Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner

Ardiyansyah[1], Panny Agustia Rahayuningsih[2], Reza Maulana[3]
Program Studi Komputerisasi Akuntansi, AMIK BSI Pontianak[1]
Jl. Abdurahman Saleh No.18A, Kota Pontianak,Indonesia
e-mail : ardiyansyah.arq@bsi.ac.id

Program Studi Komputerisasi Akuntansi, AMIK BSI Pontianak [2]
Jl. Abdurahman Saleh No.18A, Kota Pontianak,Indonesia
e-mail: panny.par@bsi.ac.id

Program Studi Komputerisasi Akuntansi, AMIK BSI Pontianak [3]
Jl. Abdurahman Saleh No.18A, Kota Pontianak,Indonesia
e-mail: reza.rza@bsi.ac.id

ABSTRAKSI

Data mining merupakan sebuah proses untuk menganalisa sebuah kasus untuk menemukan performa terbaik dari algoritma yang diuji. Salah satu cara untuk mendapatkan informasi atau pola dari kumpulan data yang besar adalah dengan menggunakan teknik-teknik dalam data mining. Ada banyak metode klasifikasi yang di gunakan untuk menghasilkan nilai akurasi yang akurat. Terdapat 5 algoritma klasifikasi yang digunakan dalam mengklasifikasi dataset blogger yaitu decision tree, Naïve bayes, k-nearest neighbour, ID3, dan CHAID. Dataset menggunakan data blogger dari *UCI Machine Learning Repository*.Blog adalah media yang bergantung pada teknologi informasi dan kemajuan teknologi. Penelitian ini diuji Dengan menggunakan validasi *10-fold cross validation* dan uji *t-test*. Sehingga hasil tertinggi dari nilai akurasi yang didapat adalah sebesar 85.00% untuk algoritma KNN. Sedangkan untuk nilai AUC algoritma CHAID yang memiliki hasil tertinggi yaitu sebesar 0.758. dan dari asil uji t-test yang dilakukan bahwa algoritma ID3,CHAID dan Naive Bayes merupakan algoritma dengan performa terbaik yang diterapkan pada dataset blogger. Sedangkan untuk algoritma KNN dan C45 merupakan algoritma dengan performa yang kurang baik dengan nilai AUC 0,500%.

Kata Kunci: Data Mining, Algoritma, Klasifikasi, Dataset Blogger

ABSTRACT

Data mining is a process to analyze a case to find the best performance of the tested algorithm. One way to get information or patterns from large data sets is to use the techniques in data mining. There are many methods of classification that are used to produce accurate accuracy values. There are 5 classification algorithms used in classifying the blogger dataset of decision tree, Naïve bayes, k-nearest neighbor, ID3, and CHAID. The dataset uses blogger data from UCI Machine Learning Repository.Blog is a medium that relies on information technology and technological advancements. This study was tested by using 10-fold validation validation and t-test. So the highest result of the obtained accuracy value is 85.00% for KNN algorithm. As for the CHAID algorithm AUC value that has the highest results of 0.758. and from the t-test ac- count that ID3, CHAID and Naive Bayes algorithms are the best performing algorithms applied to the blogger dataset. As for the algorithm KNN and C45 is an algorithm with a poor performance with an AUC value of 0.500%.

Keyword: Data Mining, Algorithm, Classification, Blogger Dataset

1. PENDAHULUAN

Selama beberapa tahun terakhir perkembangan teknologi informasi menjadi sangat maju dalam hal pengumpulan dan penyimpanan data, maka munculah suatu kebutuhan untuk dapat menghasilkan informasi dari data yang telah ada tersebut. setiap informasi yang ada menjadi suatu hal yang penting untuk menentukan setiap

keputusan dalam situasi tertentu. hal ini menyebabkan penyediaan informasi menjadi sarana untuk dianalisa dan diringkas menjadi suatu pengetahuan dari data yang bermanfaat ketika pengambilan suatu keputusan dilakukan. Data mining merupakan sebuah proses ekstraksi untuk mendapatkan suatu informasi yang sebelumnya tidak diketahui dari sebuah data. data mining dapat menganalisa kasus lama untuk menemukan pola dari data dengan menggunakan teknik pengenalan pola seperti statistik dan matematika (Witten,2011).

