PELITA TEKNOLOGI

Journal homepage: jurnal.pelitabangsa.ac.id, p-ISSN: 2301-475X, e-ISSN: 2656-7059

Optimasi Genetic Algorithm pada Naïve Bayes untuk Klasifikasi Pengajuan Kredit Bank

Donny Maulana[1],Yoga Religia[2] Nanang Tedi[3]

[1,2,3]Program Studi Teknik Informatika, Universitas Pelita Bangsa
[1,2,3]Bekasi, Indonesia
[1]donny.maulana@pelitabangsa.ac.id, [2]yoga.religia@pelitabangsa.ac.id,
[3]nanang@pelitabangsa.ac.id,

Abstrak

The selection of prospective customers who apply for credit in the banking world is very important to be considered by the marketing department in order to avoid credit problems. Currently, the website www.kaggle.com has provided South German Credit data consisting of 22 attributes, 1 label and 25976 instances which are included in the supervised learning data category. Based on several previous studies, the Naïve Bayes algorithm can provide better classification performance than other algorithms. Several studies also mention that the use of Naive Bayes can be optimized using Genetic Algorithm (GA) to obtain better performance. This study aims to compare the use of the Naive Bayes algorithm for the classification of South German Credit with and without GA optimization. The data validation process used in this research is using split validation, dividing the dataset is 95% training data and 5% testing data. The test results show that the use of GA in Naive Bayes is able to improve the performance of South German Credit data classification in terms of accuracy and recall with an accuracy value of 85.99% and recall of 87.91%.

Informasi Artikel

Diterima: 14-08-2021
Direvisi: 19-09-2021
Dipublikasikan: 26-10-2021

Keywords

Data Mining, Klasifikasi, Naïve Bayes, Genetic Algorithm

I. Pendahuluan

Bagian *marketing* perbankan perlu menyeleksi calon nasabahnya untuk mengetahui pelanggan mana yang dapat diberikan pembiayaan kredit dengan mempertimbangkan berbagai faktor. Pembiayaan kredit adalah penyedian dana oleh pihak bank kepada nasabah berdasarkan persetujuan pinjam meminjam yang mewajibkan agar nasabah melunasi pinjamannya pada jangka waktu tertentu [1]. Oleh sebab itu, seleksi calon nasabah dibutuhkan agarseorang *marketing* perbankan mampu menjaga nasabahnya supaya tidak mengalami kredit bermasalah. Cara yang dapat digunakan untuk mengetahui kepuasan pelanggan adalah menggunakan teknik data mining dengan model klasifikasi [2].

Saat ini situs www.kaggle.com telah menyediakan set data *South German Credit* yang terdiri dari 21 atribut dengan 800 *instances* pengajuan kredit dan tidak terdapat *missing value*, sehingga dapat digunakan untuk membangun model klasifikasi kelayakan kredit [3]. Atribut label yang terdapat pada data *South German Credit* adalah atribut "Kredit" dengan 600 *instances* dengan keterangan "diterima" dan 200 *instances* dengan dengan keterangan "ditolak", sehingga menjadikan data *South German Credit* termasuk *imbalance* data.

Dibutuhkan suatu algoritma yang baik untuk pembuatan model klasifikasi yang optimal, salah satunya menggunakan algoritma Naïve Bayes. Berdasarkan beberapa penelitian terdahulu,algoritma Naïve Bayes dapat memberikan performa klasifikasi yang lebih baik dibandingkan algoritmaklasifikasi yang lain seperti k-NN, C4.5, Decision Tree, bahkan Neural Network [4] [5] [6]. Selain dapat memberikan performa klasifikasi yang baik, algoritma Naïve Bayes juga dapat digunakan pada data *imbalance* [7] [8], sehingga cocok digunakan untuk mengklasifikasi-kan data *South German Credit*.

Meskipun Naïve Bayes telah menunjukkan akurasi klasifikasi yang luar biasa, namun saat ini asumsi independen jarang dibahas pada klasifikasi Naïve Bayes. Salah satu cara untuk mencoba asumsi independen pada algoritma Naïve Bayes adalah dengan pembobotan atribut [9]. Hal tersebut didukung pula oleh Liangxiao Jiang (2019) yang menyebutkan bahwa perlu diusulkan metode pembobotan atribut untuk mengurangi asumsi independen [10]. Pembobotan atribut dapat dilakukan menggunakan *Genetic Algorithm* (GA) melalui *Feature Selection* [11].

