KLASIFIKASI GANGGUAN AUTISME PADA ANAK MENGGUNAKAN ALGORITMA C4.5 DENGANTEKNIK RANDOM FOREST

Andre Eko Cahyo, Agung Nilogiri

Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember
Jl. Karimata No.49 Jember Kode Pos 68121
Email : andreekocahyo06@gmail.com

ABSTRAK

Autisme merupakan gangguan yang terjadi pada otak, yang menyebabkan beberapa area berbeda di otak tidak mampu bekerjasama. Autisme memiliki beberapa gejala yang berbeda untuk setiap jenis gangguan yang diterima. Dalam klasifikasi memiliki beberapa metode salah satunya adalah pohon keputusan (decision tree). Data yang digunakan pada penelitian ini adalah data dari penelitian oleh (Sugara, Widyatmoko, Prakoso & Saputro, 2018). Akurasi sangat penting dalam pengklasifikasian, ensemble method adalah metode yang digunakan untuk meningkatkan akurasi algoritma klasifikasi dengan membangun beberapa classifier dari data training. Dari hasil penelitian, pada perhitungan yang telah dilakukan dengan 15 kali percobaan sebelum menggunakan teknik *Ensemble* didapatkan akurasi terbaik pada k-fold 6 percobaan ke-4 dengan akurasi sebesar 83,33%, dimana nilai positif gangguan interaksi sosial memiliki hasil 80%, nilai positif pada gangguan komunikasi sebesar 100%, dan pada nilai positif gangguan memiliki perilaku presisi 100%.Kemudian dilakukan perhitungan dengan 15 kali percobaan sesudah menggunakan teknik *Ensemble Random Forest* didapatkan akurasi terbaik pada k-fold 4 percobaan ke-3 dengan akurasi sebesar 88,89%. Dimana nilai positif gangguan `interaksi sosial memiliki hasil 80,00%, nilai positif pada gangguan komunikasi sebesar 100%, dan pada nilai positif gangguan memiliki perilaku presisi 100%. Dengan demikian pada penelitian yang telah dilakukan mendapatkan peningkatan akurasi sebesar 5,56%.

Kata Kunci :*Gangguan Autisme, Algoritma C4.5, Random Forest*

ABSTRACT

Autism is a disorder that occurs in the brain, which causes several different areas of the brain to be incapacitated. Autism has several different symptoms for each accepted type of disorder. One of the classification methods is a decision tree. The data used in this study is data from research by (Sugara, Widyatmoko, Prakoso & Saputro, 2018). Accuracy is very important in classification, the ensemble method is a method used to improve classification accuracy by building several classifiers from training data. From the results of the study, the calculations that have been done with 15 trials before using the Ensemble technique are accurate to the accuracy of the 4th k-fold 6 experiment with an accuracy of 83.33%, where the positive value of social interaction disorders has a result of 80%, a positive value on communication disorders. 100%, and on a positive value the interference has a 100% precision behavior. Then the calculations were carried out with 15 experiments after using the Ensemble Random Forest technique with accurate accuracy on the k-fold 4 3rd experiment with an accuracy of 88.89%. Where the positive value of social interaction disorders has a result of 80.00%, the positive value of communication disorders is 100%, and the positive value of disorders has 100% precision behavior. Thus, in the research that has been carried out, the increase in value is 5.56%.

Keywords: *Autism Disorders, C4.5 Algorithm, Random Forest*

Pendahuluan

Autisme merupakan gangguan yang terjadi pada otak seseorang (penderita), yang menyebabkan beberapa area berbeda di otak tidak mampu bekerjasama, sehingga penderita autisme sulit berkomunikasi dan berhubungan sosial dengan orang lain. Penyebab autisme terdiri dari beberapa faktor, namun pada umumnya terjadi karena faktor genetik dan lingkungan, gambaran umum seorang anak yang didiagnosis mengalami gangguan autisme menunjukkan kurang adanya respon terhadap orang lain, seperti mengalami kendala dalam kemampuan berbahasa dan berkomunikasi, serta memunculkan respon yang aneh terhadap lingkungan sekitarnya. Anak yang mengalami gangguan autisme juga kurang responstif terhadap emosi orang lain, kurang mampu mengendalikan perilaku dalam konteks sosial, kurang mampu menggunakan isyarat sosial seperti tertawa, senyum, dan melakukan kontak mata, hal ini berbeda dengan apa yang dilakukan pada sesorang yang normal pada umumnya, untuk menangani permaslahan mengenai maslah autisme perlu diadakannya tindakan oleh pihak dokter agar permasalahan autisme dapat segera ditangani dengan cara atau metode penanganan yang akan dilakukan oleh dokter.

Semakin berkembangnya teknologi dimasa kini dapat membantu kegiatan medis dalam menangani suatu penyakit agar dapat segera teratasi seperti antara lain, *machine learning* menjadi salah satu banyak diminati di bidang komputer. *Machine learning* dapat juga

digunakan dibidang kesehatan untuk mempredikisi dan juga mengklasifikasikan suatu penyakit dari data yang dialami pasien. Algoritma klasifikasi data mining dapat dimanfaatkan dan membantu dalam mendiagnosa suatu penyakit, salah satunya gangguan autisme pada anak. *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 random forest yang merupakan peningkatan dari algoritma C4.5 oleh karena itu pemilihan metode yang digunakan membuktikan kelebihan pada random forest mendapatkan akurasi yang lebih baik dibandingkan dengan metode klasifikasi biasa, dengan demikian pada penelitian ini akan dilakukan klasifikasi algoritma C4.5 dengan metode *Random Forest.*

