)* sebanyak 18 dan sisanya menjadi data uji (*testing)*. Setelah semua data telah disiapkan maka masukan data kedalam *Majority Vote* untuk diproses, berikut proses pengolahan data menggunakan teknik *Majority Vote* pada algoritma C4.5. langkah-langkah dalam implementasi *Rapid miner* akan dijelaskan pada gambar dibawah :

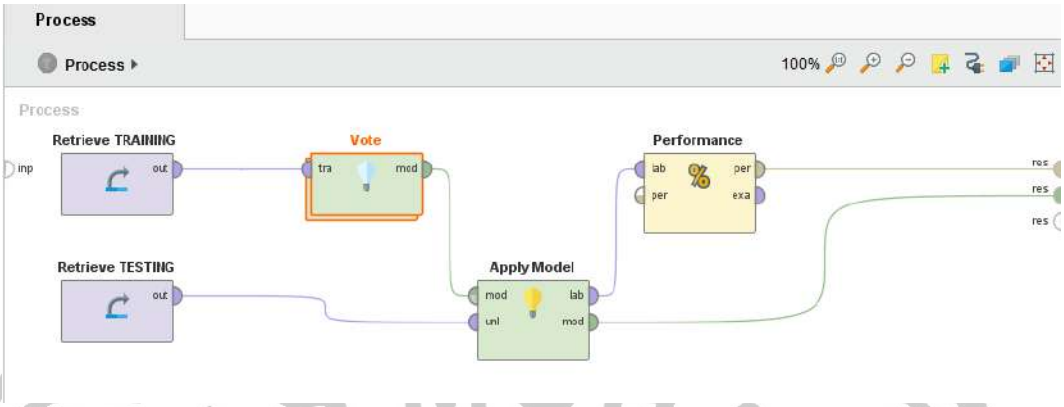


	A	B	C	D	E	F	G	H	I	J
1	GJ01	GJ02	GJ03	GJ04	GJ05	GJ06	GJ07	GJ08	GJ09	GJ10
2	Ya	Ya	Ya	Tidak	Ya	Ya	Ya	Tidak	Ya	Tidak
3	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Ya	Tidak
4	Ya	Ya	Ya	Tidak	Ya	Ya	Ya	Tidak	Ya	Tidak
5	Tidak	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak
6	Ya	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Ya
7	Ya	Tidak	Tidak	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak
8	Ya	Tidak	Tidak	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Ya
9	Tidak	Ya	Ya	Tidak	Ya	Ya	Ya	Tidak	Ya	Tidak
10	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Ya	Tidak	Tidak
11	Ya	Ya	Ya	Tidak	Ya	Ya	Ya	Tidak	Ya	Tidak
12	Ya	Tidak	Tidak	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Ya
13	Ya	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak
14	Tidak	Ya	Ya	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Ya

Gambar 4 Proses *Import* Data

Pada proses ini data yang di impor sesuai dengan pembagian dan untuk atribut gangguan yang di dalam *role* dirubah menjadi “label” agar *Rapid miner* mengetahui sebagai *output* yang digunakan. Lalu dilakukan pemodelan

seperti berikut :



Gambar 5 Proses Pengujian

Pada proses ini akan dilakukan pengujian pada data latih (*training)* untuk mendapatkan model dan data uji (*testing)* untuk mendapatkan nilai akurasi dan presisi. Setelah semuanya selesai maka proses akan dilakukan untuk mendapatkan akurasi dan presisi, didapatkan hasil sebagai berikut:

accuracy: 73.68%

	true Gangguan Komunikasi	true Gangguan Interaksi Sosial	true Gangguan Perilaku	class precision
pred. Gangguan Komunikasi	4	0	0	100.00%
pred. Gangguan Interaksi Sos...	0	6	2	75.00%
pred. Gangguan Perilaku	3	0	4	57.14%
class recall	57.14%	100.00%	66.67%	

Gambar 6 Hasil Akurasi Dan Presisi Percobaan Ke-1

Pada hasil pengujian pada gambar 4.3 nilai dari akurasi langsung dihitung dengan menggunakan 3 output tanpa merubah data dan mendapatkan akurasi sebesar 73,68 %. Maka akan dihitung menggunakan *confusion matrix* 2 output dengan 3 hasil.

4.6.2 Hasil Uji 3-Fold

Berdasarkan scenario uji pada pada pengujian dengan k = 3 yaitu membagi data latih (*training)* sebanyak 24 dan sisanya menjadi data uji (*testing)*. Setelah semua data telah disiapkan maka masukan data ke dalam *Rapidminer* untuk diproses. Proses dilakukan sama dengan pengujian 2-fold. Akurasi dan presisi yang didapatkan pada 3-fold sebagai berikut :

accuracy: 76.92%

	true Gangguan Interaksi Sosial	true Gangguan Komunikasi	true Gangguan Perilaku	class precision
pred. Gangguan Interaksi Sos...	5	0	0	100.00%
pred. Gangguan Komunikasi	0	2	0	100.00%
pred. Gangguan Perilaku	1	2	3	50.00%
class recall	83.33%	50.00%	100.00%	

Gambar 7 Hasil Akurasi Dan Presisi Percobaan Ke-2

Pada hasil pengujian pada gambar 4.5 nilai dari akurasi langsung dihitung dengan menggunakan 3 *output* tanpa merubah data dan mendapatkan akurasi sebesar 76,92%. Maka akan dihitung menggunakan *confusion matrix* 2 *output* dengan 3 hasil.