Salah satu cara untuk mendapatkan informasi atau pola dari kumpulan data yang besar adalah dengan menggunakan teknik-teknik dalam data mining. algoritma yang digunakan dalam penelitian ini adalah algoritma klasifikasi. Dalam klasifikasi membutuhkan sebuah data training untuk menemukan sebuah pola. kemudian dari data training tersebut akan diketahui performa disetiap algoritma klasifikasi. sehingga dapat menentukan performa yang terbaik diantara algoritma yang digunakan.

Blog merupakan sebuah media sosial yang baru-baru ini berada di ruang cyber adalah salah satu layanan internet dan web (Zafarani,2008) (Wyld,2007) yang menyediakan komponen perangkat lunak gratis bagi pengguna untuk membiarkan mereka berpartisipasi sebagai anggota jaringan dan komunitas virtual (Soleimanian,2012). sehingga menyebabkan hubungan dinamis dan interaktif yang tidak terbatas, dan opini tentang masalah yang diberikan (Juffinger,2009). penyebab kecenderungan pada blogger dan parameter utama pendekatan mereka adalah di antara isu-isu utama perencanaan untuk negara-negara yang ditentukan berdasarkan teknologi modern. jadi, pentig untuk memberikan solusi yang tepat untuk menentukan faktor-faktor utama kecenderungan pada blogging (Soleimanian,2012).

Algoritma klasifikasi data mining adalah suatu metode pembelajaran untuk memprediksi nilai dari sekelompok attribut dalam menggambarkan dan membedakan kelas data atau konsep yang bertujuan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui. beberapa algoritma klasifikasi yang sering digunakan antara lain adalah naïve bayes, decision tree, neural network, k-nn, random forest dan lain sebagainya. performa algoritma data mining dalam banyak kasus tergantung pada kualitas dataset, karena data training berkualitas rendah dapat menyebabkan klasifikasi yang lemah.komparasi algoritma klasifikasi sudah banyak dilakukan oleh para peneliti dengan hasil yang berbeda-beda. dengan menggunakan data public dari uci repository yang memiliki 6 atribut dan 100 record. dimana data yang digunakan adalah data blogger, yaitu untuk mengklasifikasi blogger profesional.

Terdapat lima algoritma klasifikasi yaitu C45, ID3, Chaid, K-NN dan Naive Bayes. Dari berbagai algoritma yang digunakan, penelitian ini bertujuan untuk mengetahui performa mana yang lebih baik diantara lima algoritma tersebut dengan menggunakan uji t-test dan tools yang digunakan adalah rapid miner sehingga dapat mengetahui performa yang baik dari algoritma-algoritma tersebut.

Penelitian mengenai data mining dengan membandingkan algoritma klasifikasi sudah banyak dipublikasikan. Dalam penelitian ini, diperlukannya referensi-referensi dari penelitian-penelitian yang sebelumnya sehingga dapat mengetahui metode yang digunakan.

Penelitian yang pertama yang dijadikan sebagai referensi atau acuan dalam penelitian ini yaitu penelitian yang dilakukan oleh Soleimanian et al Menjelaskan bahwa Blog adalah media baru yang muncul yang bergantung pada teknologi informasi dan kemajuan teknologi. Karena media massa di beberapa negara kurang berkembang dan berkembang berada di layanan pemerintah dan kebijakan mereka dikembangkan berdasarkan kepentingan pemerintah, jadi blog disediakan untuk ide dan bertukar pendapat. simulasi dari informasi yang diperoleh dari 100 pengguna dan blogger di Kohkiloye dan Boyer Ahmad Province dan menggunakan alat bantu Weka 3.6 dan algoritme c4.5 dengan menerapkan pohon keputusan dengan lebih dari% 82 presisi untuk mengantisipasi kecenderungan pengguna di masa depan untuk ngeblog dan menggunakan di area strategis (Soleimanian,2012).

Penelitian selanjutnya yaitu penelitian yang dilakukan oleh Khafizh Hastuti (Khafizh,2012) penelitian ini menerapkan algoritma klasifikasi untuk evaluasi serta untuk mengetahui algoritma klasifikasi yang paling akurat dengan menggunakan dataset mahasiswa non aktif.

Penelitian selanjutnya yaitu penelitian yang dilakukan oleh M. Adib Alkaromi (Adip,2012) penelitian ini menerapkan algoritma klasifikasi dalam membandingkan performa dari masing-

masing algoritma yang digunakan dengan menggunakan rapid miner.