Rumusan Masalah

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

1. Berapa tingkat presisi sebelum menggunakan teknik *ensemblerandom forest* dan sesudah menggunakan teknik *ensemble Random Forest?*
2. Berapa tingkat akurasi sebelum menggunakan teknik *ensembleRandom Forest* dan sesudah menggunakan teknik *ensembleRandom Forest?*

Tinjauan Pustaka

Menurut Yuwono (2009:26) autis merupakan gangguan perkembangan neurobiologis yang sangat kompleks/berat dalam kehidupan yang panjang, yang meliputi gangguan pada aspek interaksi sosial, komunikasi dan bahasa dan perilaku serta gangguan emosi dan persepsi sensori bahkan aspek motoriknya. Gejala autistik muncul pada usia sebelum 3 tahun.

Autisme merupakan gangguan perkembangan pervasive yang ciri utamanya adalah gangguan kualitatif pada perkembangan komunikasi baik secara verbal (berbicara dan menulis) dan non verbal (kurang bias mengekspresikan perasaan dan kadang menunjukkan ekspresi yang kurang tepat) Peeters, 2004). Hal ini ditandai dengan kurangnyaatau tidak adanya bahasa yang (diucapkan, tidak adanya inisiatif untuk konversasi, dan pembalikan dalam penggunaan kata terutama kata ganti (Monks, 2002: 378)).

Machine learning mempelajari teori pembelarajan suatu mesin atau komputer sehingga komputer dapat "belajar" dari data. Pembelajaran mesin melibatkan berbagai

disiplin ilmu statistik pada matematika, sains komputer, dan bahkan neurologi. Istilah pembelajaran mesin pertama didefinisikan oleh Arthur Samuel tahun 1959. Terdapat dua penerapan dalam *machine learning,* yaitu klasifikasi dan prediksi. Klasifikasi adalah metode dalam pembelajaran mesin yang digunakan oleh mesin untuk mengurutkan atau mengklasifikasikan objek berdasarkan karakteristik khusus yang mencoba untuk membedakan objek satu sama lain. Sedangkan prediksi atau regresi digunakan oleh mesin untuk memperkirakan output dari suatu input data berdasarkan data yang telah diperoleh dalam poses taining. Terdapat dua macam pendekatan dalam *machine learning* yaitu *supervised learning dan unsupervised learning.*

a. Supervised Learning

Supervised learning adalah pembelajaran menggunakan input data training yang telah diberi label atau variabel yang ditargetkan, dan mengelompokkan kedalam data yang sudah ada

b. Unsupervised Learning

Unsupervised learning adalah pembelajaran menggunakan input data pembelajaran yang tidak diberi label atau variabel yang ditargetkan, dan mengelompokkan atau membagi berdasarkan kriteria tertentu.

Dalam pendangan ini, klasifikasi dan regresi adalah dua jenis masalah prediksi, dimana klasifikasi digunakan untuk memprediksi nilai-nilai diskrit atau nominal, sedangkan regresi digunakan untuk memprediksi nilai-nilai yang kontinyu. Untuk selanjutnya penggunaan istilah prediction untuk memprediksi kelas yang berlabel disebut classification, dan penggunaan istilah prediksi untuk memprediksi nilai-nilai yang kontinyu sebagai prediction.

Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Selain itu dapat diekspresikan dalam bentuk bahasa basis data seperti Structure Query Language untuk mencari record pada kategori tertentu (Kusrini dan Emha, 2009).

C4.5 adalah algoritma yang sudah banyak dikenal dan digunakan untuk klasifikasi data yang memiliki atribut-atribut numerik dan kategorikal. Hasil dari proses klasifikasi yang berupa aturan-aturan dapat digunakan untuk memprediksi nilai atribut bertipe *diskret* dari

record yang baru. Algoritma C4.5 sendiri merupakan pengembangan dari algoritma ID3, dimana pengembangan dilakukan dalam hal bisa mengatasi *missing* data, bisa mengatasi data *kontinyu*, dan *pruning*. 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.

Algoritma random forest merupakan model klasifikasi yang dilakukan dengan mengembangkan berapa decision tree berdasarkan seleksi data dan variabel yang dilakukan secara acak.

Skema ini pertama kali dicetuskanh Leo Breiman pada tahun 2000 untuk membangun prediktor dengan sekumpulan decision tree yang berkembang secara acak pada subruang data.Kelas yang dihasilkan dari proses klasifikasi dipilih dari kelas yang paling banyak yang dihasilkan oleh pohon keputusan yang ada (Biau, 2012).

Dalam pembentukan *tree*, algoritma *Random Forest* akan melakukan training terhadap sampel data. Pengambilan sampel dilakukan dengan cara *sampling with replacement*. Variabel yang akan digunakan untuk menentukan pemisahan (*split*) terbaik ditentukan secara acak. Setelah seluruh *tree* terbentuk, maka proses klasifikasi akan berjalan. Penentuan kelas dilakukan dengan cara *voting* dari masing masing *tree*, kelas dengan jumlah *vote* terbanyak akan menjadi pemenangnya.

Cross validation adalah metode statistik yang digunakan untuk mengevaluasi dan membandingkan algoritma pembelajaran dengan cara membagi data menjadi dua bagian: satu digunakan untuk belajar atau melatih model, satu untuk menguji model tersebut (Refaeilzadeh, et al., 2009).

