OPTIMASI ALGORITMA C4.5 MENGGUNAKAN TEKNIK *BAGGING* PADA DATA KADAR KARAT EMAS

Siti Mutmainnah[1)], Ginanjar Abdurrahman[2)], Habibatul Azizah Al Faruq[3)]
Email: Stmutmainnah71@gmail.com [1)], abdurrahmanginanjar@unmuhjember.ac.id[2)], habibatul azizah@unmuhjember.ac.id [3)]

ABSTRAK

Emas merupakan salah satu logam mulia yang sangat diminati dikalangan masyarakat baik sebagai perhiasan maupun sebagai penimbun kekayaan. Untuk menentukan kualitas emas membutuhkan waktu yang lama. Dengan ini data mining dapat dimanfaatkan untuk mengklasifikasikan suatu kadar karat emas. Salah satu metode yang dapat digunakan adalah Algoritma C4.5. Namun metode ini memiliki tingkat akurasi yang kurang tinggi untuk itu digunakan teknik *bagging* guna meningkatkan akurasi dari Algoritma C4.5. Dari hasil penelitian, dengan menerapkan teknik *bagging* untuk klasifikasi berbasis *ensemble* pada algoritma C4.5 dapat meningkatkan akurasi sebesar 12.86%. Dengan akurasi awal 80%, setelah diterapkan teknik *bagging* menjadi 92.86%. sedangkan untuk presisinya terbagi menjadi 2 dimana nilai positif ≥ 22 memiliki hasil 80% meningkat sebanyak 3,33% menjadi 83,33%, lalu jika nilai positif pada < 22 sebesar 88,89% meningkat sebanyak 11,11% menjadi 100%.

Kata Kunci: kadar karat, emas, Algoritma *C4.5, Bagging.*

Abstract

Gold is one of the precious metals that is in great demand among the public both as jewelery and as a hoarder of wealth. To determine the quality of gold requires a long time. With this data mining can be used to classify a karat gold content. One method that can be used is the C4.5 Algorithm. However, this method has a level of accuracy that is not high enough for it to use bagging techniques to improve the accuracy of the C4.5 Algorithm. From the results of the study, by applying bagging techniques for ensemble-based classification on the C4.5 algorithm can increase accuracy by 12.86%. With an initial accuracy of 80%, after applying the bagging technique to 92.86%. while for precision it is divided into 2 where a positive value ≥ 22 has an 80% yield increasing by 3.33% to 83.33%, then if a positive value at <22 of 88.89% increases by 11.11% to 100%.

Keywords: carat content, gold, C4.5 Algorithm, Bagging.

I. PENDAHULUAN

Mendengar tentang emas tentu terlebih dahulu harus memahami tentang kadar dalam emas, kadar merupakan suatu tingkat keaslian pada emas atau bisa disebut jumlah kandungan kemurnian emas yang dinyatakan dalam karat sedangkan emas merupakan sebuah logam mulia yang banyak didambakan oleh manusia, dikategorikan sebagai logam mulia karena emas memiliki karakter yang unik sehingga emas lebih menarik dari logam lainnya dan harganya terus naik tiap waktunya (Apriyanti, 2011).

Seiring dengan perkembangan ilmu pengetahuan dan teknologi informasi, kehadiran *machine learning* di bidang komputer telah menarik banyak perhatian. *Machine learning* menjadi sebuah dari penggunaan data itu sendiri. *Machine learning* memainkan peran luas dalam pengembangan terutama dalam pengembangan data analitik (Alarifi & Young, 2018). Salah satu metode yang ada pada *machine learning* yaitu klasifikasi. *Decision tree* merupakan salah satu metode klasifikasi yang umum

digunakan. Salah satunya menggunakan metode C4.5.

Pada bidang ini, *machine learning* dapat dimanfaatkan mengklasifikasikan suatu kadar karat emas. Algoritma klasifikasi *machine learning* tersebut dapat dimanfaatkan dan membantu orang yang mempunyai emas untuk menentukan kadar karatnya tersebut.

Pada penelitian yang dilakukan oleh (Septiani, 2016) dengan judul "Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Hepatitis". untuk hasil Algoritma C4.5 menghasilkan akurasi 77,29% dan nilai AUC 0,846 yang termasuk dalam *Good Classification*.