Penelitian terakhir yang digunakan sebagai referensi panduan dalam penelitian terkait yaitu penelitian yang dilakukan oleh Wahono, Suryana, dan Ahmad (Wahono,2014), penelitian dilakukan pada Software Defect Prediction dengan menggunakan beberapa jenis algoritma klasifikasi dalam memprediksi kerusakan perangkat lunak.

Tujuan dari penelitian ini adalah menentukan dari ke lima algoritma tersebut, algoritma mana yang menghasilkan nilai akurasi dan AUC yang lebih baik.

2. TINJAUAN PUSTAKA

2.1. Data Mining

Data mining adalah suatu disiplin ilmu yang bertujuan untuk menemukan, menggali atau menambahkan pengetahuan dari data atau informasi yang kita miliki. Menurut Gartner Group menyebutkan bahwa data mining adalah proses menelusuri pengetahuan baru, pola dan tren yang dipilih dari jumlah data yang besar yang disimpan dalam repositori atau tempat penyimpanan dengan menggunakan teknik pengenalan pola serta statistik dan tehnik matematika (Widiastuti,2012). Data Mining atau sering juga disebut Knowledge Discovery in Database (KDD) adalah sebuah bidang ilmu yang banyak membahas tentang pola sebuah data.Serangkaian proses guna mendapatkan pengetahuan atau pola dari kumpulan data disebut dengan data mining (Witten,2011). Sebuah data yang besar bisa saja tidak berguna dan hanya akan menjadi sampah bila kita tidak dapat memanfaatkannya. Data mining menjawab masalah ini dengan menganalisa data yang besar tersebut kemudian membuat sebuah aturan, pola, ataupun model tertentu untuk mengenali data baru yang tidak berada dalam baris data yang tersimpan (Prasetyo,2012).

2.2. Pohon Keputusan

Pohon keputusan adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan record yang lebih kecil dengan menrapkan serangkaian aturan keputusan.

Algoritma yang dapat dipakai dalam pembentukan pohon keputusan.

1.ID3
2.CART
3.C4.5

C4.5 Merupakan pengembangan dari algoritma ID3 (Larose,2005) yang dikembangkan oleh Quinlan (Han dan Kamber, 2006). Algoritma C4.5 banyak digunakan peneliti untuk melakukan tugas klasifikasi. Output dari algoritma C4.5 adalah sebuah pohon keputusan atau sering dikenal dengan decissin tree. Dalam beberapa penelitian algoritma C4.5 ini menjadi pilihan terbaik dibandingkan dengan beberapa algoritma klasifikasi lain (Wu,2007). Tahapan Algoritma C4.5 adalah, sebagai berikut:

1. Pilih atribut sebagai akar.

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

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

2. Buat cabang untuk tiap-tiap nilai.
3. Bagi kasus dalam cabang.
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

2.3. Naive Bayes

Naive Bayes merupakan metode yang tidak memiliki aturan, Naive Bayes menggunakan cabang matematika yang dikenal dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi tiap klasifikasi pada data training. Klasifikasi Naive Bayes adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. Klasifikasi bayesian memiliki kemampuan klasifikasi serupa dengan decision tree dan neural network menurut Awwalu. Bayes rule digunakan untuk menghitung probabilitas suatu class. Algoritma Naive Bayes memberikan suatu cara mengkombinasikan peluang terdahulu dengan syarat kemungkinan menjadi sebuah formula yang dapat digunakan untuk menghitung peluang dari tiap kemungkinan yang terjadi. Bentuk umum dari teorema bayes seperti dibawah ini (Rizal,2014).

$$P(H|X) \quad \frac{P(X|H)P(H)}{P(X)}$$

Dimana:
X : Data dengan class yang belum diketahui
H : Hipotesis data X merupakan suatu class spesifik.
P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posteriori probability)

P(H) : Probabilitas hipotesis H (prior probability)
P(X|H) : Probabilitas X berdasar kondisi pada hipotesis H
P(X) : Probabilitas dari X
Naïve bayes adalah penyederhanaan metode bayes. Teorema bayes disederhanakan menjadi:

$$P(H|X)=P(X|H)P(X)$$

Bayes rule diterapkan untuk menghitung posterior dan probabilitas dari data sebelumnya. Dalam analisis bayesian, klasifikasi akhir dihasilkan dengan menggabungkan kedua sumber informasi (prior dan posterior) untuk menghasilkan probabilitas menggunakan aturan bayes (Rizal, 2014).