Confusion matrix merupakan alat pengukuran yang dapat digunakanuntuk menghitung kinerja atau tingkat kebenaran proses klasifikasi. Dengan confusion matrix dapat dianalisa seberapa baik classifier dapat mengenali record dari kelas-kelas yang berbeda. Tabel confusion matrix ditunjukkan pada tabel berikut ini:

Tabel 1 Confusion Matrix

		Prediksi	
		Positif	**Negatif**
Aktual	**Positif**	**TP**	**FN**
	Negatif	**FP**	**TN**

Keterangan :

a. TP (True Positive) merupakan banyaknya data yang kelas aktualnya adalah kelas positif dengan kelas prediksinya merupakan kelas positif.
b. FN (False Negative) merupakan banyaknya data yang kelas aktualnya adalah kelas positif dengan kelas prediksinya merupakan kelas negatif.
c. FP (False Positive) merupakan banyaknya data yang kelas aktualnya adalah kelas negatif dengan kelas prediksinya merupakan kelas positif.
d. TN (True Negative) merupakan banyaknya data yang kelas aktualnya adalah kelas negatif dengan kelas prediksinya merupakan kelas negatif.

Akurasi merupakan metode pengujian berdasarkan tingkat kedekatan antara nilai prediksi dengan nilai aktual. Dengan mengetahui jumlah data yang diklasifikasikan secara benar maka dapat diketahui akurasi hasil prediksi. Persamaan akurasi seperti pada persamaan berikut.

$$\textbf{Akurasi} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}}$$

Presisi merupakan metode pengujian dengan melakukan perbandingan jumlah informasi relevan yang didapatkan sistem dengan jumlah seluruh informasi yang terambil oleh sistem baik yang relevan maupun tidak. Persamaan presisi ditunjukkan pada persamaan berikut.

$$\textbf{Presisi} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

Metodologi Penelitian

Penerapan teknik Algoritma C4.5 dengan metode *Random Forest* untuk mengetahui akurasi terbaik dari hasil klasifikasi menggunakan teknik pada dataset gangguan autisme pada anak. Metode yang digunakan pada penelitian ini terdiri dari beberapa tahap yaitu pengumpulan data, analisis data, dan evaluasi. Diagram balok metode *Random Forest* dengan langkah-langkah sebagai berikut; masukkan dataset gangguan autisme lalu tentukan k fold,lalu bagi data menjadi data training dan data testing. Tentukan iterasi awal dan batas iterasi dengan rumus $\sqrt{jumlah\ data}$. Selanjutnya hitung entropy dan gain hingga mendapat pohon model, lalu iterasi = itersai + 1, jika iterasi masih belum lebih besar dari banyak iterasi maka akan dilakukan perhitungan ulang atau looping entropy dan gain hingga terpenuhi. Jika sudah terpenuhi maka akan menghasilkan pohon model sebanyak jumlah iterasi, kemudian gabungkan semua tree dan suara atau vote terbanyak itu yang akan menjadi klasifikasi akhir, lalu dilakukan pengujian model kepada data testing. Selama k-fold belum terpenuhi ulangi proses pembagian data hingga terpenuhi sehingga menghasilkan akurasi terbaik dengan menggunakan proses confusion matrix.

Data yang digunakan pada penelitian ini adalah data dari penelitian yang dilakukan oleh (Sugara et al., 2018). Data yang digunakan terdiri dari 24 parameter dan juga 3 output yang di hasilkan. 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

Analisis yang digunakan adalah deskriptif untuk data gejala pada gangguan autisme pada anak menggunakan software Microsoft excel 2010. Kemudian digunakan teknik *Random Forest* pada algoritma C4.5 dengan proses sebagai berikut:

Proses Random Forest

Adapun tahapan-tahapan perhitungan algoritma *Random Forest* yang akan diterapkan pada penelitian ini adalah :

a. Menyiapkan Data Training

Data training yang digunakan dalam penelitian ini menggunakan data gejala gangguan

autisme pada anak. Yang mana dalam pemilihan data dilakukan secara acak (random).

b. Membagi Data

Pada tahap ini data training akan dibagi menjadi beberapa data dengan menggunakan rumus $\sqrt{jumlah\ data}$

c. Perhitungan *Random Forest* untuk data 1

Pertama tentukan akar berdasarkan pada nilai Gain tertinggi dari atribut-atribut yang ada. Sedangkan untuk mendapatkan nilai Gain harus menentukan nilai Entropy terlebih dahulu. Dengan menggunakan rumus dasar dalam menentukan 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

Gambar 1 Hasil Perhitungan node 1 pada data 1

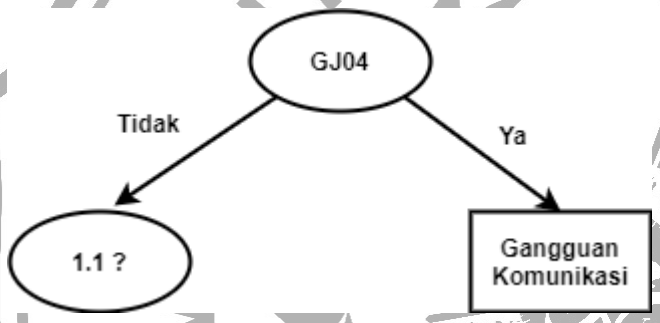
Gejala Autism		Jml Kasus	Gangguan Perilaku	Gangguan Komunikasi	Gangguan Interaksi Sosial	Entropy	Gain
Total		5	1	2	2	1,521928095	
GJ01						1,3509775	0,170950594
	Ya	2	0	1	1	1	
	Tidak	3	1	1	1	1,584962501	
GJ02						0,5509775	0,970950594
	Ya	3	1	0	2	0,918295834	
	Tidak	2	0	2	0	0	
GJ03						0,5509775	0,970950594
	Ya	3	1	0	2	0,918295834	
	Tidak	2	0	2	0	0	
GJ04						0,5509775	0,970950594
	Ya	2	0	2	0	0	
	Tidak	3	1	0	2	0,918295834	
GJ05						0,8	0,721928095
	Ya	1	1	0	0	0	
	Tidak	4	0	2	2	1	
GJ06						0,8	0,721928095
	Ya	1	1	0	0	0	
	Tidak	4	0	2	2	1	
GJ07						0,8	0,721928095
	Ya	1	1	0	0	0	