Ensemble method adalah metode yang digunakan untuk meningkatkan akurasi klasifikasi dengan membangun beberapa *classifier* dari data *training* kemudian pada saat klasifikasi metode ini menggunakan *voting/aggregating* dari *classifiers-classifiers* tersebut salah satu contohnya yaitu *ensemble method* adalah *boostrapp aggregating* yang biasa disingkat "*bagging*".

Pada penelitian yang dilakukan oleh (Prasetio & Pratiwi, 2015) menggunakan algoritma C4.5 dan C4.5 berbasis *bagging* yang akan digunakan untuk menganalisa data medis. Pada penelitiannya algoritma C4.5 *information gain* mendapatkan akurasi 73,7% dan saat digunakan *bagging* pada algoritma C4.5 *information gain* akurasinya meningkat 2,6% menjadi 76,3%. Dengan ini pada penelitian yang akan dilakukan diharapkan dapat meningkatkan akurasi dari algoritma C4.5 pada data kadar karat emas dengan menggunakan teknik *bagging*.

II. TINJAUAN PUSTAKA

A. Algoritma C4.5

Secara umum Algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut :

1. Mempersiapkan data *training*, dapat diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokan dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon dengan menghitung nilai *gain* yang tertinggi dari masing-masing atribut atau berdasarkan nilai index *entropy* terendah. Sebelumnya dihitung terlebih dahulu nilai *index entropy*, dengan rumus:

$$Entropy(i) = -\sum_{j=1}^{m} f(i,j).log2\, f[(i,j)]$$

Keterangan:
i = Himpunan Kasus
m = Jumlah partisi "i"
$f(i,j)$ = proposi j terdapat i

3. Hitung nilai *gain* dengan rumus:

$$Entropy\ split = \sum_{i=1}^{p} \frac{n1}{n} . IE(i)$$

Keterangan:
p = jumlah "p"artisi atribut
ni = proporsi "ni" terhadap "i"
n = jumlah kasus dalam "n "

4. Ulangi langkah ke-2 hingga semua *record* terpartisi proses partisi pohon keputusan akan berhenti disaat:
   a. Semua tabel dalam *record* dalam simpul m mendapat kelas yang sama.
   b. Tidak ada atribut dalam *record* yang dipartisi lagi.
   c. Tidak ada *record* di dalam cabang yang kosong.

B. Klasifikasi

Klasifikasi merupakan proses menemukan sebuah model atau fungsi yang mendeskripsikan dan membedakan data ke dalam kelas-kelas. Klasifikasi melibatkan proses pemeriksaan karakteristik dari objek dan memasukkan objek ke dalam salah satu kelas yang sudah didefinisikan sebelumnya (Han dan Kamber, 2006).

Menurut Han dan Kamber (2006) secara umum, klasifikasi terdiri dari dua tahap. Tahap pertama yaitu *learning* (proses belajar), merupakan sebuah model dibuat untuk menggambarkan himpunan kelas atau konsep data yang telah ditentukan sebelumnya. Model tersebut dibangun dengan menganalisa *record-record* pada basis data yang digambarkan dalam bentuk atribut. Setiap *record* diasumsikan masuk ke dalam suatu kelas yang telah ditentukan sebelumnya, yang dinamakan atribut kelas. Model itu sendiri bisa berupa aturan *IF-THEN, decision tree,* formula matematis atau *neural network*.

C. Algoritma-algoritma dalam *decision tree*

Ada banyak algoritma pada klasifikasi *decision tree* ini yang dapat dipakai dalam pembentukan pohon keputusan, suatu algoritma biasanya dikembangkan untuk meningkatkan kinerja algoritma yang sudah ada beberapa algoritma antara lain ID3, CART, dan C4.5 Algoritma C4.5 merupakan pengembangan dari algoritma ID3(Larose, 2005).

D. *Cross validation*

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, dkk. 2008).

Salah satu bentuk dari *cross validation* adalah K-*fold cross validation,* dalam metode K-*fold cross validation.* Data dibagi ke dalam beberapa partisi yang disebut dengan *fold.* Masing-masing *fold* memiliki jumlah data dengan ukuran yang sama atau mendekati sama. Selama K iterasi, dipilih salah satu *fold* sebagai data uji, sedangkan sisa K-1 *fold* dijadikan data latih (Refaeilzadeh, dkk. 2008).

