OPTIMASI ALGORITMA NAÏVE BAYES DENGAN INFORMATION GAIN RATIO UNTUK MENANGANI DATASET BERDIMENSI TINGGI

M. Adib Al Karomi [1)], Abdul Kharis [2)], Ivandari [3)]
[1] STMIK WIDYA PRATAMA PEKALONGAN
email: adib.comp@gmail.com
[2] STMIK WIDYA PRATAMA PEKALONGAN
email: abdulkharis.stmik@gmail.com
[3] STMIK WIDYA PRATAMA PEKALONGAN
email: ivandarialkaromi@gmail.com

Abstract

The development of computer science now allows the recording of all business processes in all fields with large storage media. Data in the fields of atronomy, health, economy, government and so on is widely recorded and is increasing from year to year. Data mining is a science that can process data into a representation of knowledge using several mathematical methods or algorithms. One of the main functions of data mining is classification. In the process of classification all old data is used as learning data to infer new data that is not yet fully known. Data which previously has no meaning can become a new knowledge by using data mining classification. Many algorithms can be used in the classification process. One of the algorithms that is proven to be good for the process of classifying high-dimensional data is Naïve Bayes. In high-dimensional data the many data attributes can affect the results of classification. The number of relevant data attributes can improve algorithm performance. While the number of irrelevant data attributes can reduce the level of accuracy of an algorithm. From the results of this study note that the selection feature of information gain ratio can improve the performance of Naive Bayes classification.

Keywords: *Bayes performance improvement, information gain ratio, public data*

1. PENDAHULUAN

Data mining merupakan bidang ilmu yang mempelajari data lampau untuk diambil menjadi sebuah pengetahuan baru (Witten, Frank, and Hall 2011). Salah satu fungsi utama data mining adalah klasifikasi. Seiring perkembangan teknologi komunikasi, objek penelitian data mining menjadi sangat luas. Data mining banyak digunakan dalam proses klasifikasi modern dengan menggunakan kumpulan data yang ada. Tipe data dapat mempengaruhi performa suatu algoritma klasifikasi data mining (Amancio et al. 2013). Algoritma klasifikasi terbaik untuk sebuah data belum tentu baik apabila digunakan untuk mengolah data yang lain (Patel, Vala, and Pandya 2014). Perbedaan performa algoritma ini dikarenakan adanya karakteristik yang berbeda dalam sebuah data (Ragab et al. 2014) (Ashari, Paryudi, and Tjoa 2013).

Salah satu algoritma klasifikasi terbaik dan banyak digunakan peneliti adalah naive bayes (Wu 2009). Naive bayes terbukti dapat menangani atribut data nominal dengan memanfaatkan perhitungan probabilitasnya. Dalam penelitian lain yang melakukan komparasi algoritma klasifikasi naive bayes terbukti menjadi algoritma terbaik dengan tingkat akurasi tertinggi (Kurniawan and Ivandari 2017). Dalam sebuah proses klasifikasi dilakukan perhitungan untuk setiap atribut data yang ada. Perbedaan jumlah atribut data yang digunakan dapat mempengaruhi tingkat akurasi sebuah

algoritma (Maimoon and Rokach 2010). Beberapa atribut yang relevan dan sesuai dapat meningkatkan performa suatu algoritma, sedangkan adanya atribut data yang tidak relevan dapat membuat performa algoritma menurun dan berkurangnya tingkat akurasi sebuah algoritma (Han and Kamber 2006). Tipe dari atribut dataset yang digunakan dalam proses klasifikasi juga dapat mempengaruhi tingkat akurasi sebuah algoritma (Alpaydin 2010).

