Research : Journal of Computer, information system, & technology management **Online** ISSN: 2615-7357
Vol. 1, **No. 2**. **April 2018, Pages 7-11** **Print** ISSN: 2615-7233

UJI PENGARUH KARAKTERISTIK DATASET PADA PERFORMA ALGORITMA KLASIFIKASI

Moch. Ali Machmudi[1)]
[1)] Stmik Bina Patria
[1)]Jurusan Manjemen Informatika-D3
Email : [1]aliadhinata@gmail.com[1)]

Abstrak

Tujuan utama penelitian ini adalah untuk mengetahui pengaruh karakteristik set data pada performa algoritma klasifikasi. Pada penelitian ini digunakan tiga set data yang memiliki variasi tipe data, jumlah atribut, dan jumlah instan yang berbeda. Set data dibelajarkan pada algoritma klasifikasi, seperti SMO, Adaboost, CART, C4.5, dan Naïve Bayes. Penelitian ini menggunakan 10 fold cross validation sebagai metode evaluasi. Hasil penelitian menunjukkan bahwa, tipe data, jumlah atribut, dan ukuran set data mempengaruhi performa algoritma klasifikasi. Semakin banyak jumlah atribut, kecenderungan akurasi kelima algoritma uji semakin tinggi. Algoritma klasifikasi yang terbaik digunakan pada tipe data numerik adalah C4.5, sedangkan untuk data nominal adalah SMO. Algoritma klasifikasi yang terbaik digunakan pada small dataset atau set data dengan jumlah instan kecil adalah Naïve Bayes dan SMO, sedangkan yang terbaik digunakan pada big dataset adalah SMO dan C4.5.

Kata Kunci – data set ; small dataset ; klasifikasi ; peforma; data mining

1. PENDAHULUAN

Dewasa ini, *data mining* telah menjadi tren dalam dunia bisnis dan riset teknologi informasi. Perkembangan *data mining* yang pesat tidak lepas dari perkembangan teknologi informasi yang memungkinkan data dalam jumlah besar terakumulasi. Lima komponen utama pada proses *data mining* adalah *input*, metode, *output*, dan evaluasi [1]. *Input* proses *data mining* adalah set data yang merupakan koleksi data dalam jumlah yang diperoleh dari aktivitas transaksi bisnis, pencatatan data-data kesehatan, transaksi sistem, dll.

Set data yang biasa digunakan pada penelitian di bidang *data mining* terdiri dari dua tipe, prifat dan publik. Set data prifat merupakan set data yang diambil dari organisasi yang akan dijadikan objek penelitian, seperti data bank, rumas sakit, pabrik, dan perusahan jasa. Set data publik adalah set data yang dapat diambil dari repositori publik yang disepakati oleh peneliti *data mining*. Misalnya Uci Repository [2] dan ACM KDD [3].

Ekstraksi pengetahuan dari koleksi data merupakan tugas utama pada proses data mining. Tiga proses utama ektraksi pengetahuan ini adalah explorasi, pembangunan model, dan pengembangan. [4]. Pada tahapan eksplorasi, dilakukan proses preprosesing data, transformasi data, seleksi atribut, dll. Pada pembangunan model dan evaluasi, dilakukan proses pemilihan metode dan model yang terbaik agar performa prediksi yang dihasilkan tinggi. Tahapan terakhir adalah menerapkan model pada data yang akan diprediksi yang akan menjadi keluaran sistem *data mining*. Keluaran *output data mining* sangat tergantung pada set data yang dibelajarkan dan algoritma yang digunakan. Terkadang, data tidak terklasifikasi dengan baik karena pemilihan algoritma tidak sesuai dengan set data pembelajaran. Oleh karena itu, langkah awal sebelum melakukan proses *data mining* adalah mepelajari dan memahami data yang nantinya akan digunakan pada pemilihan metode atau algoritma terbaik.