4.6.3 Hasil Uji 4-Fold

Berdasarkan skenario uji pada pada pengujian dengan k = 4 yaitu membagi data latih (*training)* sebanyak 27 dan sisanya menjadi data uji (*testing)*. Setelah semua data telah disiapkan maka masukan data kedalam *Rapid miner* untuk diproses. Proses dilakukan sama dengan pengujian 2-fold. Akurasi dan presisi yang didapatkan pada 4-fold sebagai berikut :

accuracy: 88.89%

	true Gangguan Perilaku	true Gangguan Komunikasi	true Gangguan Interaksi Sosial	class precision
pred. Gangguan Perilaku	2	0	1	66.67%
pred. Gangguan Komunikasi	0	3	0	100.00%
pred. Gangguan Interaksi Sos...	0	0	3	100.00%
class recall	100.00%	100.00%	75.00%	

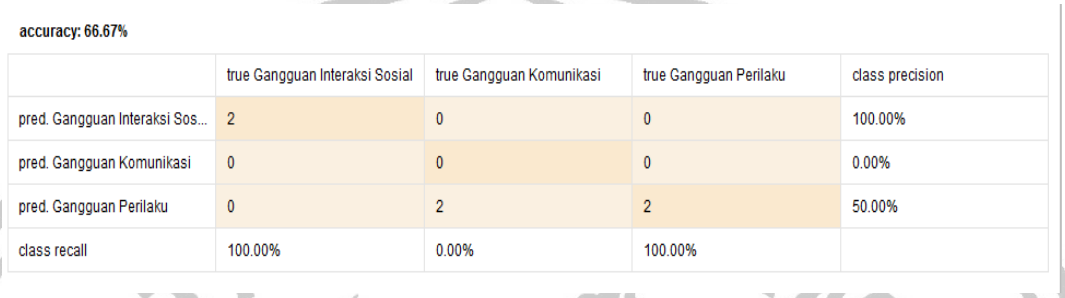
Gambar 8 Hasil Akurasi Dan Presisi Percobaan Ke-3

Pada hasil pengujian pada gambar 4.6 nilai dari akurasi langsung dihitung dengan menggunakan 3 output tanpa merubah data dan mendapatkan akurasi sebesar 88,89 %. Maka akan dihitung menggunakan *Confusion matrix* 2 output dengan 3 hasil.

4.6.4 Hasil Uji 6-Fold

Berdasarkan scenario uji pada pada pengujian dengan k = 6 yaitu membagi data latih (*training)* sebanyak 30 dan sisanya menjadi data uji (*testing)*. Setelah semua data

telah disiapkan maka masukan data kedalam *Rapidminer* untuk diproses. Proses dilakukan sama dengan pengujian sebelumnya. Akurasi dan presisi yang didapatkan pada 6-fold sebagai berikut :



accuracy: 66.67%

	true Gangguan Interaksi Sosial	true Gangguan Komunikasi	true Gangguan Perilaku	class precision
pred. Gangguan Interaksi Sos...	2	0	0	100.00%
pred. Gangguan Komunikasi	0	0	0	0.00%
pred. Gangguan Perilaku	0	2	2	50.00%
class recall	100.00%	0.00%	100.00%	

Gambar 9 Hasil Akurasi Dan Presisi Percobaan Ke-3

Pada hasil pengujian pada gambar 4.7 nilai dari akurasi langsung dihitung dengan menggunakan 3 output tanpa merubah data dan mendapatkan akurasi sebesar 66,67 %. Maka akan dihitung menggunakan *Confusion matrix* 2 *output* dengan 3 hasil.

Pembahasan

Pengujian data yang sudah dilakukan dengan menghitung jumlah keseluruhan data sebanyak 15 kali menggunakan teknik *cross validation* yang digunakan untuk mengevaluasi kinerja model dimana data dibagi menjadi data latih (*training*) dan data uji (*testing*). Melihat hasil jumlah k rata-rata memperoleh nilai yang berbeda beda pada tiap pengujian nilai k, namun juga terdapat mayoritas hasil yang sama sehingga diambil nilai akurasi dan presisi yang tertinggi, adapun setiap k-fold terdapat nilai yang tertinggi, maka hasil dari uji skenario sebanyak 4 kali dimana jumlah keseluruhan data dihitung 15 kali menghasilkan nilai akurasi yang tertinggi pada k-fold 4 percobaan ke-3 yaitu nilai akurasi 88,89% dan presisi yang didapatkan sebanyak 3 dimana nilai positif gangguan interaksi sosial memiliki hasil 66,67%, nilai positif pada gangguan komunikasi sebesar 100%, dan pada nilai positif gangguan perilaku memiliki presisi 100%.

