Implementasi Teknik Bagging untuk Peningkatan Kinerja J48 dan Logistic Regression dalam Prediksi Minat Pembelian Online

Eka Rahmawati
Program Studi Sistem Informasi
Universitas Bina Sarana Informatika
Jakarta, Indonesia
eka.eat@bsi.ac.id

Candra Agustina
Program Studi Sistem Informasi Akuntansi
Universitas Bina Sarana Informatika
Jakarta, Indonesia
candra.caa@bsi.ac.id

Abstract*—The rapid growth of online shopping sites makes business in the virtual world very promising. Purchasing intentions is one of the keys to success in an online store. There are several data mining methods for making predictions on online purchase intentions datasets. Data can represent the characteristics or habits of each user who has visited a site whether it ends with a transaction or not. Some popular algorithms with good performance in data mining include J48 and Logistic Regression. However, in data sometimes there is a problem of class imbalance, so the ensemble technique needs to be applied. One technique that can be applied is bagging. This research examines data using bagging techniques to improve the performance of the J48 algorithm and Logistic Regression. The results of improving the performance of data mining algorithms with these techniques have an accuracy value of 89.68% for the J48 algorithm and 88.50% for the Logistic Regression algorithm. This figure shows an increase when compared with initial testing without using ensemble techniques. Increases were also experienced in Recall, F-Measure, and AUC values.

Keywords—purchasing intentions; J48; Logistic Regression; Bagging;

***Abstrak*— Pesatnya situs pembelanjaan online menjadikan bisnis di dunia virtual sangat menjanjikan. Minat pembelian menjadi salah satu kunci kesuksesan pada sebuah toko online. Terdapat beberapa metode data mining untuk melakukan prediksi pada dataset minat pembelian online. Data dapat mewakili karakteristik atau kebiasaan dari setiap user yang telah mengunjungi suatu situs baik berakhir dengan melakukan transaksi ataupun tidak. Beberapa algoritma yang populer dengan kinerja yang baik dalam data mining diantaranya J48 dan Logistic Regreession. Namun, dalam sebuah data terkadang terdapat masalah ketidakseimbangan kelas, sehingga perlu diterapkan teknik ensemble. Salah satu teknik yang dapat diterapkan adalah teknik bagging. Penelitian kali ini mengujikan data dengan teknik bagging untuk meningkatkan kinerja algoritma J48 dan Logistic Regression. Hasil dari peningkatan kinerja algoritma data mining dengan teknik tersebut memiliki nilai akurasi 89.68% untuk algoritma J48 dan 88.50% untuk algoritma Logistic Regression. Angka tersebut menunjukan adanya peningkatan jika dibandingkan dengan pengujian awal tanpa menggunakan teknik ensemble. Peningkatan juga dialami pada nilai Recall, F-Measure, dan AUC.**

Keywords—Minat Pembelian, J48, Logistic Regression, Bagging

PENDAHULUAN

Peningkatan penyedia situs belanja online dengan berbagai fitur yang menguntungkan membuat konsumen beralih dari konsep pasar konvensional. Jika dahulu pembelian hanya dapat dilakukan dengan tatap muka secara *real-time*, maka saat ini sebuah gambar dapat mewakili deskripsi produk yang dibutuhkan. Pembelian dengan menggunakan situs online juga memiliki tantangan baik untuk pembeli maupun penjual. Untuk mendapatkan pelanggan dengan menggunakan sistem online, diperlukan berbagai strategi seperti user interface yang menarik. Selain itu, kepastian pembelian terhadap user yang telah mengunjungi website juga diperlukan. Tidak jarang user sudah memasukan produk ke keranjang, namun tidak melakukan checkout dan batal melakukan transaksi. Kepastian pembelian pelanggan perlu diperhatikan agar dapat dilakukan evaluasi terhadap layanan, tampilan ataupun faktor lain yang memberikan pengaruh terhadap intensitas pembelian. Proses evaluasi dapat dilakukan dengan menggunakan metode data minig. Efektifitas dari algoritma dapat diterapkan pada data untuk menentukan tingkat akurasinya. Penelitian kali ini akan menguji kinerja dari algoritma J48 dan Logistic Regression pada prediksi minat pembelian online.