	Tidak	4	0	2	2	1	
GJ08						0,8	0,721928095
	Ya	1	1	0	0	0	
	Tidak	4	0	2	2	1	
GJ09						0,9509775	0,570950594
	Ya	2	1	0	1	1	
	Tidak	3	0	2	1	0,918295834	
GJ10						1,2	0,321928095
	Ya	1	0	1	0	0	
	Tidak	4	1	1	2	1,5	
GJ11						1,2	0,321928095
	Ya	1	0	1	0	0	
	Tidak	4	1	1	2	1,5	
GJ12						1,2	0,321928095
	Ya	1	0	1	0	0	
	Tidak	4	1	1	2	1,5	
GJ13						0,5509775	0,970950594
	Ya	2	0	2	0	0	
	Tidak	3	1	0	2	0,918295834	
GJ14						0,9509775	0,570950594
	Ya	3	0	2	1	0,918295834	
	Tidak	2	1	0	1	1	
GJ15						1,2	0,321928095
	Ya	1	0	1	0	0	
	Tidak	4	1	1	2	1,5	

GJ16						1,2	0,321928095
	Ya	1	0	1	0	0	
	Tidak	4	1	1	2	1,5	
GJ17						1,2	0,321928095
	Ya	1	0	1	0	0	
	Tidak	4	1	1	2	1,5	
GJ18						0,5509775	0,970950594
	Ya	2	0	0	2	0	
	Tidak	3	1	2	0	0,918295834	
GJ19						1,2	0,321928095
	Ya	1	0	0	1	0	
	Tidak	4	1	2	1	1,5	
GJ20						1,2	0,321928095
	Ya	1	0	0	1	0	
	Tidak	4	1	2	1	1,5	
GJ21						0,5509775	0,970950594
	Ya	2	0	0	2	0	
	Tidak	3	1	2	0	0,918295834	
GJ24						1,3509775	0,170950594
	Ya	2	0	1	1	1	
	Tidak	3	1	1	1	1,584962501	
GJ23						0,5509775	0,970950594
	Ya	2	0	0	2	0	
	Tidak	3	1	2	0	0,918295834	
GJ24						1,2	0,321928095

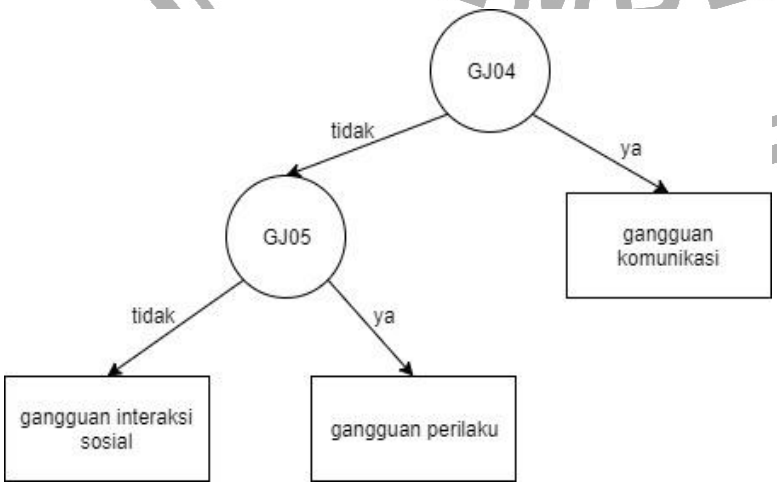
	Ya	1	0	0	1	0	
	Tidak	4	1	2	1	1,5	

Dari tabel perhitungan diatas terdapat beberapa nilai *gain* tertinggi karena itu akan dipilih salah satu untuk menjadi akar yaitu pada atribut GJ04 dengan nilai 0,970950594, dengan demikian atribut GI04 dijadikan node akar. Atribut GJ04 memiliki dua nilai yaitu ya dan tidak. Pada atribut GJ04 masih diperlukan perhitungan lanjutan di karenakan nilai "tidak" terdapat dua *output*. Dari hasil tersebut dapat digambarkan pohon keputusan sementara seperti Gambar berikut:



Gambar 2 Pohon Keputusan Sementara pada data 1

Dari gambar diatas akan dihitung entropy dan gain untuk menentukan node 1.1. Selanjutnya akan di hitung lagi pada GJ04 dengan nilai "tidak" dengan menggunakan tabel perhitungan seperti pada perhitungan sebelumnya, dengan begitu didapatkan nilai gain tertinggi pada node 1.1 yaitu GJ05. Dikarenakan hasil telah didapatkan dan nilai *entropy* 0 maka perhitungan telah selesai dan pohon keputusan akhirnya sebagai berikut :