Salah satu proses untuk memilih atribut yang akan digunakan dalam proses klasifikasi adalah dengan melakukan pre processing yaitu seleksi fitur. Seleksi fitur adalah perlakuan terhadap dataset untuk menghitung kepentingan seluruh atribut data yang ada. Seleksi fitur ini dapat menghasilkan tingkat kepentingan atau biasa disebut dengan bobot untuk semua atribut dalam dataset. Atribut dengan nilai bobot tinggi berarti memiliki kepentingan yang tinggi, begitu juga sebaliknya. information gain ratio merupakan metode seleksi fitur yang banyak digunakan dan terbukti dapat menangani dataset berdimensi tinggi (Koprinska 2010). Metode ini merupakan pembaruan dari metode lama yaitu information gain. Penelitian akan menggunakan information gain ratio untuk optimasi algoritma naive bayes dalam melakukan klasifikasi dataset berdimensi tinggi.

Dalam penelitian ini digunakan dataset publik terpopuler yang biasa digunakan dalam penelitian klasifikasi data mining. Data publik yang akan digunakan diambil dari UCI Machine Learning Repository. UCI Repository merupakan penyedia dataset publik yang telah teruji dan banyak digunakan dalam penelitian tingkat internasional. Website dari penyedia dataset dapat diakses di: https://archive.ics.uci.edu/ml/index.php

2. KAJIAN LITERATUR DAN PENGEMBANGAN HIPOTESIS

Naïve bayes merupakan salah satu algoritma data mining terbaik dan banyak digunakan dalam proses klasifikasi (Wu 2009). Banyak penelitian yang melakukan komparasi algoritma klasifikasi untuk berbagai jenis dataset. Beberapa penelitian menghasilkan naïve bayes sebagai algoritma terbaik. Diantaranya adalah:

1) Tahun 2017 dilakukan komparasi algoritma data mining untuk klasifikasi penyakit kanker payudara (Kurniawan and Ivandari 2017). Penelitian ini membandingkan algoritma K-Nearest Neighbour, Decission Tree C4.5 serta naïve bayes untuk deteksi penyakit kanker payudara. Dataset yang digunakan dalam penelitian ini adalah data breast cancer winconsin yang merupakan dataset public dari UCI Machine Learning Repository. Hasil dari penelitian ini naïve bayes merupakan algoritma terbaik dengan tingkat akurasi sebesar 95,85%. Sedangkan algoritma K-Nearest Neighbour mendapatkan tingkat akurasi sebesar 94,71% dan Decission Tree C4.5 sebesar 94,70%.
2) Peningkatan akurasi algoritma KNN dengan seleksi fitur gain ratio untuk klasifikasi penyakit diabetes mellitus (Indrayanti, Devi, and Al Karomi 2017). Algoritma information gain ratio terbuki dapat meningkatkan performa algoritma klasifikasi dengan melakukan pre processing seleksi fitur. Penelitian yang dilakukan Indrayanti dkk pada tahun 2017 ini menggunakan dataset public diabetes mellitus. Hasil penelitian ini membuktikan gain ratio dapat menaikkan tingkat akurasi algoritma klasifikasi.

Dari beberapa literatur yang ada dapat disimpulkan bahwa penggunaan information gain ratio untuk seleksi fitur dapat meningkatkan performa algoritma. Penelitian ini menerapkan metode information gain ratio untuk meningkatkan performa algoritma naive bayes. Perhitungan ini dilakukan khususnya untuk dataset berdimensi tinggi.