GA merupakan salah satu algoritma optimasi yang dibuat untuk meniru beberapa proses yang diamati dalam evolusi alam [12]. Optimasi yang dilakukan oleh GA adalah dengan memprediksi jumlah iterasi yang tepat, sehingga tidak diperlukan lagi perhitungan dengan jumlah iterasi yang berbeda untuk mendapatkan kemunculan yang lengkap dari jalur independen [13]. Keuntungan paling signifikan dari GA adalah kemampuannya dalam pencarian global serta kemampuan beradaptasi terhadap spektrum masalah yang luas [14]. Berdasarkan beberapa penelitian sebelumnya menyebutkan bahwa penggunaan GA mampu meningkatkan performa klasifikasi dari Naïve Bayes [15] [16].

Berdasarkan penelitian sebelumnya menunjukkan bahwa GA mampu untuk

meningkatkan performa klasifikasi pada Naïve Bayes, akan tetapi belum ditemukan penerapan GA pada Naïve Bayes untuk klasifikasi kepuasan pelanggan maskapai penerbangan. Penelitian ini melakukan analisa optimasi GA pada Naïve Bayes untuk untuk klasifikasi data *South German Credit.*

II. Metodologi

A. Data yang digunakan

Penelitian ini menggunakan data sekunder berupa set data *South German Credit* yang diambil dari situs www.kaggle.com [3]. Adapun jumlah *instances* data yang terdapat pada data *South German Credit* adalah sebanyak 800 *instances* yang terdiri dari 21 atribut dan tidak terdapat *missing value*, sehingga tidak memerlukan *pre-processing* data. Berdasarkan 21 atribut yang ada, terdapat 1 atribut label yang terdapat pada data *South German Credit* yaitu atribut “Kredit”. Pada label tersebut terdapat 600 *instances* dengan keterangan “good” dan 200 *instances* dengan dengan keterangan “bad”, sehingga menjadikan data *South German Credit* termasuk *imbalance* data.

Tabel 1. Atribut Data *South German Credit*

Atribut	Keterangan
status	status of the debtor's checking account with the bank (categorical)
duration	credit duration in months (quantitative)
credit history	history of compliance with previous or concurrent credit contracts (categorical)
purpose	purpose for which the credit is needed (categorical)
amount	credit amount in DM (quantitative; result of monotonic transformation; actual data and type of trans...
savings	debtor's savings (categorical)
employment duration	duration of debtor's employment with current employer (ordinal; discretized quantitative)
installment rate	credit installments as a percentage of debtor's disposable income (ordinal; discretized quantitative...
personal status sex	combined information on sex and marital status; categorical; sex cannot be recovered from the variab...
other debtors	Is there another debtor or a guarantor for the credit? (categorical)
present residence	length of time (in years) the debtor lives in the present residence (ordinal; discretized quantitati...
property	the debtor's most valuable property, i.e. the highest possible code is used. Code 2 is used, if code...
age	age in years (quantitative)
other installment plans	installment plans from providers other than the credit-giving bank (categorical)
housing	type of housing the debtor lives in (categorical)
number credits	number of credits including the current one the debtor has (or had) at this bank (ordinal, discretiz...
job	quality of debtor's job (ordinal)
people liable	number of persons who financially depend on the debtor (i.e., are entitled to maintenance) (binary,d...
telephone	Is there a telephone landline registered on the debtor's name? (binary; remember that the data are f...
foreign worker	Is the debtor a foreign worker? (binary)
credit risk	Has the credit contract been complied with (good) or not (bad) ? (binary)

