KLASIFIKASI DATA MINING MENGGUNAKAN NAÏVE BAYES *CLASSIFIER* DENGAN ALGORITMA C5.0

(Classification Data Mining using Naïve Bayes Classifier with C5.0 Algorithm)

Aini Ayu Wulandari[1*)], Dewi Retno Sari Saputro[2)]

[1,2)] Program Studi Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sebelas Maret, Jl. Ir Sutami No. 36A, Surakarta 57126, Indonesia
e-mail: ainiayu111@student.uns.ac.id[1*)], dewiretnoss@staff.uns.ac.id[2)]
*)penulis korespondensi

Abstract. Data mining is a process of detecting interesting patterns and knowledge from large amounts of data. Data mining has several tasks, one of them is classification. Classification is a process of grouping data into certain classes based on the variables. There are various methods to complete classification. The method that is often used for classification is Naïve Bayes Classifier (NBC). This is because NBC is considered an easy and efficient method. NBC is a combination of naïve (the condition between variables are assumed to be independent) and Bayes theorem. The assumption of independent variables in NBC can sometimes result in unfavorable results in classification. This can be avoided by adding the C5.0 algorithm to NBC. C5.0 algorithm is an algorithm that is useful for selecting variables based on information gain value. The algorithm is run before classifying with NBC. This study discusses about theory of classification using NBC with C5.0 algorithm. C5.0 algorithm added to NBC can optimize classification and know the most influential variables.

Keywords: C5.0 Algorithm, Classification, Data Mining, Naïve Bayes Classifier.

1. Pendahuluan

Data mining merupakan proses tentang memecahkan masalah dengan menganalisis suatu data [14]. Menurut Han *et. al*. [2], sumber data dapat mencakup *database*, data *warehouse*, *web*, penyimpanan informasi lain, atau data yang otomatis tersimpan di dalam sistem secara dinamis. Dalam data mining terdapat berbagai teknik seperti klasifikasi, asosiasi, *clustering*, prediksi, estimasi, dan analisis deviasi [12]. Klasifikasi merupakan proses mengelompokkan suatu data menjadi kelas-kelas tertentu dengan memanfaatkan struktur data. Pengklasifikasian data dapat diselesaikan dengan berbagai metode, contohnya Naïve Bayes *Classifier* (NBC). Penggunaan metode dalam klasifikasi berguna untuk mempermudah proses klasifikasi.

NBC merupakan gabungan antara naïve dan teorema Bayes. Teorema tersebut dikemukakan oleh seorang ilmuwan asal Inggris bernama Thomas Bayes yang menyatakan bahwa kejadian di masa depan didasarkan pada kejadian masa lalu. Menurut Muller dan Guido [5], NBC adalah metode yang efisien karena NBC melihat setiap fitur (variabel) secara individual dan mengumpulkan statistik kelas per kelas dari setiap

variabel. Menurut Rish [10], dalam prakteknya NBC bersaing baik dengan metode lain walaupun asumsi saling bebas umumnya kurang baik. Tidak bisa dipungkiri jika asumsi saling bebas pada NBC dapat mengakibatkan hasil klasifikasi yang tidak selalu baik. Untuk mengatasi hal tersebut ditambahkan algoritma C5.0 pada NBC. Algoritma tersebut bertujuan mendapatkan variabel yang relevan sebelum dilakukan klasifikasi menggunakan NBC.