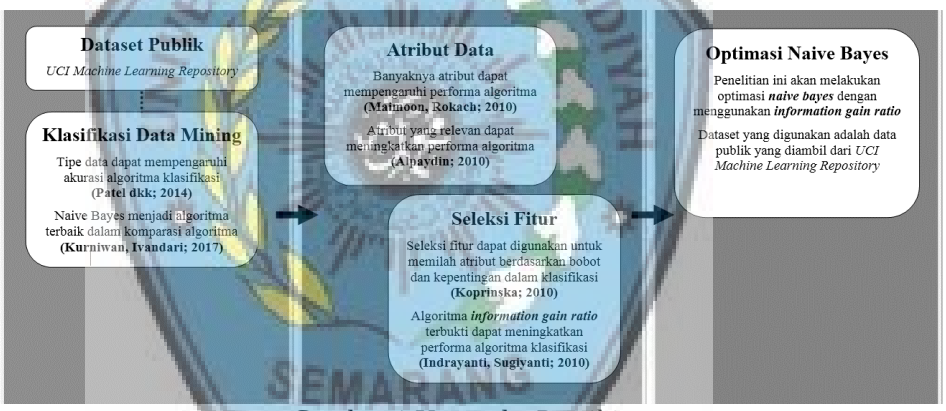
3. METODE PENELITIAN

Metode penelitian dalam penelitian ini adalah eksperimental. Dalam penelitian ini akan dilakukan pengukuran akurasi semua dataset yang ada menggunakan algoritma naïve bayes. Setelah diketahui performa algoritma dalam melakukan klasifikasi pada tiap dataset, berikutnya akan dilakukan optimasi menggunakan information gain ratio untuk meningkatkan akurasi algoritma serta memperbaiki performa dari naïve bayes. Gambar 1 berikut merupakan kerangka pemikiran dalam penelitian ini:

3.1 Pengumpulan Data

Dalam tahapan ini akan dikumpulkan dataset yang merupakan data terpopuler yang digunakan dalam proses penelitian. Data yang digunakan dalam penelitian ini adalah dataset public dari UCI repository. UCI repository merupakan salah satu sumber dataset terpercaya yang menyediakan lebih dari 440 dataset machine learning. Dataset dari UCI banyak digunakan oleh peneliti bidang ilmu komputer untuk menguji metode atau model suatu algoritma. Dataset tersebut dapat diunduh pada laman: https://archive.ics.uci.edu/ml/datasets.html.

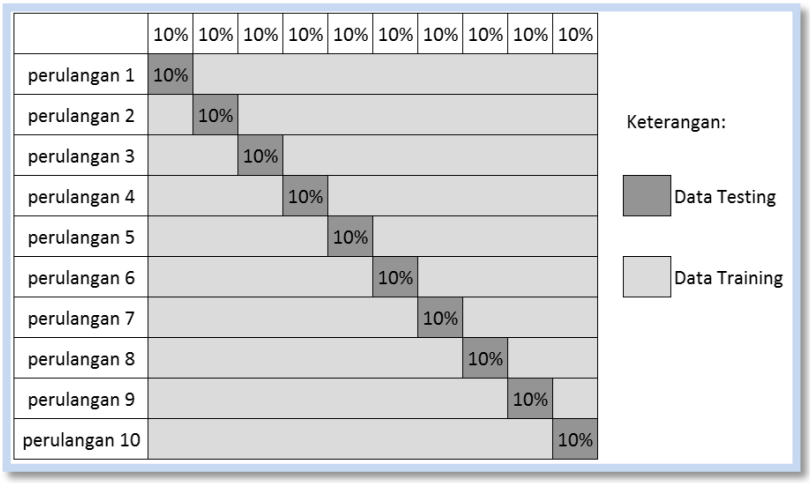
Selain pengumpulan dataset, dalam tahapan ini juga akan dilakukan analisa terhadap semua dataset yang telah diperoleh. Analisa tersebut terkait jenis atribut serta tipe dataset terpilih. Analisa ini juga didasarkan dari referensi terkait yaitu dari artikel ilmiah yang membahas tentang semua dataset terpilih.



Gambar 1 Kerangka Pemikiran

3.2 Desain dan Pemodelan Algoritma

Dalam tahapan desain dan pemodelan algoritma ini akan dilakukan menggunakan pre processing seleksi fitur dilanjutkan dengan klasifikasi naive bayes. Dalam tahapan ini juga akan digunakan X-Validation untuk melakukan validasi dataset. Validasi dilakukan dengan cara membagi data menjadi 10 bagian, 9 diantaranya dijadikan data training dan 1 bagian yang lain digunakan sebagai data testing. Proses ini diulang sebanyak 10 kali sehingga semua bagian data pernah menjadi data testing. Proses ini biasa disebut dengan 10 folds cross validation. Gambar 2 adalah representasi dari 10 folds cross validation.