2.4. K-Nearest Neighbour (K-NN)

Algoritma K-Nearest Neighbor adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Nearest Neighbor adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dan kasus lama yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada menurut Obbie.

Langkah-langkah untuk menghitung metode Algoritma K-Nearest Neighbor:

a. Menentukan Parameter K (Jumlah tetangga paling dekat).
b. Menghitung kuadrat jarak Euclid (queri instance) masing-masing objek terhadap data sampel yang diberikan.
c. Kemudian mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak Euclid terkecil.
d. Mengumpulkan kategori Y (KlasifikasiNearest Neighbor).
e. Dengan menggunakan kategori Nearest Neighbor yang paling mayoritas maka dapat diprediksi nilai queri instance yang telah dihitung.

2.5. ID3

Algoritma ID3 atau iterative Dichotomiser 3 (ID3) merupakn sebuah metode yang digunakan untuk membangkitkan pohon keputusan. Algoritma pada metode ini menggunakan konsep dari entropy informasi menurut Obbie. Pemilihan atribut dengan menggunakan Informastion Gain. Pemilihan atribut pada ID3 dilakukan dengan properti statistik yang disebut dengan information gain. Gain mengukur seberapa baik suatu atribut memisahkan training example ke dalam kelas target. Atribut dengan informasi tertinggi akan dipilih, dengan tujuan untuk mendefenisikan gain. Entropy bisa dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas. Semakin kecil nilai entropy maka semakin baik digunakan dalam mengekstraksi suatu kelas.

2.6. CHAID

CHAID adalah singkatan dari Chi-squared Automatic Interaction Detector. CHAID bekerja untuk menduga sebuah variabel tunggal, disebut sebagai variabel dependen yang didasarkan pada sejumlah variabel-variabel independen. CHAID merupakan suatu teknik iteratif yang menguji satu persatu variabel independen yang digunakan dalam klasifikasi dan menyusunkan berdasarkan pada tingkat signifikansi statistik chi-square terhadap variabel dependennya (Gallacgher,2000).

3. METODOLOGI

Penelitian yang dilakukan dengan mengusulkan lima algoritma dan sebagai alat bantu dalam menghitung performa dari setiap algoritmanya adalah repid miner. Lima algoritma yang digunakan yaitu C45, Naive Bayes, K-NN, ID3 dan CHAID. Perbandingan yang dilakukan untuk mengetahui salah satu algoritma yang paling baik performanya diantara kelima algoritma tersebut dengan menggunakan dataset Blogger. Penelitian ini menggunakan sebuah software yaitu rapid miner. Pada penelitian ini dilakukan beberapa langkah atau tahapan penelitian antara lain: pengumpulan data,pengolahan data awal,metode yang diusulkan,pengujian metode dan hasil penelitian.

3.1 Dataset

Dataset blogger merupakan salah satu data publik yang ada pada web UCI Machine Learning repository.Database bloggers ini dari Kohkiloye dan Boyer Ahmad Province dari Iran. Atribute yang terdapat pada data Blogger yaitu: Local Political Social Space (LPSS), Local Media Turnover (LMT), Topic, Caprice, Degree. Sedangkan untuk class yang terdapat pada dataset blogger yaitu Professional Blogger (PB) adalah yes dan no.