Beberapa algoritma klasifikasi pada *data mining* adalah Decision Tree, Support Vector Machine, Bayesian Network, Neural Network, Logistic Regression, dll

Pada penelitian ini digunakan tiga set data yang memiliki karakteristik berbeda seperti jenis tipe data, jumlah atribut, dan ukuran set data. Set data dibelajarkan dengan beberapa algoritma klasifikasi untuk menjawab beberapa pertanyaan berikut:

1. Apakah ukuran data set mempengaruhi performa algoritma klasifikasi lanjut?
2. Adakah pengaruh tipe data dan jumlah atribut pada performa algoritma kalsifikasi lanjut?
3. Algoritma apa yang memiliki performa lebih baik pada *small dataset*?
4. Algoritma apa yang memiliki performa lebih baik pada dataset bertipe data numerik dan pada data nominal?

2. PENELITIAN TERKAIT

Beberapa penelitian yang sama mengenai pengujian set data telah dilakukan pada [6], yang menguji perbandingan algoritma *clustering* dengan menguji beberapa varian jumlah instan dan klaster yang dihasilkan. Pada penelitian [7], dilakukan uji performa algoritma klasifikasi dan *clustering* di weka dengan membandingan mode *testing* 10 fold cross validation dan Percentage Split. Pada Penelitian [4] dilakukan uji performa algoritma klasifikasi decision tree pada set data dengan dua tipe data yang berbeda, yaitu nominal dan numeric. Pada Penelitian [8], dilakukan penelitian membandingan algoritma SVM dan KNN pada *medical data set* dengan berbagai ukuran set data atau jumlah instan.

3. LANDASAN TEORI

A. Set Data

Beberapa faktor yang menjadi pertimbangan karakteristik set data adalah atribut, class, tipe data, dan jumlah instan. Atribut adalah faktor atau parameter yang menyebabkan class/label/target terjadi. *Class* adalah atribut yang akan dijadikan target, sering juga disebut dengan label. Tipe data untuk variabel pada statistik terbagi menjadi empat: nominal, ordinal, interval, ratio tetapi secara praktis, tipe data untuk atribut pada *data mining* hanya menggunakan dua: Nominal (Diskrit) dan Numeric (Kontinyu atau Ordinal).

B. Metode Klasifikasi

1) SMO

SVM merupakan metode klasifikasi yang berusaha menemukan *hyperplane* terbaik pada *input* space. Prinsip dasar SVM adalah linear classifier, dan selanjutnya dikembangkan agar dapat bekerja pada problem non-linear. dengan memasukkan konsep *kernel trick* pada ruang kerja berdimensi tinggi. SMO merupakan salah satu algoritma SVM [9].

2) CART

CART adalah metode partisi rekursif yang menggunakan dua metode regresi dan klasifikasi [4]. CART, penerapan metode algoritma ini banyak digunakan dalam berbagai bidang yang membutuhkan pengolahan data yang komprehensif. Hanya saja mekanismenya terdiri dari beberapa tahap yang bertingkat meliputi *automatic class balancing, automatic missing, value handling cost-sensitive learning, dynamic feature construction* dan probabilitas estimasi *tree* sehingga tingkat kompleksitas menjadi pertimbangan para peneliti pemula. Hasil akhirnya adalah gambaran atribut berdasarkan prioritas kebutuhan proses [10].

3) C4.5

C4.5 adalah sebuah *decision tree* yang digunakan untuk klasifikasi dengan konsep *information entropy* [4]. Sebuah pohon keputusan adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan record yang lebih kecil dengan menerapkan serangkaian aturan keputusan .

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut [5] :

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

4) Naive bayes

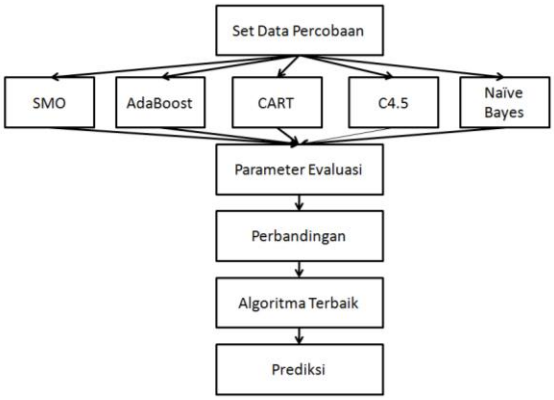
Naive Bayesian *classifier* adalah metode klasifikasi yang berdasarkan probabilitas dan Teorema Bayesian dengan asumsi bahwa setiap variabel X bersifat bebas (independent). Dengan kata lain, Naïve Bayesian *classifier* mengansumsikan bahwa keberadaan sebuah atribut (variabel) tidak ada kaitannya dengan beradaan atribut yang lain [9]. Naïve Bayes memiliki keunggulan untuk pengembangan *data mining* yaitu kemudahan konstruksinya dan tidak membutuhkan parameter skema pengulangan yang kompleks sehingga mudah dalam membaca data dalam jumlah yang besar [10]. Hal ini terjadi karena desain rancangan penuntunan klasifikasi terhadap data. Selain itu, metode ini dinyatakan sebagai algoritma yang mempunyai sifat *simplicity, elegance* dan *robustnes.*