Gambar 2 representasi 10 folds cross validation

3.3 Seleksi Fitur dan Perhitungan Akurasi Algoritma

Tahapan selanjutnya adalah seleksi fitur menggunakan information gain ratio. Dalam tahap ini akan dilakukan perhitungan bobot seluruh atribut data. Selanjutnya diterapkan treshold atau batas dari bobot yang akan digunakan. Atribut dengan bobot diatas treshold selanjutnya akan digunakan dalam proses klasifikasi. Sedangkan atribut yang memiliki bobot dibawak treshold nantinya tidak akan digunakan dalam proses klasifikasi. Aplikasi bantu dalam penelitian ini adalah rapid miner.

Proses perhitungan tingkat akurasi menggunakan confussion matrix atau matrix kebingungan. Dalam matrix ini nantinya akan muncul klasifikasi yang sesuai dengan data asli dan yang tidak sesuai dengan aslinya. Proses perhitungan akurasi didasarkan pada matrix ini. Semakin banyak klasifikasi yang sesuai dengan data asli maka akurasi semakin baik.

3.4 Pengujian

Dalam tahapan ini akan dicatat perubahan akurasi yang terjadi pada algoritma klasifikasi naive bayes. Perbandingan ini dilakukan antara hasil akurasi naive bayes menggunakan seluruh atribut dataset dengan hanya menggunakan atribut terpilih setelah dilakukan seleksi fitur information gain ratio. Pencatatan dilakukan terhadap seluruh dataset untuk mengetahui peningkatan performa naive bayes dengan melakukan seleksi fitur information gain ratio. Dalam tahapan pengujian ini akan diketahui optimasi algoritma naive bayes dengan menggunakan algoritma information gain ratio.

4. HASIL PENELITIAN

Penelitian ini menggunakan dataset publik yang diambil dari uci repository. Dataset yang digunakan antara lain breast cancer dataset, primary tumor dataset serta lympography dataset. Tabel 1 merupakan hasil akurasi klasifikasi naive bayes dari dataset breast cancer. Sedangkan tabel 2 merupakan hasil akurasi naive bayes untuk dataset breast cancer dengan seleksi fitur information gain ratio.

Tabel 1. Hasil akurasi klasifikasi naive bayes dataset breast cancer

	true no-recurrence-events	true recurrence-events	class precision
pred. no-recurrence-events	164	47	77.73%
pred. recurrence-events	37	38	50.67%

class recall	81.59%	44.71%	**70.64%**

Tabel 2. Hasil akurasi klasifikasi naive bayes + IGR dataset breast cancer

	true no-recurrence-events	true recurrence-events	class precision
pred. no-recurrence-events	167	43	79.52%
pred. recurrence-events	34	42	55.26%
class recall	83.08%	49.41%	**73.07%**

Dari hasil yang terperinci pada tabel 1 diatas, diketahui bahwa tingkat akurasi klasifikasi naive bayes untuk dataset breast cancer adalah 70,64%. Sedangkan dari tabel 2 diketahui bahwa tingkat akurasi naive bayes untuk dataset breast cancer setelah dilakukan seleksi fitur menggunakan information gain ratio adalah 73,07%. Artinya untuk dataset breast cancer ini seleksi fitur information gain ratio dapat memberikan peningkatan akurasi. Dalam penggunaan information gain ratio ini digunakan treshold 0,06.

Untuk dataset primary tumor dilakukan perhitungan dengan kondisi yang sama. Hasil akurasi dari naive bayes dapat dilihat pada tabel 3. Sedangkan hasil naive bayes setelah disisipi information gain ratio pada dataset primary tumor dapat dilihat pada tabel 4. Hasil perhitungan juga menunjukkan hal yang sama yaitu peningkatan nilai akurasi naive bayes setelah dataset dilakukan seleksi fitur. Nilai treshold yang digunakan dalam information gain ratio ini adalah 0,06.