Tabel 1
Blogger Database of Kohkiloye and Boyer Ahmad Province in Iran

No	Degree	Caprice	Topic	LMT	LPSS	PB
1	high	left	Impression	yes	yes	yes
2	high	left	political	yes	yes	yes
3	medium	middle	Tourism	yes	yes	yes
4	high	left	political	yes	yes	yes
5	medium	middle	News	yes	yes	yes
6	medium	middle	News	yes	yes	yes
7	high	left	political	yes	yes	yes
8	high	right	political	yes	no	yes
9	high	right	political	yes	no	no
10	medium	right	Tourism	yes	no	yes
11	high	right	Tourism	yes	yes	yes
12	medium	left	News	yes	no	yes
13	high	left	political	yes	yes	no
14	low	right	news	no	yes	no
15	high	left	political	yes	yes	yes
16	medium	left	impression	yes	yes	yes
17	medium	left	political	yes	yes	yes
18	high	right	political	yes	yes	yes
19	medium	left	impression	yes	yes	yes
20	high	right	tourism	yes	yes	no
21	high	left	political	yes	yes	yes
22	medium	left	news	yes	yes	yes
23	high	right	political	no	yes	no
24	low	left	tourism	yes	no	no
25	high	left	news	yes	yes	yes
26	high	left	political	yes	yes	yes
27	low	right	impression	no	no	yes
28	high	right	political	yes	yes	yes
29	high	left	impression	no	no	yes
30	medium	left	scientific	yes	yes	no
31	high	right	political	yes	yes	yes
32	low	left	scientific	yes	yes	no
33	medium	right	tourism	yes	yes	no
34	Low	right	political	yes	yes	yes
35	High	left	impression	yes	no	yes
36	medium	left	tourism	yes	no	yes
37	medium	middle	scientific	yes	no	yes
38	medium	middle	impression	no	yes	no
39	medium	right	scientific	yes	yes	no
40	medium	left	impression	no	no	yes
41	High	left	political	yes	yes	no
42	medium	left	news	no	yes	yes
43	High	left	political	yes	yes	yes
44	medium	right	news	yes	yes	no
45	medium	left	tourism	yes	no	yes
46	medium	middle	news	yes	yes	yes
47	Low	middle	impression	yes	no	no
48	Low	right	impression	yes	no	no
49	medium	right	news	yes	yes	no
50	medium	left	impression	yes	yes	yes
51	High	left	political	yes	yes	yes
52	High	left	political	yes	yes	yes
53	medium	middle	tourism	yes	yes	yes
54	High	left	political	yes	yes	yes
55	medium	middle	news	yes	yes	yes

56	medium	middle	news	yes	yes	yes
57	High	left	political	yes	yes	yes
58	High	right	political	yes	no	yes
59	High	right	political	yes	no	no
60	medium	right	tourism	yes	no	yes
61	medium	right	tourism	yes	yes	yes
62	medium	left	news	yes	no	yes
63	High	left	impression	yes	yes	no
64	Low	right	news	no	yes	no
65	High	left	political	yes	yes	yes
66	medium	left	impression	yes	yes	yes
67	medium	left	political	yes	yes	yes
68	High	right	political	yes	yes	yes
69	medium	left	political	yes	yes	yes
70	High	right	impression	yes	yes	no
71	medium	left	political	yes	yes	yes
72	medium	left	news	yes	yes	yes
73	medium	right	political	no	yes	no
74	Low	left	tourism	yes	no	no
75	High	left	news	yes	yes	yes
76	High	left	political	yes	yes	yes
77	Low	right	impression	no	no	yes
78	High	right	political	yes	yes	yes
79	High	left	impression	no	no	yes
80	medium	left	scientific	yes	yes	no
81	High	right	political	yes	yes	yes
82	Low	left	scientific	yes	yes	no
83	medium	right	tourism	yes	yes	no
84	Low	right	political	yes	yes	yes
85	high	left	impression	yes	no	yes
86	medium	left	tourism	yes	no	yes
87	medium	middle	impression	yes	no	yes
88	medium	middle	impression	no	yes	no
89	medium	right	scientific	yes	yes	no
90	medium	left	impression	no	no	yes
91	high	left	political	yes	yes	no
92	medium	left	news	no	yes	yes
93	high	left	political	yes	yes	yes
94	medium	right	news	yes	yes	no
95	medium	left	tourism	yes	no	yes
96	medium	middle	impression	yes	yes	yes
97	low	middle	impression	yes	no	no
98	low	right	impression	yes	no	no
99	medium	right	news	yes	yes	no
100	medium	left	impression	yes	yes	yes

3.2 *Cross Validation*

Kemudian untuk validasi, penelitian ini menggunakan 10-*fold cross validation*. 10-fold cross-validation akan mengulang pengujian sebanyak 10 kali. Berikut tabel 10-fold Cross Validation:

TABEL 2
10-FOLD CROSS VALIDATION

Validation	Dataset								
1	■								
2	■								
3		■							
4			■						
5				■					
6					■				
7						■			
8							■		
9								■	
10									■

3.3 *Confusion Matrix*

Evaluasi model klasifikasi didasarkan pada pengujian untuk memperkirakan obyek yang benar dan salah [14], urutan pengujian ditabulasikan dalam confusion matrix dimana kelas yang diprediksi ditampilkan dibagian atas matriks dan kelas yang diamati disisi kiri. Setiap sel berisi angka yang menunjukkan berapa banyak kasus yang sebenarnya dari kelas yang diamati untuk diprediksi.