Penelitian pada data customer churn[1] telah menggunakan algoritma decision tree dan logistic regression karena kedua algoritma tersebut memiliki performance yang bagus. Penelitian [2] telah melakukan pengujian terhadap data intensitas minat beli pada sebuah data toko online dengan algoritma decision tree C45 dan Random Forest dengan nilai akurasi 82.34 % dan 82.29 %. Algoritma decision tree juga memiliki nilai akurasi yang baik pada penelitian loyalitas pelanggan[3]. Penggunaan decision tree pada penelitian peningkatan kredit bank[4] juga telah dilakukan dengan tingkat akurasi mencapai 82%. Penelitan dengan logistic regression pada dataset resiko kartu kredit[5] menunjukkan bahwa nilai akurasinya mencapai 74% dan 70% pada algoritma Decision Tree. Kemudian peningkatan kinerja dilakukan dengan menggunakan teknik bagging. Implementasi teknik logistic regression juga dilakukan pada penelitian loyalitas pelanggan[6] dengan nilai AUC 0,7320. Kemudian penelitian tersebut juga menerapkan teknik ensemble untuk meningkatkan kinerja algoritma. Pentingnya teknik ensemble untuk mengatasi ketidakseimbangan kelas juga dilakukan pada penelitian[7] dalam klasifikasi harga

DOI: https://doi.org/10/25047/jtit.v7i1.123

listrik yang menunjukan adanya peningkatan setelah diterapkannya teknik ensemble.

Penelitian menggunakan data minat pembelian user yang diperoleh dari salah satu platform toko online dengan mengidentifikasi setiap pembeli. Dari data tersebut dapat dilakukan klasifikasi menggunakan teknik data mining. Untuk menangani ketidakseimbangan kelas, maka diperlukan penerapan teknik tertentu pada klasifikasi data mining. Salah satu teknik yang dapat digunakan untuk mengatasi imbalance class adalah teknik bagging. Penggunaan teknik bagging untuk mengatasi *imbalanced class* telah dilakukan pada penelitian[5] untuk memprediksi resiko dari kartu kredit. Selain itu, teknik *bagging* juga dilakukan pada deteksi penipuan kartu kredit[8]. Penelitian juga menunjukan bahwa tingkat akurasi meningkat ketika teknik bagging yang diterapkan dengan algoritma C45. Penggunaan teknik bagging untuk menangani *imbalanced class* juga telah dilakukan pada klasifikasi harga listrik[7]. Penggunaan teknik *bagging* untuk menangani ketidak seimbangan kelas juga dilakukan pada penelitian[9] pada prediksi cacat software.

TINJAUAN PUSTAKA

A. *Bagging*

Ketidakseimbangan kelas sering terjadi pada dataset sehingga proses klasifikasi tidak dapat dilakukan dengan masimal. Bagging menjadi salah satu teknik yang dapat digunakan untuk mengatasi ketidakseimbangan kelas. Teknik ensemble bekerja untuk mengatasi komponen classifier dengan menentukan bobot terkecil[7]. Masalah imbalance class seringkali mempengaruhi kinerja dari algoritma. Kombinasi dari algoritma dengan teknik ensemble juga dilakukan untuk mendapatkan performa yang tinggi. Terdapat beberapa teknik ensemble yang dapat diterapkan pada klasifikasi suatu data. Teknik bagging juga menjadi salah satu cara yang tepat untuk menangani kasus ketidakseimbangan kelas[10]. Penggunaan teknik ini telah dilakukan pada penelitian sebelumnya dan berhasil meningkatkan kinerja algoritma.

Bagging merupakan sebuah mesin ensemble aggregate bootstrap yang pada awalnya dikembangkan oleh Breiman untuk diterapkan pada peningkatan akurasi klasifikasi[11]. Penerapan teknik bagging dapat dilakukan dengan mengkombinasikan set pelatihan untuk kemudian diset secara kasual. Keefektifan bagging untuk meningkatkan akurasi dari suatu algoritma klasifikasi telah dibuktikan oleh beberapa penelitian. Kelebihan dari penggunaan teknik bagging adalah dapat mengurangi varians dari algoritma dengan penyesuaian antara estimasi dan hasil yang diinginkan dari peningkatan akurasi suatu model[12]. Prediksi dari out-of-bag menunjukan H(X) pada vektor X. Learner yang tidak dilatih X yang akan terlibat pada prosesnya[13]. Adapun rumusnya adalah sebagai berikut:

$$\mathrm{H}(X)^{\mathrm{OOB}} = \underset{y \in Y}{argmax} \sum_{t=1}^{T} \mathbb{I}\left(h_t(X) = y\right).\mathbb{I}(X \notin N_t) \quad (1)$$

dimana *X* adalah vector, *x* adalah variabel, *y* adalah output spaces, *N* adalah data sample, *T* adalah jumlah learner (*t=1,....,T), H* adalah learner dan $\mathbb{I}$ (.) adalah indikator dari fungsi yang mengambil 1 jika. true dan o or false.