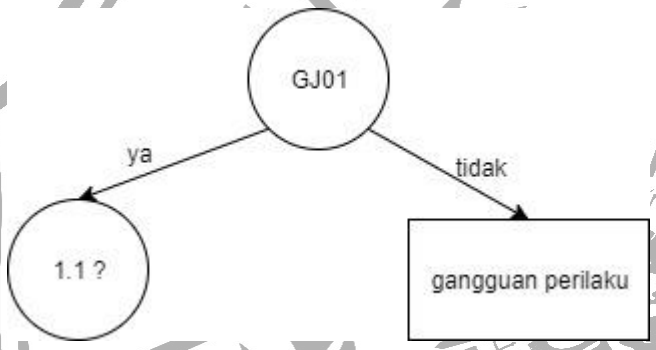


Gambar 3 Pohon keputusan Akhir pada data 1

d. Pehitungan *Random Forest* untuk data 2

Pada data 2 dilalukan langkah-langkah seperti pada data 1. Pertama akan di hitung *entropy* dan *gain* untuk menentukan akar dari pohon keputusan dari *gain* tertinggi. Rumus yang digunakan seperti pada data 1 dan perhitungan *entropy* dan *gain* sebagai berikut :

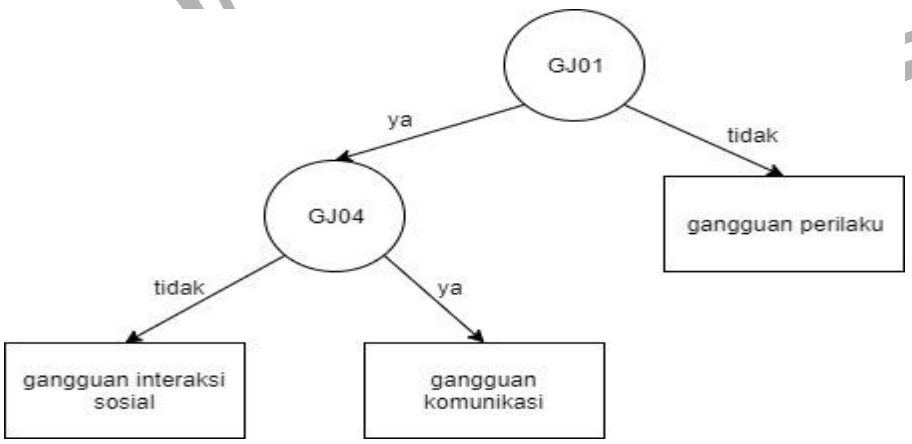
Hasil dari perhitungan node 1 pada data 2 yang dilakukan pada langkah langkah yang dilakukan sama halnya dengan data 1, didapatkan nilai gain tertinggi pada atribut GJ01. Maka atribut GJ01 dijadikan *node* akar, pada atribut GJ01 diperlukan perhitungan lanjutan dikarenakan nilai ya terdapat dua *output*, digambarkan pohon keputusan sementara seperti gambar berikut :



Gambar 4 Pohon Keputusan Sementara pada data 2

Dari gambar diatas akan di hitung *entropy* dan *gain* untuk menentukan node 1.1

Selanjutnya akan di hitung lagi pada GJ01 dengan nilai "ya" dengan menggunakan tabel perhitungan seperti pada perhitungan sebelumnya, dengan begitu didapatkan nilai gain tertinggi pada node 1.1 yaitu GJ04. Dikarenakan hasil telah didapatkan dan nilai *entropy* 0 maka perhitungan telah selesai dan pohon keputusan akhirnya sebagai berikut :



Gambar 5 Pohon Keputusan Akhir pada data 2

e. Pehitungan *Random Forest* untuk data 3

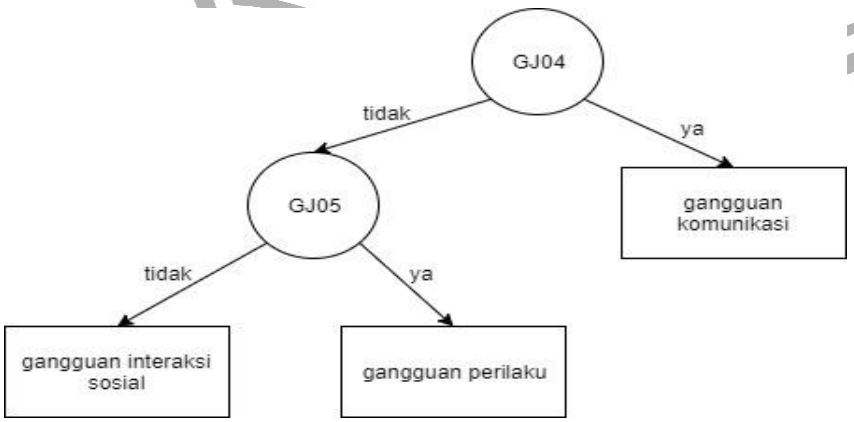
Pada data 3 dilalukan langkah-langkah seperti pada data 1 dan data 2. Pertama akan di hitung *entropy* dan *gain* untuk menentukan akar dari pohon keputusan dari *gain* tertinggi. Rumus yang digunakan seperti pada data 1 dan perhitungan *entropy* dan *gain* sebagai berikut :

Hasil dari perhitungan node 1 pada data 3 yang dilakukan pada langkah langkah yang dilakukan sama halnya dengan data 1 dan data 2, didapatkan nilai gain tertinggi pada atribut GJ04. Maka atribut GJ04 dijadikan *node* akar, pada atribut GJ04 diperlukan perhitungan lanjutan dikarenakan nilai "tidak" terdapat dua *output*, digambarkan pohon keputusan sementara seperti gambar berikut :