Algoritma C5.0 merupakan salah satu algoritma yang digunakan dalam *Decision Tree* (*DT*). Hasil yang baik diperoleh ketika menggabungkan *DT* dengan metode lain. Hal ini sejalan dengan penelitian yang dilakukan oleh Cardie [1] serta Ratanamahatana dan Gunopulos [11]. Penelitian yang dilakukan Cardie [1] pada tahun 1993 adalah menggunakan *DT* untuk improvisasi *Case-Based Learning* (*CBL*) dan hasilnya adalah *CBL* memiliki performa yang lebih baik dibandingkan dengan hanya menggunakan *CBL* atau *DT*. Pada tahun 2003, Ratanamahatana dan Gunopulos [11] meneliti tentang penggunaan algoritma C4.5 pada NBC. Hasilnya menunjukkan bahwa kombinasi tersebut membutuhkan *training* lebih sedikit untuk mencapai akurasi yang tinggi pada klasifikasi. Pada tahun 2015, Pandya dan J. Pandya [6] meneliti tentang penggunaan algoritma C5.0 untuk seleksi fitur dan mengurangi kesalahan teknik pemangkasan. Pada penelitian ini dilakukan kajian kembali teori tentang klasifikasi data mining menggunakan NBC dengan algoritma C5.0.

2. Metodologi

Penelitian ini merupakan penelitian teori yakni melakukan kajian terhadap teori NBC dan algoritma C5.0. Langkah-langkah yang dilakukan dalam penelitian ini adalah mempelajari dan menganalisis kajian pustaka dari beberapa referensi buku, jurnal, serta tulisan terkait teori tentang NBC dan algoritma C5.0.

3. Hasil dan Pembahasan

3.1 Naïve Bayes *Classifier* (NBC)

NBC adalah suatu metode probabilitas dan statistik yang digunakan untuk klasifikasi. NBC terbukti memiliki kecepatan yang tinggi saat diaplikasikan pada data dengan jumlah besar [3]. Sebagaimana ditulis oleh Mountassir *et. al*. [4], NBC tidak sensitif terhadap data set yang tidak seimbang. NBC menggunakan perhitungan probabilitas sederhana yang didasarkan pada teorema Bayes. Menurut Wilmott [13], jika terdapat kejadian terpisah yang dimisalkan C dan X maka rumus Bayes dituliskan sebagai

$$P(X) = \frac{P(C)P(C)}{P(X)} \qquad (1)$$

dengan C merupakan kelas, X merupakan variabel, $P(C|X)$ adalah probabilitas *posterior* untuk C, $P(X|C)$ merupakan *likelihood*, $P(C)$ adalah probabilitas awal (*prior*) C, dan $P(X)$ adalah probabilitas awal (*prior*) X atau *evidence*.

Probabilitas *posterior* adalah probabilitas setelah adanya sampel variabel tertentu dalam kelas C. Nilai probabilitas tersebut akan digunakan untuk menentukan klasifikasi di kelas mana sampel variabel berada. *Likelihood* adalah probabilitas munculnya sampel variabel pada kelas C. Probabilitas *prior* C adalah probabilitas sebelum adanya variabel tertentu dalam kelas C. Probabilitas *prior* X adalah probabilitas munculnya variabel secara luas atau disebut *evidence*. Pada NBC memerlukan $X = (X_1, \cdots, X_n)$ yang merupakan karakteristik petunjuk (variabel) dan kelas C_k dengan $k = 1, 2, \cdots, m$ sehingga rumus (1) menjadi

$$P(X_1, \cdots, X_n) = \frac{P(C_k)P(C_k)}{P(X_1, \cdots, X_n)}. \qquad (2)$$

Oleh karena NBC memiliki asumsi bahwa variabel dalam tiap kelas saling bebas yang kuat (*naïve*) maka $P(C_k)$ ditulis sebagai

$$P(C_k) = P(C_k) \times \cdots \times P(C_k) = \prod_{i=1}^{n} \quad P(X_i|C_k). \qquad (3)$$

Dalam NBC, *evidence* ($P(X_1, \cdots, X_n)$) selalu konstan untuk setiap kelas sehingga rumus (2) dan (3) dapat ditulis menjadi

$$P(X_1, \cdots, X_n) = P(C_k) \prod_{i=1}^{n} \quad P(X_i|C_k).$$

Langkah-langkah klasifikasi dengan NBC diuraikan sebagai berikut.