B. *Logistic Regression*

Logistic Regression merupakan algoritma klasifikasi yang digunakan pada data mining. Algoirtma ini dikenal memiliki performance tinggi[14]. Algoritma logistic regressin menjadi populer implementasinya dalam klasifikasi data mining karena performance yang dimiliki sangat baik. Logsitic regresi menjadi salah satu analisis yang menggunakan konsep multivariate untuk mendeteksi dependen variabel yang dilakukan berdasarkan variabel independen.

Logistic Regression mempunyai keterbatasan, dalam menangani variabel dependen yang bersifat dikotomi ataupun kategori. Dalam kenyataannya, pada penelitian terdapat banyak variabel dikotomi yang menarik untuk diteliti. Regresi Logistik menjadi algoritma dengan pendekatan model yang dapat digunakan untuk menggambarkan relasi beberapa variabel independen X terhadap variabel dependen D yang bernilai biner[15]. Penelitian ini menggunakan Logistic Regression yang akan diimplementasikan dengan teknik bagging sebagai cara untuk meningkatkan kinerja algoritma.

C. *Decision Tree*

Decision Tree merupakan proses klasifikasi pohon keputusan dari kelas yang diberi label tuple pelatihan. Pohon keputusan berbentuk bagan alur sederhana seperti struktur pohon, di mana simpul paling atas dalam pohon adalah simpul akar[16]. Decision tree menjadi algoritma yang biasa digunakan untuk memprediksi model, dan juga untuk mengetahui informasi berharga melalui sejumlah besar klasifikasi data. Algoritma klasifikasi ini yang telah banyak diterapkan dalam kesalahan diagnosa. Ini mengadopsi regulasi rekursif top-down dan nilai-nilai atribut dibandingkan di node internal pohon keputusan.

Prinsip klasifikasi decision tree lebih sederhana dan mudah dipahami. Pada saat yang sama, model klasifikasi berdasarkan decision tree menghitung dengan cepat[17]. Untuk meningkatkan kelengkapan dan kegunaan decision tree dalam proses pembentukanya, metode ensemble dapat diterapkan untuk meningkatkan nilai akurasi yang dicapai. Penerapan decison tree untuk berbagai penelitian juga menghasilkan nilai akurasi yang sangat baik. Penggunaannya untu klasifikasi minat pembelian pada toko online digunakan agar dapat membantu proses pengambilan keputusan.

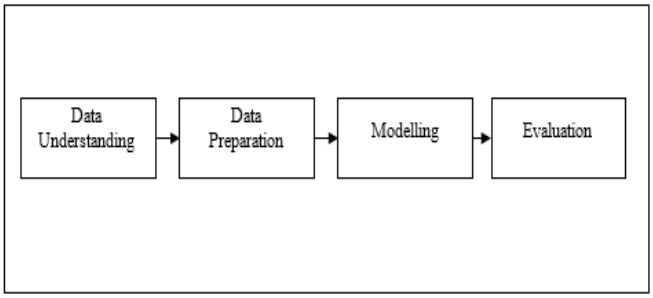
J48 menjadi salah satu pengembangan dari algoritma decision tree yang memiliki kinerja yang baik. Algritma ini membuat pohon keputusan yang tergantung pada nilai-nilai atribut dari data pelatihan yang tersedia terhadap klasifikasi item baru[18]. Algoritima akan menganalisis semua item pelatihan dengan mengenali berbagai atribut untuk membedakan beragam contoh dengan lebih jelas. Fitur yang terdapat pada algoritma J48 ini membedakan semua instance yang digunakan untuk mengklasifikasikannya, yang terbaik seharusnya berisi pencapaian informasi tertinggi. Pohon keputusan yang dihasilkan oleh J48 dapat digunakan untuk klasifikasi. Di setiap simpul pohon, J48 memilih atribut data yang paling efektif membagi pengaturan tes menjadi himpunan bagian yang ditingkatkan dalam satu kelas atau yang lain. Kriteria pemisahan adalah perolehan informasi terstandarisasi (berbeda dengan entropi). Atribut dengan perolehan informasi terstandarisasi tertinggi yang layak dibuat berdasarkan keputusan[19]. Penerapan algoritma J48 dalam data mining juga harus ditingkatkan akurasinya. Ketika terdapat ketidakseimbangan kelas, maka teknik

DOI: https://doi.org/10.25047/jtit.v7i1.123

ensemble menjadi salah satu cara yang dapat diimplementasikan.