C. Evaluasi performa

Dalam penelitian ini peforma masing-masing algoritma klasifikasi terhadap dua seleksi fitur akan diukur berdasarkan *accuracy*, *time comsumtion,* dan *root mean square error.*

4. METODE

Langkah Penelitian digambarkan pada bagan dibawah ini.



Gambar 1: Alur Penelitian

A. Set Data

Bahan penelitian yang digunakan pada penelitian ini adalah tiga set data berbeda yang diambil dari UCI *machine learning repository* yang bersumber dari University of Wisconsin Hospitals, Madison dari Dr. William H. Wolberg http://archive.ics.uci.edu/ml/datasets.html).

Tabel 1. Set Data

Nama Data set	Jumlah Atribut	Jumlah Instan	Karakteristik Atribut	Karakteristik Data set	Missing Value
1. Glass	10	214	Numeric	Multivariate	No
2. Ionosphere	34	351	Numeric	Multivariate	No
3. Soybean	35	683	Nominal	Multivariate	Yes

Set data 1 dan 2 memilik tipe data yang sama, tetapi jumlah atribut yang berbeda. Set data 2 dan 3 memiliki jumlah atribut sama, dengan tipe yang berbeda.

B. Model Analisis Data

Penelitian ini menggunakan algoritma klasifikasi SMO, Adaboost, CART, C4.5, dan Naïve Bayes pada tool weka.

C. Mode testing

Metode pengujian yang digunakan adalah 10 *fold cross validation* .

D. Parameter Evaluasi

Parameter Evaluasi yang akan digunakan adalah akurasi, *built time, root mean square error*.

E. Percobaan

1) Percobaan 1 membandingkan akurasi pada set data dengan jumlah atribut yang berbeda.
2) Percobaan 2 membandingkan akurasi pada set data dengan tipe data berbeda (numeric-nominal)
3) percobaan 3 membandingkan akurasi pada set data dengan ukuran pada set data dengan tipe data nominal
4) percobaan 4 membandingkan akurasi pada set data dengan ukuran pada setdata dengan tipe data numerik

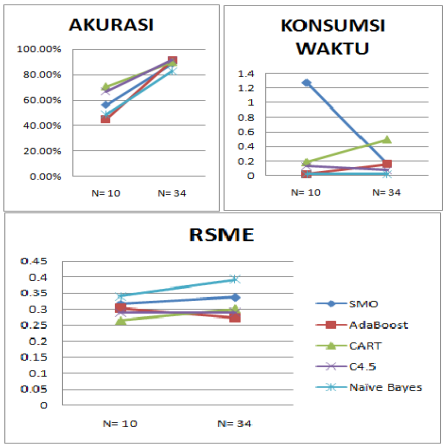
5. HASIL DAN PEMBAHASAN

A. Percobaan pada set data dengan jumlah atribut berbeda

Pada percobaan ini digunakan set data glass dan ionospher yang memiliki tipe data numerik dengan jumlah atribut yang berbeda. Berikut hasil percobaan.