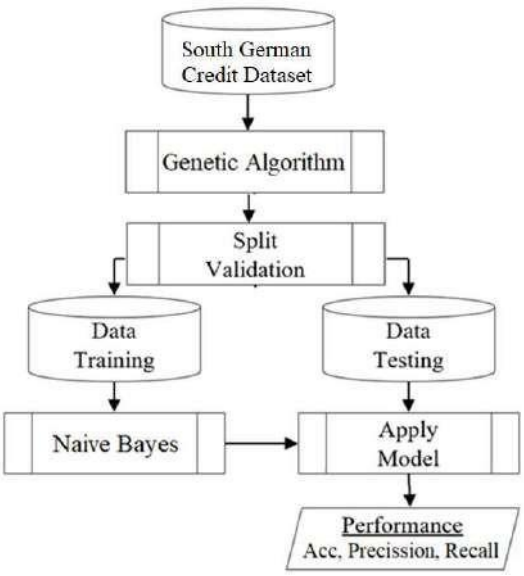
Data *South German Credit* dipilih karena telah bebas dari *missing value*, sehingga tidak perlu lagi prepocessing data untuk digunakan dalam pembuatan model klasifikasi.

B. Model Penelitian

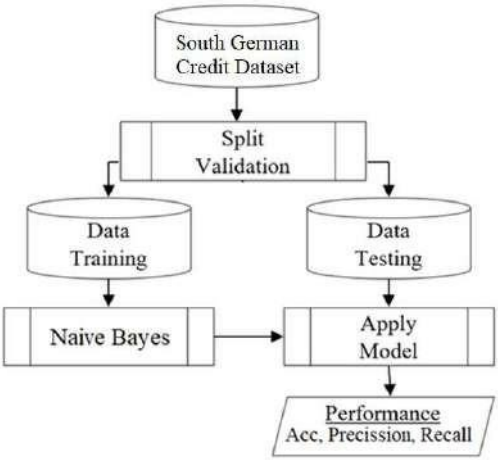
Data *South German Credit* digunakan untuk membentuk model klasifikasi. Label yang digunakan adalah pada atribut “*Credit Risk*” dengan nilai “Good” dan “Bad”. Sebanyak 77% *instances*

yang ada pada data *South German Credit* merupakan *instance* dengan kelas label “good”, sedangkan sisanya adalah *instances* dengan label “bad”. Penelitian ini melakukan pengujian sebanyak 2 kali yang nantinya akan dianalisa hasil yang diperoleh. Pengujian pertama dilakukan menggunakan optimasi GA, sedangkan pengujian yang kedua dilakukan tanpa optimasi GA.

Model klasifikasi yang dibangun pada penelitian ini menggunakan proses *spit validation* untuk membagi data kedalam data *training* dan data *testing*. Data training yang digunakan pada penelitian ini adalah sebanyak 95% dari seluruh data *South German Credit*, sedangkan untuk data *testing* menggunakan 5% sisanya. Data *training* yang diperoleh dari proses validasi akan digunakan untuk pemodelan klasifikasi dengan dengan algoritma Naïve Bayes. Model yang dihasilkan kemudian dijadikan sebagai apply model untuk digunakan pada data *testing.* Setelah klasifikasi telah dilakukan, kemudian diukur kinerja dari model klasifikasi yang dibentuk berdasarkan nilai akurasi, presisi, dan *recall*.



Gambar 1. Pengujian Pertama Klasifikasi Naïve Bayes Menggunakan Genetic Algorithm



Gambar 2. Pengujian Kedua Klasifikasi Naïve Bayes Tanpa Menggunakan Genetic Algorithm

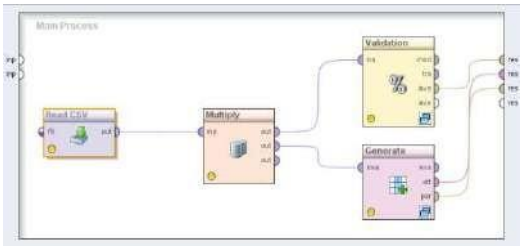
Pada Gambar 1 dan Gambar 2 menunjukkan bahwa pada penelitian ini pengujian dilakukan sebanyak 2 kali, yaitu: (1) Klasifikasi data *South German Credit* menggunakan Naïve Bayes dengan optimasi Genetic Algorithm, (2) Klasifikasi data *South German Credit* menggunakan Naïve Bayes tanpa optimasi Genetic Algorithm. Hasil performa dari kedua pengujian tersebuat akan dibandingkan untuk

kemudian dianalisa untuk menunjukkan temuan penelitian.