E. *Bagging*

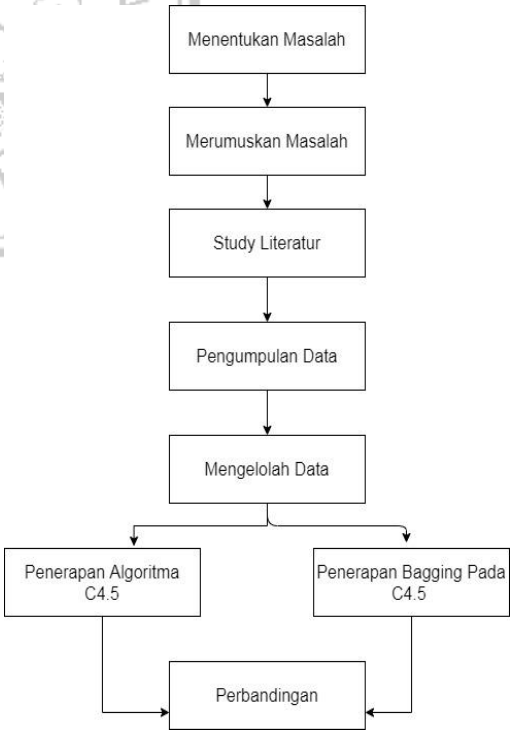
Bagging adalah singkatan dari *bootstrap aggregating*, menggunakan subdataset (*bootstrap*) untuk menghasilkan set pelatihan L (*learning*), L melatih dasar belajar menggunakan prosedur pembelajaran yang tidak stabil, dan kemudian, selama pengujian, mengambil rata-rata (Breiman, 1996). *Bagging* baik digunakan untuk klasifikasi dan regresi.

Model ensemble berlaku dengan penggabungan berbagai teknik pengambilan sampel seperti *bagging*, *boosting*, dan lain-lain untuk menjamin keragaman di kolom *classifier*. Penanganan noise data merupakan masalah penting untuk proses pembelajaran klasifikasi, sejak terjadinya noise yang tinggi dalam proses pelatihan atau pengujian (klasifikasi) pada dataset, mempengaruhi keakuratan prediksi pengklasifikasi yang dipelajari.

III. METODOLOGI PENELITIAN

a. Perancang Metode Penelitian

Rancangan metode penelitian yang dilakukan dalam penelitian ini adalah sebagai berikut, yang mana memiliki tahapan – tahapan:

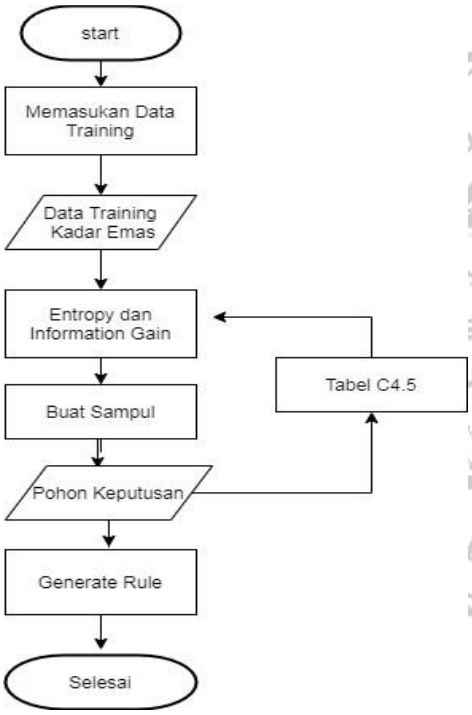


Gambar 1 Metode Penelitian

b. Jenis Penelitian

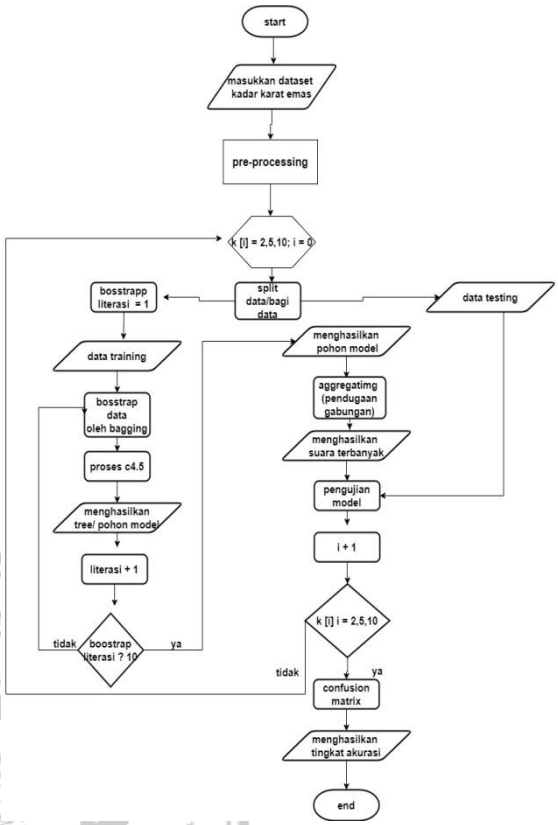
Penjelasan mengenai tahapan-tahapan dalam algoritma C4.5 menggunakan teknik bagging untuk menentukan kadar karat emas. Metodelogi yang digunakan pada penelitian ini terdiri dari beberapa tahapan yaitu study pendahuluan, pengumpulan data, *pre-processing* analisis data, perancangan sistem dan penarikan kesimpulan.