Tabel 3. Hasil akurasi klasifikasi naive bayes dataset primary tumor

	true 1.0	true 2.0	class precision
pred. 1.0	81	3	96.43%
pred. 2.0	3	17	85.00%
class recall	96.43%	85.00%	**94.09%**

Tabel 4. Hasil akurasi klasifikasi naive bayes + IGR dataset primary tumor

	true 1.0	true 2.0	class precision
pred. 1.0	81	0	100.00%
pred. 2.0	3	20	86.96%
class recall	96.43%	100.00%	**97.09%**

Peningkatan nilai akurasi untuk datasetprimary tumor yaitu 3%. Sebelumnya tingkat akurasi naive bayes untuk dataset primary tumor adalah 94,09%. Setelah dilakukan seleksi fitur dengan menggunakan information gain ratio dengan treshold 0,06 nilai akurasi naive bayes meningkat menjadi 97,09%.

Dataset lymphography memiliki 4 varian label yaitu: 1.0, 2.0, 3.0, serta 4.0. hasil penelitian menunjukkan bahwa information gain ratio juga dapat memberikan peningkatan akurasi naive bayes untuk dataset lymphography ini. Treshold yang digunakan adalah 0,02. Secara lebih terperinci hasil tingkat akurasi naive bayes untuk dataset lymphography dapat dilihat pada tabel 5 dan tabel 6 berikut.

Tabel 5. Hasil akurasi klasifikasi naive bayes dataset lymphography

	true 3.0	true 2.0	true 4.0	true 1.0	class precision
pred. 3.0	40	12	2	0	74.07%
pred. 2.0	20	69	1	2	75.00%
pred. 4.0	1	0	1	0	50.00%

pred. 1.0	0	0	0	0	0.00%
class recall	65.57%	85.19%	25.00%	0.00%	**74.38%**

Tabel 6. Hasil akurasi klasifikasi naive bayes + IGR dataset lymphography

	true 3.0	true 2.0	true 4.0	true 1.0	class precision
pred. 3.0	41	10	3	0	75.93%
pred. 2.0	19	71	0	2	77.17%
pred. 4.0	1	0	1	0	50.00%
pred. 1.0	0	0	0	0	0.00%
class recall	67.21%	87.65%	25.00%	0.00%	**76.29%**

5. SIMPULAN

Dari hasil penelitian yang dilakukan dapat disimpulkan bahwa penggunaan information gain ratio untuk seleksi fitur dapat meningkatkan performa algoritma naive bayes. Peningkatan yang ada bervariasi dan juga dipengaruhi oleh treshold yang digunakan.

6. REFERENSI

Alpaydin, Ethem. 2010. *Introduction to Machine Learning Second Edition*. London: The MIT Press.

Amancio, D. R., C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. a. Rodrigues, and L. Da F. Costa. 2013. "A Systematic Comparison of Supervised Classifiers," October. http://arxiv.org/abs/1311.0202v1.

Ashari, Ahmad, Iman Paryudi, and A Min Tjoa. 2013. "Performance Comparison between Naïve Bayes , Decision Tree and K-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool" 4 (11): 33–39.

Azhagusundari, B, and Antony Selvadoss Thanamani. 2013. "Feature Selection Based on Information Gain," no. 2: 18–21.

Brammer, Max. 2007. *Principles of Data Mining*. London: Springer.

Christobel, Angeline, and D.r Sivaprakasam. 2011. "An Empirical Comparison of Data Mining Classification Methods" 3 (2): 24–28.

Deng, Houtao, and George Runger. 2012. "Feature Selection via Regularized Trees," January. http://arxiv.org/abs/1201.1587v3.