III. Hasil dan Pembahasan

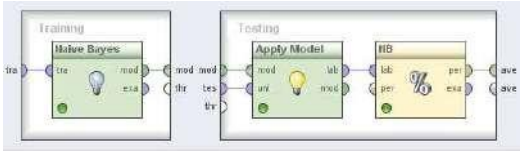
A. Langkah Pengujian

Tools RapidMiner versi 5.0 digunakan pada penelitian ini untuk melakukan pengujian. RapidMiner dapat digunakan untuk penelitian, pembuatan prototipe yang cepat, serta mendukung semua langkah proses penambangan data seperti persiapan data, visualisasi hasil, validasi, dan optimasi [17], sehingga dianggap cocok digunakan pada penelitian ini. Tahap pertama dalam pembuatan model penelitian adalah dengan memanggil data *South German Credit* tools RapidMiner, kemudian dilakukan fungsi multiply untuk melakukan dua pengujian sekaligus, yaitu pengujian dengan menggunakan GA dan pengujian tanpa menggunakan GA. Proses validasi data dilakukan menggunakan split validation untuk membagi data kedalam 95% data training dan 5% data testing. Secara lebih jelasnya tentang pemanggilandata dan proses validasi dapat dilihat pada Gambar 3.



Gambar 3. Pemanggilan Data dan Proses Validasi

Pada setiap proses validasi yang ditampilkan pada Gambar 3, didalamnya berisi proses pembelajaran dengan algoritma Naïve Bayes yang kemudian diterapkan pada apply model untuk di ukur performa akurasi, presisi dan recall. Adapun proses pembelajaran yang ada pada penelitian ini dapat dilihat pada Gambar 4.



Gambar 4. Proses Pembelajaran Naïve Bayes dan Apply Model

Langkah selanjutnya setelah seluruh model penelitian telah terbentuk adalah menjalankan model yang telah dibangun pada RapidMiner, kemudian akan diperoleh hasil akurasi, presisi dan recall untuk di analisa hasil.

B. Hasil Pengujian

Setelah dilakukan 2 kali pengujian, diperoleh nilai akurasi, presisi, dan recall dari kedua model. Secara lebih lengkap hasil pengujian dapat dilihat pada Tabel 2.

Tabel 2. Hasil Pengujian

No	Algoritma	Akurasi	Presisi	Recall
1	Naïve Bayes	84,53%	88,47%	84,90%
2	GA + Naïve Bayes	85,99%	87,43%	87,91%

Berdasarkan Tabel 2 dapat diketahui bahwa GA ternyata mampu meningkatkan akurasi dan recall dari Naïve Bayes, akan tetapi GA belum mampu meningkatkan nilai presisi dari Naïve Bayes. Hasil pengujian menunjukkan bahwa dengan akurasi 85,99%, optimasi GA memberikan Naïve Bayes peningkatan nilai akurasi sebesar 1,46% dan peningkatan nilai recall sebasar 3,01% untuk klasifikasi data *South German Credit.* Namun demikian, ternyata penggunaan GAjuga menurunkan nilai presisi sebesar 1,04% dari penggunaan Naïve Bayes

untuk klasifikasi dat *South German Credit.* Hal tersebut diduga karena dari 22 atribut yang ada pada data *South German Credit*, ternyata hanya ada 4 atribut saja yang diberikan pembobotan oleh GA. Hasil pembobotan ini juga menjelaskan kenapa peningkatan akurasi dan recall yang diberikan oleh GA tidak terlalu besar. Hasil pembobotan atribut dari GA dapatdilihat pada Tabel 3.