Flowchart pengujian pada penilitan ini diperuntukkan untuk prosess algoritma C4.5 menggunakan teknik *bagging* yang berisi tahapan-tahapan yang dilakukan dalam penelitian ini ditunjukkan pada gambar 3.2 di bawah ini.



Gambar 2 *Flowchart* pengujian Algoritma C4.5

Pertama memulai perhitungan *information gain* dengan memasukkan data *training* kadar karat emas kemudian menentukan akar dari pohon dengan mengitung *entropy* dan *gain* pembentukan simpul yang berisi atribut tersebut ulangi perhitungan *information gain* terus dilakukan sampai semua data telah termasuk dalam kelas yang sama. Atribut yang telah dipilih tidak diikutkan lagi dalam perhitungan nilai *information gain.*



Gambar 3 *Flowchart* Pengujian

Pertama memulai dengan memasukkan dataset kadar karat emas lalu di *pre-processing* kemudian membagi K-*fold* setelah itu data dibagi 2 dataset yang pertama data *training* yang kedua data testing untuk data *training* yang pertama menentukan iterasi awal setelah itu di proses *bootstrap* data oleh *bagging* selanjutnya proses perhitungan Algoritma C4.5 kemudian akan menghasilkan *tree/*pohon keputusan lalu iterasi awal tadi ditambah 1 jika *bootstrap* tidak sesuai yang ditentukan diawal maka proses akan diulang kembali pada *bootstrap* data oleh *bagging* jika hasil *bootstrap* sesuai dengan iterasi awal maka proses akan dilanjutkan dengan menggabungkan semua hasil *tree* kemudian setelah digabungkan maka tentukan suara terbanyak untuk dijadikan klasifikasi akhir pada bagian pengujian model digabung *output* data *testing* dengan klasifikasi akhir untuk bisa mengitung *confusion matrix* yang nantinya

bisa menghasilkan akurasi yang didapatkan.

c. Pengumpulan Data

Data yang digunakan pada penelitian ini adalah data kadar karat emas dari penelitian oleh (Prakoso & Sutanto, 2018). Data ini terdiri dari 5 parameter dengan 2 *output* yang dihasilkan nantinya. Data yang digunakan 120 *record.* Data yang akan digunakan untuk contoh perhitungan manual adalah 20 *record* yang 15 *record* untuk data *training* dan untuk 5 data digunakan untuk *testing.*

d. Implementasi *K-Fold Cross Validation*

K-Fold Cross Validation adalah teknik validasi dengan membagi data secara acak kedalam k bagian. Dengan data berjumlah 100 record, data tersebut di bagi menjadi 2 bagian yaitu 80% sebagai data training dan 20% sebagai data testing kemudian data tersebut diproses menggunakan metode K-Fold Cross Validation. Dalam penelitian ini nilai K-Fold yang digunakan sebanyak 2, 5, 7, dan 10 bagian karena untuk dipartisi menjadi data training dan data testing. Pemilihan model disesuaikan dengan kebutuhan sesuai dengan tujuan penelitian.

e. Confusion Matrix

Confusion matrix merupakan dataset yang hanya memiliki dua kelas, kelas yang satu sebagai positif dan kelas yang lain sebagai negatif. *Confusion matrix* berisi informasi perbandingan tabel hasil klasifikasi dengan tabel sebenarnya. Terlihat pada gambar tabel dibawah ini:

Tabel 1 Confusion Matrix

Classification		*Classification Class*	
Real Class	*Class = Yes*	TP	FN
	Class = No	FP	TN

Keterangan:

TP : Hasil prediksi positif dengan kelas sebenarnya positif

FN : Hasil prediksi negatif dengan kelas sebenarnya positif

FP : Hasil prediksi negatif dengan kelas sebenarnya negatif

TN : Hasil prediksi negatif dengan kelas sebenarnya negatif

Untuk menghitung akurasi digunakan persamaan dibawah ini:

$$\text{Presisi} = \frac{\text{TP}}{\text{FP}+\text{TP}} \; X \; 100\%$$

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

IV. IMPLEMENTASI DAN PENGUJIAN

a. Implementasi bagging menggunakan algoritma C4.5 pada Rapidminer

Tahap implementasi tools Rapidminer. RapidMiner merupakan perangakat lunak yang bersifat terbuka (open source). RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. Maka dilakukan implementasi Rapidminer untuk teknik *bagging* pada algoritma C4.5 pada data gangguan autisme.

b. Pengujian Data

Pada penelitian ini data yang sudah dikumpulkan akan dilakukan preprocessing data dimana tahap ini dilakukan proses *balancing* data dimana data yang tidak b*alance* dijadikan *balance*. Data yang akan digunakan sebagai pengujian pada tugas akhir ini sebanyak 70 data. Sebelumnya data akan dibagi menjadi 4 kategori pengujian data yaitu pembagian pengujian dengan data latih dan data uji yang berbeda-beda :

1. Hasil Uji 2-Fold

Pada hasil pengujian k = 2 mendapatkan akurasi sebesar 91.43% dan presisi yang didapatkan sebanyak 2 dimana nilai positif < 22 memiliki hasil 100%, lalu jika nilai positif > 22 sebesar 86.96%,

2. Hasil Uji 5-Fold

Pada hasil pengujian k = 4 mendapatkan akurasi sebesar 92.86% dan presisi yang didapatkan sebanyak 2 dimana nilai positif < 22 memiliki hasil 100%, lalu jika nilai positif > 22 sebesar 83.33%,

3. Hasil Uji 7-Fold

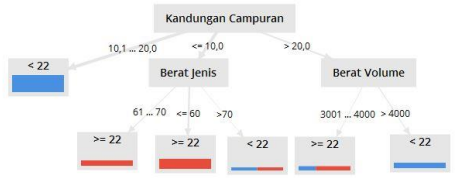
Pada hasil pengujian k = 5 mendapatkan akurasi sebesar 90% dan presisi yang didapatkan sebanyak 2 dimana nilai positif < 22 memiliki hasil 100%, lalu jika nilai positif > 22 sebesar 80%,

4. Hasil Uji 10-Fold

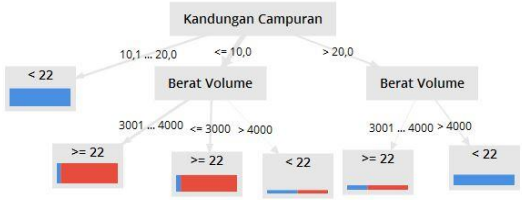
Pada hasil pengujian k = 8 mendapatkan akurasi sebesar 91.43% dan presisi yang didapatkan sebanyak 2 dimana nilai positif < 22 memiliki hasil 100%, lalu jika nilai positif > 22 sebesar 86.96%,

c. *Tree* penerapan algoritma C4.5

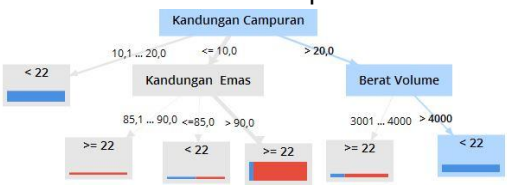
Melalui penelitian ini telah menghasilkan model keputusan dari 24 percobaan yang di setiap K-*fold* memiliki percobaan terbaik diantaranya bisa dilihat gambar dibawah ini:



Gambar 4 *Tree* K-*fold* 2 percobaan ke-2



Gambar 5 *Tree* K-*fold* 5 percobaan ke-4



Gambar 6 *Tree* K-*fold* 7 percobaan ke-5

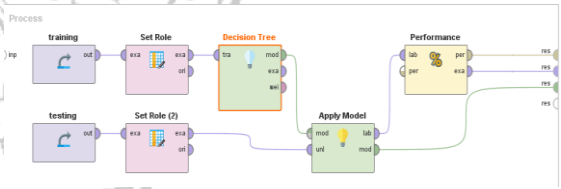
Gambar 7 *Tree* K-*fold* 10 percobaan ke-8