Clasifier	N= 10			N=34		
	Akurasi	Time	RSME	Akurasi	Time	RSME
SMO	56.07%	1.27	0.3166	88.60%	0.16	0.3376
AdaBoost	44.86%	0.02	0.3027	90.88%	0.16	0.2733
CART	70.56%	0.19	0.2642	89.74%	0.5	0.3011
C4.5	66.82%	0.13	0.2897	91.45%	0.08	0.2901
Naïve Bayes	48.60%	0.02	0.3399	82.62%	0.02	0.3935

Tabel 2. Percobaan 1



Gambar 2. Grafik evaluasi percobaan 1

Dari grafik diatas dapat dilihat tren yang menunjukkan bahwa semakin besar jumlah

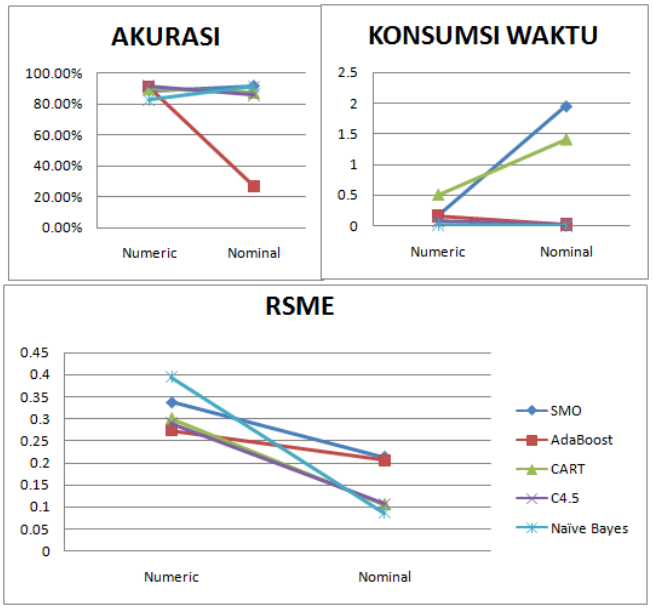
atribut, hasil akurasi pun juga meningkat. Pada dataset yang memiliki jumlah atribut banyak lebih baik menggunakan algoritma C.45, sedangkan untuk jumlah atribut sedikit menggunakan algortima CART.

B. Percobaan pada set data dengan tipe data berbeda

Pada percobaan ini digunakan set data ionospher dan soybean yang memiliki dengan jumlah atribut yang sama dan tipe data berbeda. Berikut hasil percobaan.

Tabel 3. Percobaan 2

Clasifier	Numeric			Nominasl		
	Akurasi	Time	RSME	Akurasi	Time	RSME
SMO	88.60%	0.16	0.3376	91.48%	1.95	0.2131
AdaBoost	90.88%	0.16	0.2733	26.99%	0.02	0.2057
CART	89.74%	0.5	0.3011	87.50%	1.41	0.1064
C4.5	91.45%	0.08	0.2901	85.80%	0.02	0.1065
Naïve Bayes	82.62%	0.02	0.3935	91.19%	0	0.0871



Gambar 3. Grafik evaluasi percobaan 2

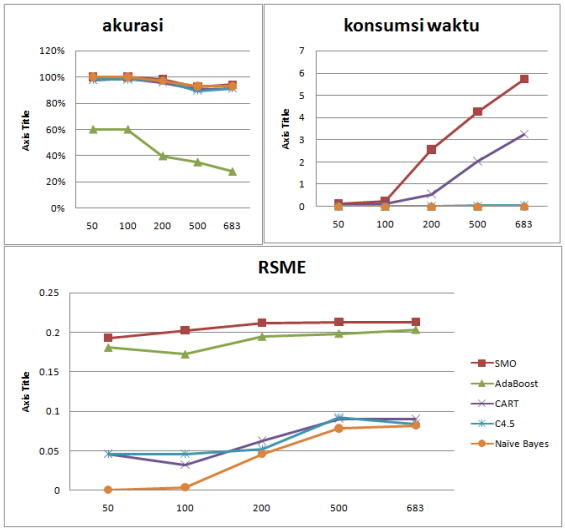
Algoritma klasifikasi yang terbaik digunakan pada data nominal adalah SMO. Algoritma klasifikasi yang terbaik digunakan pada data numerik adalah C4.5

C. Percobaan pada set data dengan tipe nominal dan ukuran set data yang berbeda.

Pada percobaan ini digunakan set data soybean dengan ukuran instan dibedakan menjadi beberapa tingkatan. Berikut hasil percobaan dengan parameter akurasi.