Gambar 6 Pohon Keputusan Sementara pada data 3

Dari gambar diatas akan di hitung *entropy* dan *gain* untuk menentukan node 1.1

Selanjutnya akan di hitung lagi pada GJ04 dengan nilai "tidak" dengan menggunakan tabel perhitungan seperti pada perhitungan sebelumnya, dengan begitu didapatkan nilai gain tertinggi pada node 1.1 yaitu GJ05. Dikarenakan hasil telah didapatkan dan nilai *entropy* 0 maka perhitungan telah selesai dan pohon keputusan akhirnya sebagai berikut :



Gambar 7 Pohon Keputusan Akhir pada data 3

Evaluasi

Evaluasi akan dilakukan dengan membandingkan hasil klasifikasi pada 3 pohon keputusan 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 et al., 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 di ambil 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 pengaplikasiaannya. Sebelum digunakan pada *toolRapid miner* data terlebih dahulu di bagi menjadi data *training* dan data *testing*. Dimana dengan jumlah data yang berbeda sesuai dengan k-fold cross validation. . Data akan di bagi 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	2fold percobaan ke1	1 – 18	50%	19 – 36	50%
2	2fold percobaan ke2	19 – 36	50%	1 – 18	50%
3	3fold percobaan ke1	1 – 24	66%	25 - 36	34%
4	3fold percobaan ke2	1 – 12 & 25 – 36	66%	13 – 24	34%
5	3fold percobaan ke3	13 – 36	66%	1 – 12	34%
6	4fold percobaan ke1	1 – 27	75%	28 -36	25%
7	4fold percobaan ke2	1 – 18 & 28 – 36	75%	19 – 27	25%
8	4fold percobaan ke3	1 – 9 & 19 – 36	75%	10 – 18	25%
9	4fold percobaan ke4	10 – 36	75%	1 – 9	25%
10	6fold percobaan ke1	1 – 30	83%	31 - 36	17%
11	6fold percobaan ke2	1 – 24 & 31 – 36	83%	25 – 30	17%
12	6fold percobaan ke3	1 – 18 & 25 – 36	83%	19 – 24	17%
13	6fold percobaan ke4	1 – 12 & 19 – 36	83%	13 – 18	17%
14	6fold percobaan ke5	1 – 6 & 13 – 36	83%	7 – 12	17%
15	6fold	7 – 36	83%	1 – 6	17%

	percobaan ke6				

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 Algoritma C4.5

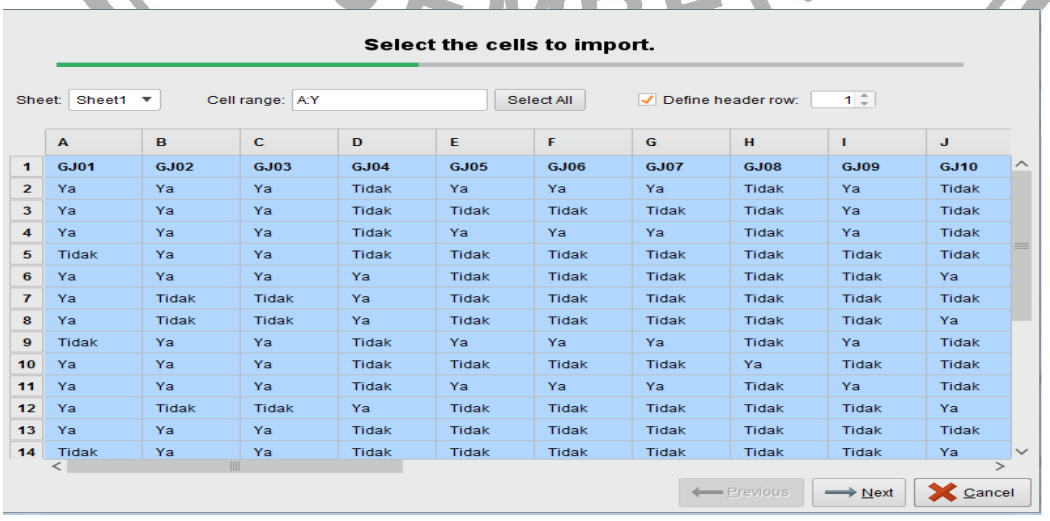
Pada hasil pengujian ini yang akan perlihatkan adalah hasil terbaik dari *cross validation* sebelum menggunakan teknik *Ensemble,*didapatkan akurasi terbaik pada k-fold 6 percobaan ke 4 sebesar 83,33%, denganpresisi dimana nilai positif gangguan interaksi sosial memiliki hasil 100%, nilai positif pada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