TABEL 3
CONFUSION MATRIX

CLASSIFICATION		
	Class = True	**Class = False**
Class = True	a (true-positive-TP)	b (false-positive-TP)
Class = False	c (true-positive-TP)	d (false-positive-TP)

2.4. *ROC Curve*

Kurva ROC dibagi dalam dua dimensi, dimana tingkat TP diplot pada sumbu Y dan tingkat FP diplot pada sumbu X. Tetapi untuk merepresentasikan grafis yang menentukan klasifikasi mana yang lebih baik, digunakan metode yang menghitung luas daerah dibawah kurva ROC yang disebut AUC (Area Under the ROC Curve) yang diartikan sebagai probabilitas[14].

AUC mengukur kinerja diskriminatif dengan memperkirakan probabilitas output dari sampel yang dipilih secara acak dari populasi positif atau negatif, semakin besar AUC, semakin kuat klasifikasi yang digunakan. Karena AUC adalah bagian dari daerah unit persegi, nilainya akan selalu antara 0,0 dan 1,0.

TABEL 4
NILAI AUC

Nilai AUC	Klasifikasi
0.90 - 1.00	Paling Baik
0.80 - 0.90	Baik
0.70 - 0.80	Adil atau Sama
0.60 - 0.70	Rendah
0.50 - 0.60	Gagal

2.6. *T-Test*

T-Test adalah metode pengujian hipotesis dengan menggunakan satu individu (objek penelitian) dengan menggunakan dua perlakuan yang berbeda. Walaupun dengan menggunakan objek yang sama tetapi sampel tetap terbagi menjadi dua yaitu data dengan perlakuan pertama dan data dengan perlakuan kedua. Performance dapat diketahui dengan cara membandingkan kondisi objek penelitian pertama dan kondisi objek pada penelitian kedua.

4. HASIL DAN PEMBAHASAN

Perbandingan performance masing-masing dari algoritma, sebagai berikut:

Berdasarkan tabel diatas, dapat diketahui bahwa algoritma KNN memiliki nilai accuracy tertinggi yaitu 85.00%, ID3 82.00%, CHAID 75.00%, Naive Bayes 71,00% dan Decision Tree 68.00%. Sedangkan pada uji ROC curve menunjukkan bahwa CHAID dan ID3 mencapai nilai AUC yang terbaik yaitu 0.758 dan 0.757, kemudian Naive Bayes 0.730 dan KNN serta Decision Tree 0.500.

TABEL 5
PERBANDINGAN PERFORMANCE LIMA ALGORITMA

Model	Accuracy	AUC
Decision Tree	68.00%	0.500
Naive Bayes	71.00%	0.730
KNN	85.00%	0.500
ID3	82.00%	0.757
CHAID	75.00%	0.758

Kemudian, pengujian t-test akan didapatkan perbandingan, sebagai berikut:

TABEL 6
UJI STATISTIK T-TEST

A	B	C	D	E	F
	0.680 +/- 0.160	0.740 +/- 0.092	0.800 +/- 0.173	0.690 +/- 0.070	0.700 +/- 0.063
0.680 +/- 0.160		0.317	0.125	0.858	0.717
0.740 +/- 0.092			0.346	0.187	0.271
0.800 +/- 0.173				0.079	0.104
0.690 +/- 0.070					0.741
0.700 +/- 0.063					

Keterangan :
B : Decision Tree
C : Naive Bayes
D : KNN
E : ID3
F : Chaid

Dari pengujian t-test diatas, bahwa hasil perbandingan antara algoritma KNN dengan ID3 ada perbedaan yang signifikan (H1). Kemudian untuk perbandingan algoritma antara C45 dengan Naive Bayes, KNN, ID3 dan CHAID tidak ada perbedaan yang signifikan (H0). Begitu juga dengan Naive Bayes dengan KNN, ID3 dan CHAID tidak ada perbedaan yang signifikan(H0). Sama seperti perbandingan antara KNN dengan CHAID dan ID3 dengan CHAID tidak ada perbedaan yang signifikan.

Dilihat dari hasil pengujian AUC dan T-Test, algoritma yang memiliki performance terbaik adalah Algoritma ID3, CHAID dan NB. Sedangkan algoritma KNN dan C45 merupakan algoritma yang memiliki performance yang kurang baik dalam penerapan dataset blogger.