1. Membagi data menjadi data *training* dan data *testing*.
2. Pada data *training* dilakukan perhitungan probabilitas *prior* ($P(C_k)$) dan *likehood* ($P(C_k)$) setiap variabel.
3. Menghitung probabilitas *posterior* ($P(X_1, \cdots, X_n)$) pada data *testing* dimana penentuan probabilitas *prior* dan *likelihood* didasarkan pada hasil perhitungan dari data *training*.
4. Memilih kelas dengan probabilitas *posterior* yang besar untuk menentukan dikelas mana suatu data diklasifikasikan.

3.2 *Decision Tree* (*DT*)

Menurut Han *et. al*. [2], *Decision Tree* (*DT*) adalah suatu struktur *flowchart* yang menyerupai pohon. Konsep dasar *DT* adalah mengubah data menjadi pohon keputusan sehingga hasil keputusan dapat diinterpretasikan dengan lebih mudah. *DT* memiliki tiga jenis *node* yang diuraikan sebagai berikut.

1. *Root node* adalah *node* paling atas, pada node ini tidak terdapat *input* dan bisa tidak memiliki *output* atau *output* lebih dari satu.
2. *Internal node* adalah *node* percabangan, pada *node* ini hanya terdapat satu *input* dan memiliki *output* minimal dua.
3. *Leaf node* atau terminal *node* adalah *node* akhir, pada *node* ini hanya terdapat satu *input* dan tidak memiliki *output*.

Terdapat beberapa algoritma yang digunakan dalam *DT*. Quinlan [9] memperkenalkan algoritma ID3, kemudian algoritma ID3 berkembang menjadi algoritma C4.5 dan akhirnya menjadi algoritma C5.0.

3.3 Algoritma C5.0

Algoritma C5.0 merupakan pengembangan dari algoritma C4.5. Sebagaimana ditulis oleh Patel dan Rana [7], algoritma C5.0 lebih baik dari algoritma C4.5 dalam hal kecepatan, memori, dan efisiensi. Pada algoritma C5.0 menggunakan memori yang sedikit. Patel dan Rana [7] menyatakan bahwa algoritma C5.0 merupakan algoritma klasifikasi yang dapat diterapkan pada data set yang besar. Menurut Patil *et. al.* [8] algoritma C5.0 dapat dengan mudah mengatasi *multi value* atribut dan *missing* atribut dari data set. Berikut uraian algoritma C5.0.

1. Mengelompokkan variabel ke dalam kelas-kelas tertentu.
2. Menghitung nilai *entropy* dan *information gain*. Berikut merupakan rumus untuk menghitung *entropy*.
$$entropy(S) = -\sum_{k=1}^{m} \frac{|S_k|}{|S|}\frac{|S_k|}{|S|}$$
dengan $|Sk|$ merupakan jumlah sampel kelas k, $|S|$ adalah jumlah sampel, dan m adalah banyaknya kelas k. *Entropy* setiap variabel dihitung dengan rumus yang ditulis sebagai
$$entropy_X(S) = \sum_{j=1}^{p} \frac{|S_j|}{|S|} \times entropy(S_j)$$
dengan S_j adalah sampel variabel X yang dipartisi sebanyak j dan p merupakan banyaknya partisi j. Untuk menghitung *information gain* digunakan rumus yang ditulis sebagai
$$Gain(X) = entropy(S) - entropy_X(S).$$
3. Memilih variabel yang memiliki nilai *information gain* yang maksimal untuk dijadikan sebagai *node* akar dan variabel dengan nilai *information gain* yang maksimal berikutnya sebagai *node* cabang.
4. Perhitungan nilai *information gain* akan tetap dilakukan sampai setiap data memperoleh kelas dan variabel yang sudah terpilih tidak dihitung lagi.