Tabel 4. Percobaan 3

SIZE	SMO	AdaBoost	CART	C4.5	Naïve Bayes
50	100%	60%	98%	98%	100%
100	100%	60%	99%	98%	100%
200	98.01%	39.80%	96.02%	97.51%	97.01%
500	92.22%	35.13%	90.62%	89.42%	93.01%
683	93.85%	27.96%	91.07%	91.51%	92.97%



Gambar 4. Grafik evaluasi percobaan 3

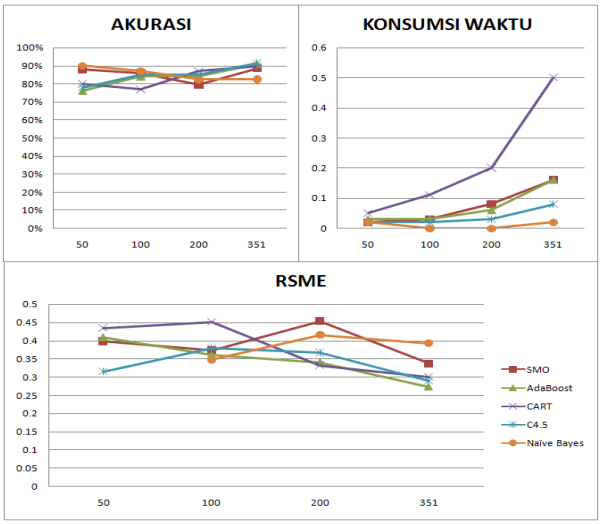
Algoritma klasifikasi yang terbaik digunakan pada data nominal untuk ukuran set data kecil (*small data set*) adalah SMO dan Naïve Bayes

D. Percobaan pada set data dengan tipe numerik dan ukuran set data yang berbeda.

Pada percobaan ini digunakan set data ionosfer dengan ukuran instan dibedakan menjadi beberapa tingkatan. Berikut hasil percobaan dengan parameter akurasi.

Tabel 5. Percobaan 4

Size	SMO	AdaBoost	CART	C4.5	Naïve Bayes
50	88%	76%	80%	78%	90%
100	86%	84%	77%	85%	87%
200	79.50%	84%	87%	85%	82.50%
351	88.60%	90.88%	89.74%	91.45%	82.62%



Gambar 5. Grafik evaluasi percobaan 4

Pada grafik diatas dapat diketahui bahwa algoritma klasifikasi yang terbaik digunakan pada data numerik untuk ukuran set data kecil adalah Naïve Bayes

6. KESIMPULAN

Dari penelitian ini dapat diketahui bahwa algoritma klasifikasi yang terbaik digunakan pada tipe data numerik adalah C4.5, seangkan untuk data nominal adalah SMO. Algoritma klasifikasi yang terbaik digunkan pada *small dataset* atau set data dengan jumlah instan kecil adalah Naïve Bayes dan SMO, sedangkan yang terbaik digunakan pada *big dataset* adalah SMO dan C4.5.

7. DAFTAR PUSTAKA

[1] Wahono, R.S. *Data mining* : Proses Data Mining. http://romisatriawahono.net. Diakses 9 Januari 2013

[2] *http://www.ics.uci.edu/~mlearn/MLRepository.html*

[3] *http://www.sigkdd.org/kddcup/*

[4] Saini, D., Rajavat, A. 2013. Performance Evaluation System For Decision Tree Algorithms. Journal: International Journal Of Computers & Technology.

[5] Fakhrurrifqi, M., Wardoyo R. 2013. Perbandingan Algoritma Nearest Network, C.45, dan LVQ untuk klasifikasi kemampuan mahasiswa. In IJCCS Universitas Gadjah Mada.

[6] Abbas, O.A. 2008. Comparisons Beetwen Data Clustering Algorithm. The International Arab Journal of informastion Technology.

[7] Tiwari, M., Jha, M.B., Yadav, O.P. 2012. Performance Analysis of *Data mining* Algorithm in weka.

[8] Raiwal, J.S., Saxena, K. 2012. Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over Medical Data set. International Journal of Computer Applications (0975 – 8887)

[9] Wahyuni, E.S., Setiawan, N.A., Nugroho, H.A. 2013. Penerapan Metode Seleksi Fitur Pada Klasifikasi Kanker Payudara.

[10] Subiyanto. A. 2008. Penggunaan Algoritma Klasifikasi Dalam Data Mining. Program Studi Sistem Informasi Fakultas Sains dan Teknologi UIN Jakarta