5. KESIMPULAN DAN SARAN

Penelitian dengan menggunakan dataset blogger yang di dapat dari uci machine learning repository dengan membandingkan 5 algoritma klasifikasi yaitu decision tree, naïve bayes, K-Nearest Neighbour, ID3, dan chaid. dengan menggunakan validasi 10-fold cross validation dan uji t-test. sehingga hasil tertinggi dari nilai akurasi sebesar sebesar 85.00% yaitu algoritma KNN. sedangkan untuk nilai AUC sebesar 0.758 untuk algoritma Chaid. dan dari asil uji t-test yang dilakukan bahwa algoritma id3, chaid dan Naive Bayes merupakan algoritma dengan performa terbaik yang diterapkan pada dataset blogger. sedangkan untuk algoritma knn dan C45 merupakan algoritma dengan performa yang kurang baik dengan nilai AUC sebesar 0,500%.

Adapun saran untuk penelitian selanjutnya adalah sebagai berikut:

1. Dapat menggunakan Dataset yang berbeda yang dapat di ambil dari *UCI Machine Learning Repository*
2. Dapat menggunakan data preprocessing seperti menambahkan fitur selection.
3. Menggunakan model Agortima yang berbeda dengan dataset yang sama.

REEFERENSI

Adip Alkaromi M. Komparasi Algoritma Klasifikasi untuk Dataset Iris dengan Repid Miner. 2012.

D. Widiastuti, "Analisa Perbandingan Algoritma SVM, Naive Bayes, dan Decision Tree dalam Mengklasifikasikan Serangan (Attacks) pada Sistem Pendeteksi Intrusi," 2012.

D. T. Larose, Discovering Knowledge in Data: an Introduction to Data Mining. John Wiley & Sons, 2005.

E. Prasetyo, Data Mining Konsep dan Aplikasi menggunakan Matlab.Yogyakarta: Andi Offset, 2012, p. 353.

Florin Gorunescu, Data Mining: Concepts, Model and Techniques, Prof. Janusz Kacprzyk and Prof. Lakhmi C. Jain, Eds. Berlin,

Gallacgher, CA. 2000. *An Iterative Approach to Classification Analysis.*

I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition. Elsevier, 2011.

J. Han and M. Kamber, Data Mining: Concepts and Techniques Second Edition. Elsevier, 2006.

J. Awwalu, A. Ghazvini, and A. A. Bakar, "Performance Comparison of Data Mining Algorithms: A Case Study on Car Evaluation Dataset."

Juffinger,A., Lex, E., 2009, Cross language Blog Mining and Trend Visualization ,WWW 2009, 2009, Madrid, Spain.1149-1150.

Khafiizh Hastuti. Analisa Komparasi Algoritma Klasifikasi Data Mining untuk Prediksi Mahasiswa Non Aktif. ISBN 979-26-0255-0, 2012.

Obbie, "Penerapan Algoritma Klasifikasi Data Mining ID3 untuk Menentukan Penjurusan Siswa SMAN 6 Semarang

Rizal Amegia Saputra, "komparasi algoritma klasifikasi data mining untuk memprediksi penyakit tuberculosis (tb)," semin. nas. inov. dan tren snit, 2014.

Soleimanian Gharehchopogh, F., & Reza Khaze, S. (2012). Data Mining Application for Cyber Space Users Tendency in Blog Writing: A Case Study. *International Journal of Computer Applications*, *47*(18), 975–888. https://doi.org/10.5120/7291-0509

Wyld,D., 2007, The Blogging Revolution: Government in the Age of Web 2.0 ,IBM Center for The Business of Government.

Wahono, R. S., Herman, N. S., & Ahmad, S. (2014). A comparison framework of classification models for software defect prediction. *Advanced Science Letters*, *20*(10–12), 1945–1950. http://doi.org/10.1166/asl.2014.5640

X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang,

H. Motoda, G. J. Mclachlan, A. Ng, B. Liu, P. S. Yu, Z. Z. Michael, S. David, and J. H. Dan, Top 10 algorithms in data mining. 2007, pp. 1–37.

Zafarani,R, Jashki, M.A, Baghi,H.R , Ghorbani,A., 2008, A Novel Approach for Social Behavior Analysis of the Blogosphere, springer-Verlag Berlin Heidelberg, S. Bergler (Ed.): Canadian AI, 356–367.