Gallager, Robert G, and Life Fellow. 2001. "Claude E . Shannon : A Retrospective on His Life , Work , and Impact" 47 (7): 2681–95.

Han, Jiawei, and Micheline Kamber. 2006. *Data Mining: Concepts and Techniques Second Edition*. *Elsevier*. Elsevier.

Hastuti, Khafiizh. 2012. "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Mahasiswa Non Aktif" 2012 (Semantik): 241–49.

Ian H Witten. Eibe Frank. Mark A Hall. 2011. *Data Mining 3rd*.

Indrayanti, Indrayanti, Sugianti Devi, and M. Adib Al Karomi. 2017. "Peningkatan Akurasi Algoritma KNN Dengan Seleksi Fitur Gain Ratio Untuk Klasifikasi Penyakit Diabetes Mellitus." *IC-TECH* XIII (2): 1–6. ejournal.stmik-wp.ac.id.

Jiawei Han and Micheline Kaber. 2006. "Data Mining:Concepts and Techniques." *University of Illinois at Urbana-Champaign*.

Koprinska, Irena. 2010. "Feature Selection for Brain-Computer Interfaces," 100–111.

Kurniawan, M. Faisal, and Ivandari. 2017. "Komparasi Algoritma Data Mining Untuk Klasifikasi Kanker Payudara." *IC Tech* I April 20: 1–8.

Kusrini, Sri Hartati, Retantyo Wardoyo, and Agus Harjoko. 2009. "Perbandingan

Metode Nearest Neighbor Dan Algoritma c4.5 Untuk Menganalisis Kemungkinan Pengunduran Diri Calon Mahasiswa Di Stmik Amikom Yogyakarta" 10 (1).

Larose, Daniel T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons.

Maimoon, Oded, and Lior Rokach. 2010. *Data Mining and Knowledge Discovery Handbook*. Vol. 40. Springer. doi:10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.

Novakovic, Jasmina. 2010. "The Impact of Feature Selection on the Accuracy of 1DwYH Bayes Classifier" 2: 1113–16.

Patel, Kanu, Jay Vala, and Jaymit Pandya. 2014. "Comparison of Various Classification Algorithms on Iris Datasets Using WEKA" 1 (1): 1–7.

Prasetyo, Eko. 2012. *Data Mining Konsep Dan Aplikasi Menggunakan Matlab*. Yogyakarta: Andi Offset.

Pudjianto, Tacbir Hendro, Faiza Renaldi, and Age Teogunadi. 2011. "Penerapan Data Mining Untuk Menganalisa Kemungkinan Pengunduran Diri Calon Mahasiswa Baru."

Ragab, Abdul Hamid M., Amin Y. Noaman, Abdullah S. Al-Ghamdi, and Ayman I. Madbouly. 2014. "A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining." *Proceedings of the 2014 Workshop on Interaction Design in Educational Environments - IDEE '14*. New York, New York, USA: ACM Press, 106–13. doi:10.1145/2643604.2643631.

Santosa, Budi. 2007. *Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis*. Edisi Pert. Yogyakarta: Graha Ilmu.

Susanto, Sani, and Dedi Suryadi. 2010. *Pengantar Data Mining: Menggali Pengetahuan Dari Bongkahan Data*. Yogyakarta: Andi Offset.

Widiastuti, Dwi. 2007. "Analisa Perbandingan Algoritma SVM, Naïve Bayes, Dan Decission Tree Dalam Mengklasifikasikan Serangan (Attack) Pada Sistem Pendeteksi Intrusi." *Jurusan Sistem Informasi Universitas Gunadarma*, 1–8.

Witten, Ian H, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition*. Elsevier.

Wu, Xindong. 2009. *The Top Ten Algorithms in Data Mining*. Edited by Vipin Kumar. New York: Taylor & Francis Group, LLC.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, et al. 2007. *Top 10 Algorithms in Data Mining*. *Knowledge and Information Systems*. Vol. 14. doi:10.1007/s10115-007-0114-2.