Tabel 2 Daftar Hasil Akurasi Dan Presisi Algoritma C4.5

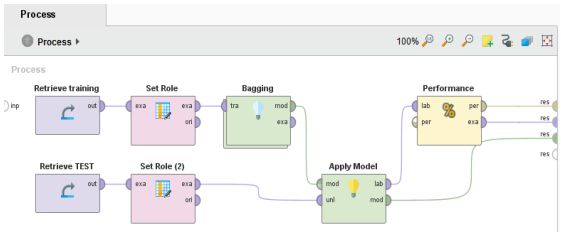
C4.5			
Pengujian ke	K = 2		
	Akurasi	Presisi	
		≥ 22	< 22
2	82.86%	85%	80%
K = 5			
4	85.71%	80%	88.89%
K = 7			
5	80%	75%	83.33%
K = 10			
8	71.43%	100%	66.67%

Tabel 3 Daftar Hasil Akurasi dan Presisi *bagging* pada C4.5

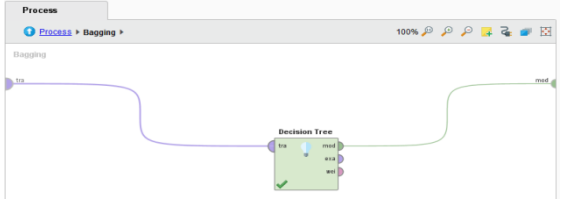
Bagging			
Pengujian ke	K = 2		
	Akurasi	Presisi	
		≥ 22	< 22
2	91.43%	86.9%	100%
K = 5			
4	92.86%	83.33%	100%
K = 7			
5	90%	80%	100%
K = 10			
8	85.71%	100%	80%



Gambar 8 Proses C4.5



Gambar 9 Proses Pengujian



Gambar 10 Proses *Bagging*

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan penelitian yang telah dilakukan dapat di ambil kesimpulan sebagai berikut:

1. Pada penelitian yang dilakukan Dari 24 kali percobaan terbaik pada K-*fold* 5 percobaan 4 dimana akurasi pada algoritma C4.5 adalah 80% Dengan menggunakan teknik *bagging* akurasi tersebut mengalami peningkatan 10% dimana hasil akhir akurasi menjadi 92,86% sedangkan untuk presisinya terbagi menjadi 2 dimana nilai positif ≥ 22 memiliki hasil 80% meningkat sebanyak 3,33% menjadi 83,33%, lalu nilai positif pada < 22 sebesar 88,89% meningkat sebanyak 11,11% menjadi 100%.
2. Dapat disimpulkan bahwa *ensemble method* sangat baik dalam meningkatkan akurasi dan presisi khususnya pada teknik bagging menggunakan klasifikasi algoritma C4.5 yang telah dibuktikan pada penelitian yang telah dilakukan.

B. Saran

Beberapa saran yang dapat dijadikan pertimbangan untuk penelitian selanjutnya yaitu:

1. Penelitian selanjutnya dapat menggunakan metode klasifikasi atau *ensamble method* lainnya untuk mengetahui metode klasifikasi seperti contohnya *Ensemble Method Voting* & *Boosting*.
2. Pada penelitian berikutnya dapat dikembangkan agar menggunakan *platform* web atau android agar dapat diakses oleh orang-orang yang membutuhkan dan berguna sebagai pengetahuan bagi masyarakat.

DAFTAR PUSTAKA

Apriyanti. 2011. *Anti Rugi dengan Berinvestasi Emas*. Yogyakarta: Pustaka Baru

Alarifi, G. S, & Young, H. S. 2018. *Using Multiple Machine Learning Algorithms to Predict Autism in Children Int'l Conf.* Artificial Intelligence | ICAI'18 |.

Septiani, D. W. 2014. Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Hepatitis. Vol. XI No. 1, Maret 2014 *Jurnal Techno Nusa Mandiri.*

Prasetio, R. T. & Pratiwi. 2015 *Penerapan Teknik Bagging Pada Algoritma Klasifikasi Untuk Mengatasi Ketidakseimbangan Kelas Dataset Medis.* INFORMATIKA. Vol. II, No. 2.

Larose, D. T. 2005. *Discovering knowledge in data an introduction to data mining.* United State of America: Jhon Wiley & Sons Inc.

Breiman, L. 1996. *Bagging Predictors. Machine Learning,* 123-140.