Tabel 3. Hasil Pembobotan GA pada Data *South German Credit*

Atribut	Pembobotan
status	0
duration	0
credit history	1
purpose	0
amount	0
savings	0
employment duration	0
installment rate	0
personal status sex	0
other debtors	0
present residence	0
property	1
age	1
other installment plans	0
housing	0
number credits	0
job	1
people liable	0
telephone	0
foreign worker	0

Pada Tabel 3 dapat diketahui bahwa hanya terdapat 4 atribut yang diberikan pembobotan oleh GA. Hal ini menjukkan bahwa, berdasarkan GA ke-4 atribut inilah yang paling penting untuk diperhatikan Ketika hendak melakukan klasifikasi data *South German Credit.* Adapun atribut-atribut tersebut yaitu: credit history, property, age, dan job.

C. Pembahasan Hasil

Berdasarkan hasil pengujian yang telah dilakukan, klasifikasi data *South German Credit* telah diketahui bahwa penggunaan optimasi GA mampu meningkatkan performa akurasi dan recall dari algoritma Naïve Bayes meskipun tidak terlalu besar. Kecilnya peningkatan performa yang diberikan diduga karena atribut yang diberikan pembobotanoleh GA jumlahnya tidak sampai 25% dari seluruh atribut yang ada pada data *South German Credit.* Hal tersebut menjadikan proses perhitungan probabilitas pada Naïve bayes menjadi kurang berpengaruh. Bahkan melihat dari sisi presisi ternyata penggunaan GA malah menjadikan performa Naïve Bayes menjadi menurun.

Meskipun optimasi dari GA tidak memberikan hasil yang maksimal, dengan menggunakan GA ternyata dapat diperoleh atribut mana saja yang dapat dijadikan sebagai prioritas evaluasi untuk melihat kepuasan dari pelanggan maskapai penerbangan. Dengan melihat atribut yang diberikan pembobotan oleh GA, dapat dijadikan sebagai acuan untuk mempertimbangkan atribut tersebut sebagai fokus utama untuk peningkatan pelayanan. Adapun atribut yang diberikan pembobotan oleh GA antara lain: Class, Inflight wifi service, On-board service dan Checkin service. Temuan ini diharapkan dapat memberikan sumbangan praktis terhapat pelayanan kedepan yang akan diberikan oleh maskapai penerbangan terhadap pelanggan mereka.

menunjukkan temuan penelitian.

IV. Kesimpulan

Penelitian ini telah menguji penggunaan algoritma Naïve Bayes untuk mengklasifikasikan data *South German Credit* serta membandingkannya dengan klasifikasi Naïve Bayes

menggunakan optimasi GA. Berdasarkan pengujian yang telah dilakukan menunjukkan beberapa hasil yaitu:

1. Akurasi dan recall paling tinggi dari klasifikasi data *South German Credit* adalah menggunakan algoritma Naïve Bayes dengan optimasi GA. Akurasi maksimal yang diperoleh yaitu sebesar 85,99% dan recall maksimal adalah sebesar 87,91%.
2. Nilai presisi yang paling maksimal dari klasifikasi data *South German Credit* adalah dengan menggunakan algoritma Naïve Bayes tanpa optimasi GA dengan nilai presisi sebesar 88,47%.
3. Algoritma GA belum mampu memberikan penambahan performa secara maksimal pada algoritma Naïve Bayes untuk mengklasifikasikan data *South German Credit.*
4. Atribut Class, Inflight wifi service, On-board service dan Checkin service adalah atribut yang perlu dipertimbangkan oleh pihak maskapai penerbangan untuk memaksimalkan kepuasan dari para pelanggan.

Hasil dari penelitian ini masih belum mampu memberikan performa yang cukup baik untuk klasifikasi data *South German Credit*, karena baik akurasi, presisi ataupun recall tidak ada yang memperoleh nilai lebih dari 90%. Hal ini membutuhkan penelitian lebih lanjut untuk memperoleh model klasifikasi data *South German Credit* yang lebih baik kedepannya. Berdasarkan temuan penelitian ini menyarankan agar pada penelitian dimasa mendatang dapat menerapkan metode optimasi yang lain untuk lebih mengoptimalkan performa dari algoritma Naïve Bayes, misalkan algoritma Particle swarm optimization (PSO) atau boostraping.

Daftar Pustaka

[1] A. T. Rahmawati, M. Saifi and R. R. Hidayat, "Analisis Keputusan Pemberian Kredit dalam Langkah Meminimalisir Kredit Bermasalah," *Jurnal Administrasi Bisnis,* vol. 35, no. 1, pp. 179-186, 2016.