Hasil dari beberapa pengujian k-fold diatas diperoleh presentase akurasi dan presisi dimana P adalah presisi dan A adalah akurasi, berikut merupakan gambaran dari tabel daftar pengujian hasil akurasi dan presisi :

Tabel 4 Daftar Hasil Akurasi Dan Presisi

Pengujian Ke-	K =2				K = 3				K = 4				K = 6			
	A	P			A	P			A	P			A	P		
		GIS	GK	GP		GIS	GK	GP		GIS	GK	GP		GIS	GK	GP
1	73,68%	75%	100 %	57,14%	75,00%	66,67%	100 %	66,67%	66,67%	66,67%	100 %	33,33%	66,67%	100 %	100 %	33,33%
2	83,33%	85,71 %	75%	100 %	76,92%	100 %	100 %	50%	66,67%	75%	0%	60%	60%	50%	100 %	50%
3					66,67%	75%	66,67%	100 %	88,89%	66,67%	100 %	100 %	66,67%	100 %	0%	50%
4									66,67%	66,67%	60%	100 %	66,67%	50%	100 %	100 %
5													83,33%	100 %	0,00 %	66,67%
6													50%	50%	50%	0%

Kesimpulan

Berdasarkan penelitian yang telah dilakukan dapat diambil kesimpulan sebagai berikut :dari 36 *record* data dengan 15 kali percobaan dimana nilai akurasi tertinggi didapatkan pada k-4 pada percobaan 3. Dimana akurasi yang didapatkan sebesar 88,89% dan presisi yang didapatkan sebanyak 3 dimana nilai positif gangguan interaksi sosial memiliki hasil 100%, lalu nilai positif pada gangguan komunikasi sebesar 100%, dan pada nilai positif gangguan perilaku memiliki presisi 66,67%. Pada perhitungan yang telah dilakukan sebelum menggunakan teknik *Ensemble* didapatkan akurasi terbaik pada k-fold 6 percobaan ke-5 dengan akurasi sebesar 83,33%, dimana nilai positif gangguan interaksi sosial memiliki hasil 100%, nilai positif pada gangguan komunikasi sebesar 0,00%, dan pada nilai positif gangguan perilaku memiliki presisi 66,67%. Kemudian dilakukan perhitungan sesudah menggunakan teknik *Majority Vote* didapatkan akurasi yang sama pada k-fold 4 percobaan ke-3 dengan akurasi sebesar 88,89%. Dimana nilai positif gangguan interaksi sosial memiliki hasil 100%, nilai positif pada gangguan komunikasi sebesar 100%, dan pada nilai positif gangguan perilaku memiliki presisi 66,67%. Dengan demikian pada penelitian yang telah dilakukan mendapatkan peningkatan akurasi sebesar 5,56%. *Ensemble method* lebih baik dalam meningkatkan akurasi dan presisi pada teknik *Majority Vote* dibandingkan dengan menggunakan metode algoritma C4.5, *Naive Bayes* dan *K-nearest Neighbor* yang telah dilakukan.

DAFTAR PUSTAKA

Budhiman, M. 2002. *Autistic spectrum disorder*. Jakarta: Yayasan Autisma Indonesia.

David Hartanto, Seng Hansun, 2014. Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa, *ULTIMATICS, Vol. VI, No.1 Juni 2014, ISSN: 2085-4552*.

Fitroh, I. 2015. *Metode K-Nearest Neighbor Berbasis Particle Swarm Optimization untuk peramalan kepadatan arus lalu lintas*. Semarang: Pasca Sarjana Universitas Dian Nuswantoro.

Gorunescu, F. 2011. *Data Mining Concept Model and Techniques*. Berlin: Springer.ISBN 978-3-642-19720-8.

Jojor Jennifer Sianipar, M. F. 2012. *Identifikasi Diagnosis Gangguan Autisme Pada Anak Menggunakan Metode Modified K-Nearest Neighbor MKNN*.

Kusnawi, 2007. *Penghantar Solusi Data Mining*. Yogyakarta : STMIK AMIKOM.

Munawaroh, Munjiati. 2013. *Manajemen Operasi*. Yogyakarta. LP3M UMY.

Rosandy, T. 2016. *Perbandingan Metode Naive Bayes Classifier dengan Metode Decision Tree Untuk Menganalisa Kelancaran Pembiayaan. Jurnal TIM Darmajaya, 02(01), 52–62*.

Saleh, Ahmad. 2015. *Klasifikasi Gejala Depresi Pada Manusia dengan Metode Naïve Bayes Menggunakan Java*, Yogyakarta.

Sunu, Christoper. 2012. *Panduan memecahkan masalah autism unlocking autism*. Yogyakarta, Lintangterbit.

Sugara, B., Widyatmoko, D., Prakoso, B. S., & Saputro, D. M. 2018. *Penerapan algoritma c4.5 untuk deteksi dini gangguan autisme pada anak. 2018(Sentika)*.

Tan, Et Al. 2007. *Intellectual capital and financial returns of companies. Journal of Intellectual Capital Vol. 8 No. 1, 2007 pp. 76-95*.

YongZhang, H. 2014. *A Weighted Voting Classifier Based on Differential Evolution*.