Hasil Pengujian Teknik *Random Forest*Pada Algoritma C4.5

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

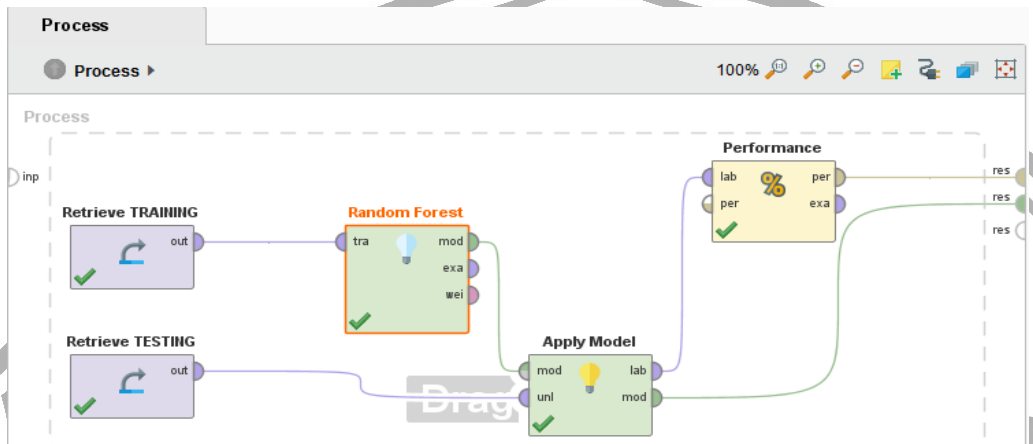
Berdasarkan skenario uji pada pada pengujian dengan k = 2 yaitu membagi data latih (*training)* sebanyak 18 dan sisanya menjadi data uji (*testing)*. Setelah semua data telah disiapkan maka masukan data kedalam *Rapidminer* untuk diproses, berikut proses pengolahan data menggunakan teknik *RandomForest* pada algoritma C4.5. langkah-langkah dalam implementasi *Rapidminer* 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 8 Proses *Import* Data

Pada proses ini data yang di import sesuai dengan pembagian dan untuk atribut gangguan yang di dalam*role* dirubah menjadi "label" agar *Rapidminer* mengetahui sebagai output yang digunakan. Lalu dilakukan pemodelan seperti berikut :



Gambar 9 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 Komuni...	true Gangguan Interaks...	true Gangguan Perilaku	class precision
pred. Gangguan Komu...	4	0	0	100.00%
pred. Gangguan Interak...	0	4	0	100.00%
pred. Gangguan Perilaku	3	2	6	54.55%
class recall	57.14%	66.67%	100.00%	

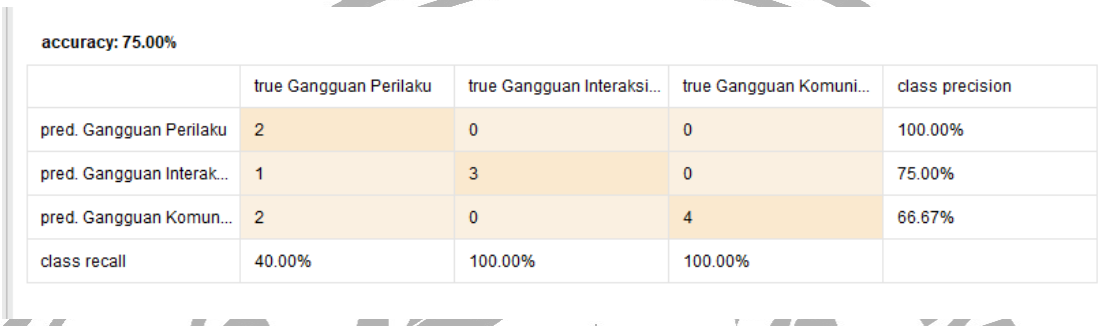
Gambar 10 Hasil Akurasi Dan Presisi Percobaan Ke-2

Pada hasil pengujian pada gambar 6 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.

Pada gambar 10 adalah akurasi terbaik pada 2-fold *cross validation*,dimana hasil itu didapatkan pada percobaan ke-1. Menggunakan data uji sebanyak 18. Dimana akurasi yang didapatkan sebesar 73,68% dan presisi yang didapatkan sebanyak 3 dimana nilai positif gangguan`interaksi sosial memiliki hasil 100%, nilai positif pada gangguan komunikasi sebesar 100%, dan pada nilai positif gangguan memiliki perilaku presisi 54.55%.

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 kedalam *Rapidminer* untuk diproses. Proses dilakukan sama dengan pengujian 2-fold. Akurasi dan presisi yang didapatkan pada 3-fold sebagai berikut:



accuracy: 75.00%

	true Gangguan Perilaku	true Gangguan Interaksi...	true Gangguan Komuni...	class precision
pred. Gangguan Perilaku	2	0	0	100.00%
pred. Gangguan Interak...	1	3	0	75.00%
pred. Gangguan Komun...	2	0	4	66.67%
class recall	40.00%	100.00%	100.00%	

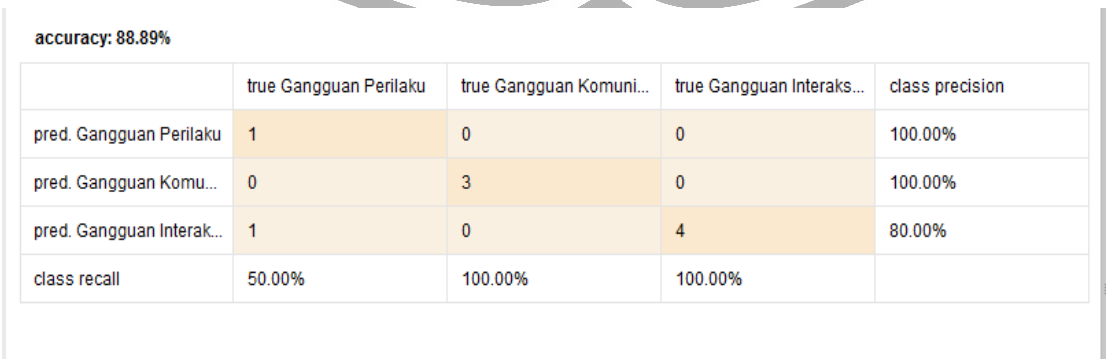
Gambar 11 Hasil Akurasi Dan Presisi Percobaan Ke-3