3.4 Penambahan Algoritma C5.0 pada Naïve Bayes *Classifier* (NBC)

Pada NBC mengasumsikan bahwa setiap variabelnya saling bebas. Asumsi tersebut dapat mengakibatkan klasifikasi yang kurang baik sehingga nilai akurasi pun akan mengalami penurunan. Oleh karena itu, algoritma C5.0 ditambahkan pada NBC. Algoritma C5.0 dipilih karena menurut Patel dan Rana [7], algoritma C5.0 memiliki kecepatan, memori, dan efisiensi yang baik. Algoritma tersebut menggunakan nilai *information gain* untuk mendapatkan variabel yang relevan. Variabel tersebut berguna untuk klasifikasi menggunakan NBC. Penambahan algoritma C5.0 pada NBC dapat mengoptimalkan klasifikasi dan dapat mengetahui variabel yang paling berpengaruh.

4. Kesimpulan

Berdasarkan pembahasan diperoleh kesimpulan bahwa untuk mengatasi NBC yang memiliki asumsi setiap variabelnya saling bebas, ditambahkan algoritma C5.0 pada NBC. Algoritma C5.0 digunakan untuk mendapatkan variabel yang relevan, kemudian dilakukan klasifikasi menggunakan NBC. Algoritma C5.0 ditambahkan pada NBC dapat mengoptimalkan klasifikasi dan mengetahui variabel yang paling berpengaruh

Daftar Pustaka

[1] Cardie, C., (1993), Using Decision Tree to Improve Case-Based Learning, *Proceedings of the Tenth International Conference of Machine Learning*, ISBN: 978-1-55860-307-3, PP: 25-32.

[2] Han, J., Kamber, M., Pei, J., (2012), *Data Mining Concepts and Techniques*, 3rd ed., Morgan Kaufmann Publisher.

[3] Han, J. and Kamber, M., (2006), *Data Mining Concepts and Techniques*, 2nd ed., Morgan Kaufmann Publishers.

[4] Mountassir.A, Benbrahin.H, Berrada.I, (2012), An Empirical Study to Address The Problem of Unbalanced Data Sets in Sentiment Classification, *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 3298-3303, https://doi.org/10.1109/ICSMC.2012.6378300.

[5] Muller, A. C. & Guido, S., (2017), *Introduction to Machine Learning with Python*, O'Reilly Media Inc.

[6] Pandya, R. and Pandya, J., (2015), C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning, *International Journal of Computer Applications* **117(16)**, 18-21, https://doi.org/10-5120/20639-3318.

[7] Patel, B.R. & Rana, K.K., (2014), A Survey on Decision Tree Algorithm for Classification, *International Journal of Engineering Development and Research (IJEDR)* **2**, 1-5.

[8] Patil, N., Lathi, R., Chitre, V., (2012), Comparison of C5.0 CART Classification Algorithms Using Pruning Technique, *International Journal of Engineering Research Technology* (IJERT), **1**, 1-5.

[9] Quinlan, J.R., (1986), Induction of Decision Trees, *Machine Learning* 1, 81-106.

[10] Rish, I., (2001), An Empirical Study of The Naïve Bayes Classifier, *Proceeding IJCAI*, PP: 41-46.

[11] Ratanamahatana, C.”ann”. & Gunopulos, D., (2003), Feature Selection for The Naïve Bayes Classifier Using Decision Trees, *Applied Artificial Intelligence* **17(5)**, 475-487, https://doi.org/10.1080/713827175.

[12] Thuraisingham, B., (2000), A primer for understanding and applying data mining, *IT Professional* **2(1)**, 28–31, https://doi.org/10.1109/6294.819936.

[13] Wilmott, P., (2019), *Machine Learning: An Applied Mathematics Introduction*, 1st ed., Panda Ohana Publishing.

[14] Witten, I. H., Frank, E., and Hall, M. A., (2011), *Data Mining Practical Machine Learning Tools and Techniques*, 3rd ed., Morgan Kaufmann Publishers, USA.