[2] M. S. Garver, "Using Data Mining for Customer Satisfaction Research," *Marketing Research*, vol. 14, no. 1, pp. 8-17, 2002.

[3] "Kaggle," kaggle.com, 2020. [Online]. Available: https://www.kaggle.com/c/south-german-credit-prediction/overview/data-overview. [Accessed 2 November 2020].

[4] I. A. A. Amra and A. Y. A. Maghari, "Students Performance Prediction Using KNN and Naïve Bayesian," in *8th International Conference on Information Technology (ICIT)*, Al-Zaytoonah University of Jordan, Jordan, 2017.

[5] F. Osisanwo, J. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi and J. Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128-138, 2017.

[6] E. N. Azizah, U. Pujianto, E. Nugraha and Darusalam, "Comparative Performance Between C4.5 and Naive Bayes Classifiers in Predicting Student

Academic Performance in A Virtual Learning Environment," in *4th International Conference on Education and Technology (ICET)*, Malang, Indonesia, 2018.

[7] K. Madasamy and M. Ramaswami, "Data Imbalance and Classifiers: Impact and Solutions from A Big Data Perspective," *International Journal of Computational Intelligence Research*, vol. 13, no. 9, pp. 2267-2281, 2017.

[8] E. M. Hassib, A. I. El-Desouky, E.-S. M. El-Kenawy and S. M. El-Ghamrawy, "An Imbalanced Big Data Mining Framework for Improving Optimization Algorithms Performance," *Journal & Magazines*, vol. 7, no. 1, pp. 170774-170795, 2019.

[9] S. Chen, G. I. Webb, L. Liu and X. Ma, "A Novel Selective Naïve Bayes Algorithm," *Knowledge-Based Systems*, vol. 192, pp. 1-15, 2020.

[10] L. Jiang, L. Zhang, L. Yu and D. Wang, "Class-Specific Attribute Weighted Naive Bayes," *Pattern Recognition*, vol. 88, no. 1, pp. 321-330, 2019.

[11] S. Ernawati, R. Wati, N. Nuris, L. S. Marita and E. R. Yulia, "Comparison of Naïve Bayes Algorithm with Genetic Algorithm and Particle Swarm Optimization as Feature Selection for Sentiment Analysis Review of Digital Learning Application," *Journal of Physics: Conference Series*, vol. 1641, pp. 1-7, 2020.

[12] S. Ernawati, E. R. Yulia, Frieyadie and Samudi, "Implementation of The Naïve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies," in *The 6th International Conference on Cyber and IT Service Management (CITSM 2018)*, Medan, Indonesia, 2018.

[13] A. Arwan and D. S. Rusdianto, "Optimization of Genetic Algorithm Performance Using Naïve Bayes for Basis Path Generation," *Kinetik*, vol. 2, no. 4, pp. 273-282, 2017.

[14] E. Stripling, S. v. Broucke, K. Antonio, B. Baesens and M. Snoecka, "Profit Maximizing Logistic Model for CustomerChurn Prediction Using Genetic Algorithms," *Swarm and Evolutionary Computation*, vol. 40, no. 1, pp. 116-130, 2018.

[15] D. K. Choubey, S. Paul, S. Kumar and S. Kumar, "Classification of Pima Indian Diabetes Dataset Using Naive Bayes With Genetic Algorithm As An Attribute Selection," in *The International Conference on Communication and Computing Systems (ICCCS)*, Ranchi, India, 2016.

[16] L. G. P. Suardani, I. M. A. Bhaskara and M. Sudarma, "Optimization of Feature Selection Using Genetic Algorithm with Naïve Bayes Classification for Home Improvement Recipients," *International Journal of Engineering and Emerging Technology*, vol. 3, no. 1, pp. 66-70, 2018.

[17] A. Jeyaraj, R. S and M. R. Raja, "A study of classification algorithms using Rapidminer," *International Journal of Pure and Applied Mathematics,* vol. 119, no. 12, pp. 15977-15988, 2018.