Pada hasil pengujian nilai dari akurasi langsung dihitung dengan menggunakan 3 output tanpa merubah data dan mendapatkan akurasi sebesar 75,00%. Maka akan dihitung menggunakan confusion matrix 2 output dengan 3 hasil,

Pada gambar 11 adalah akurasi terbaik pada 3-fold *cross validation*, dimana hasil itu didapatkan pada percobaan ke-2. Menggunakan data uji sebanyak 12. Dimana akurasi yang didapatkan sebesar 75,00% dan presisi yang didapatkan sebanyak 3 dimana nilai positif gangguan `interaksi sosial memiliki hasil 75,00%, lalu jika nilai positif pada gangguan komunikasi sebesar 66,67%, dan pada nilai positif gangguan memiliki perilaku presisi 100%.

Hasil Uji 4-Fold

Berdasarkan scenario 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 *Rapidminer* 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 Komuni...	true Gangguan Interaks...	class precision
pred. Gangguan Perilaku	1	0	0	100.00%
pred. Gangguan Komu...	0	3	0	100.00%
pred. Gangguan Interak...	1	0	4	80.00%
class recall	50.00%	100.00%	100.00%	

Gambar 12 Hasil Akurasi Dan Presisi Percobaan Ke-4

Pada hasil pengujian pada gambar 7 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.

Pada gambar 12 adalah akurasi terbaik pada 4-fold *cross validation*. Menggunakan data uji sebanyak 9. Dimana akurasi yang didapatkan sebesar 88,89% dan presisi yang didapatkan sebanyak 3 dimana nilai positif gangguan interaksi sosial memiliki hasil 80%, nilai positif pada gangguan komunikasi sebesar 100%, dan pada nilai positif gangguan memiliki perilaku presisi 100%.

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: 83.33%

	true Gangguan Interaks...	true Gangguan Komuni...	true Gangguan Perilaku	class precision
pred. Gangguan Interak...	3	0	1	75.00%
pred. Gangguan Komu...	0	2	0	100.00%
pred. Gangguan Perilaku	0	0	0	0.00%
class recall	100.00%	100.00%	0.00%	

Gambar 13 Hasil Akurasi Dan Presisi Percobaan Ke-6

Pada hasil pengujian pada gambar 4.6 nilai dari akurasi langsung dihitung dengan menggunakan 3 output tanpa merubah data dan mendapatkan akurasi sebesar 83,33 %. Maka akan dihitung menggunakan *Confusion matrix* 2 output dengan 3 hasil.

Pada gambar 13 adalah akurasi terbaik pada 6-fold *cross validation*. Menggunakan data uji sebanyak 6. Dimana akurasi yang didapatkan sebesar 83,33% dan presisi yang didapatkan sebanyak 3 dimana nilai positif gangguan `interaksi sosial memiliki hasil 75.00%, nilai positif pada gangguan komunikasi sebesar 100%, dan pada nilai positif gangguan memiliki perilaku presisi 0.00%%

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 80,00%, nilai positif pada gangguan komunikasi sebesar 100%, dan pada nilai positif gangguan memiliki perilaku 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%	100%	100 %	54,55%	75,00%	66,67%	100 %	66,67%	66,67%	100 %	100 %	40,00%	66,67%	100 %	100 %	33,33%
2	72,22%	71,43 %	75,00%	66,67%	61,54%	80,00%	100 %	33,33%	77,78%	100 %	0%	66,67%	66,67%	50%	100 %	100 %
3					75,00%	75,00%	66,67%	100 %	88,89%	80,00%	100 %	100 %	50,00%	100 %	0%	40,00%
4									66,67%	66,67%	60,00%	100 %	83,33%	75,00%	100 %	0%
5													66,67%	50,00%	100 %	50,00%
6													50,00%	50,00%	50,00%	0%

Kesimpulan

Berdasarkan penelitian yang telah dilakukan dapat di ambil 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 80%, lalu jika nilai positif pada gangguan komunikasi sebesar 100%, dan pada nilai positif gangguan memiliki perilaku presisi 100%. Pada perhitungan yang telah dilakukan

sebelum menggunakan teknik *Ensemble*didapatkan akurasi terbaik pada k-fold 6 percobaan ke-4 dengan akurasi sebesar 83,33%, dimana nilai positif gangguan interaksi sosial memiliki hasil 80%, nilai positif pada gangguan komunikasi sebesar 100%, dan pada nilai positif gangguan memiliki perilaku presisi 100%. *Ensemble method* lebih baik dalam meningkatkan akurasi dan presisi khususnya pada teknik *Random Forest* dibandingkan dengan menggunakan metode algoritma C4.5 yang telah dibuktikan pada penelitian yang telah dilakukan.

DAFTAR PUSTAKA

Biau, G. 2012. Analysis of A Random Forest Model. Journal of Machine Learning Research 13. Paris : *Universite Pierre et Marie Curie.*

Kusrini, luthfi taufiq Emha, (2009), Algoritma Data Mining, Penerbit Andi, Yogyakarta.

Peeters, T. (2004). Autisme: Hubungan Pengetahuan Teoritis dan Intervensi Pendidikan Bagi Penyandang. Jakarta: Dian Rakyat.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In L. LIU & M. T. ÖZSU (Eds.), Encyclopedia of Database Systems (pp. 532–538).https://doi.org/10.1007/978-0-387-39940-9_565.

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

Yuwono, J. 2009. Memahami Anak Autistik. Bandung: CV Alfabeta