METODE PENELITIAN

Penelitian ini melakukan pendekatan terhadap data untuk membandingkan kinerja yang lebih baik antara decision tree dan logistic regression. Dua algoritma tersebut menjadi algoritma yang memiliki performa yang baik jika dilihat dari penelitan sebelumnya. Data yang digunakan adalah data pengelompokan minat beli dari pengguna toko online. Terdapat 12330 data yang berhasil dikumpulkan. Data tersebut memiliki 17 atribut untuk menentukan kepastian pembelian.



Gambar 1 Metode Penelitian

A. Data Understanding

Data yang digunakan adalah data yang diperoleh dari dataset publik dari UCI repository. Dari 12.330 sesi dalam dataset, 84,5% (10.422) adalah sampel kelas negatif yang tidak berakhir dengan belanja, dan sisanya (1908) adalah sampel kelas positif yang berakhir dengan belanja.

B. Data Preparation

Persiapan data dilakukan dengan menerapkan teknik ensemble untuk mengatasi ketidakseimbangan kelas. Teknik ini juga digunakan untuk mengoptimalkan kinerja dari algoritma klasifikasi. Teknik yang digunakan dalam data prepartaion adalah teknik bagging.

C. Modelling

Selanjutnya tahapan modelling dilakukan dengan mengimplementasikan algoritma decision tree dan logistic regression.

D. Evaluation

Agar dapat dilakukan evaluasi dari pengujian algoritma, maka nilai akurasi, AUC, Recall dan F-measure. Nilai akan diperbandingkan untuk mengetahui efektifitas penggunaan algoritma.

HASIL DAN PEMBAHASAN

Penelitian [2] telah melakukan prediksi secara realtime pada dataset *online shoppers intention* dengan menggunakan algoritma C45 dan *Random Forest*. Hasil penelitian menunjukan nilai akurasi untuk algoritma C45 82,34% dan 82,29% untuk algoritma *Random Forest*. Peningkatan nilai akurasi masih dapat dilakukan dengan menggunakan algoritma *decision tree* J48 dan *Logistic Regression* serta penambahan teknik bagging untuk mengatasi *imbalanced class.*

Pada penelitian ini, data diujikan dengan menggunakan algoritma J48 dan logistic regression. Pengujian awal dilakukan untuk mengetahui hasil dari kinerja algoritma sebelum menerapkan metode untuk peningkatan kinerja. Pengujian yang dilakukan memperoleh hasil seperti yang terdapat pada Tabel 1.

TABLE I. HASIL PENGUJIAN AWAL

Algoritma	*Accuracy*	*Recall*	*F-Measure*	*AUC*
J48	89.53%	0.895	0.892	0.784
LR	88.40%	0.884	0.868	0.894

Hasil pengujian awal menunjukan bhawa J48 memiliki nilai akurasi yang lebih tinggi 1.13% jika dibandingkan dengan algoritma logistic regression. Nilai Recall dari klasifikasi J48 juga lebih tinggi 0.011 dari logistic regression. F-measure yang dihasilkan dari algoritma J48 juga lebih baik 0.024 jika dibandingkan dengan logistic regression. Untuk nilai AUC, logistic regression lebih tinggi 0.011 dari J48. Secara keseluruhan hingga tahap ini semua algoritma dapat menguji data secara efektif melihat dari nilai AUC yang melebih 0.6 sehingga termasuk dalam kategori good classification.

Selanjutnya, upaya peningkatan kinerja algoritma dilakukan dengan menggunakan teknik bagging. Implementasi teknik bagging untuk kedua algoritma tersebut kemudian menghasilkan data klasifikasi seperti yang terdapat pada Table 2.

TABLE II. HASIL PENGUJIAN DENGAN BAGGING

Algoritma	*Accuracy*	*Recall*	*F-Measure*	*AUC*
J48	89.68%	0.897	0.893	0.924
LR	88.50%	0.885	0.869	0.924

Hasil pengujian dengan teknik bagging menunjukan bahawa J48 memiliki nilai akurasi yang lebih tinggi 1.18% jika dibandingkan dengan algoritma logistic regression. Nilai Recall dari klasifikasi J48 juga lebih tinggi 0.012 dari logistic regression. F-measure yang dihasilkan dari algoritma J48 juga lebih baik 0.024 jika dibandingkan dengan logistic regression. Jumlah ini masih sama seperti pengujian awal. Untuk nilai AUC, logistic regression dan J48 memiliki nilai yang sama.

Selanjutnya perbandingan pengujian dengan dan tanpa bagging dapat dilihat pada Table 3.

TABLE III. PERBANDINGAN HASIL PENGUJIAN

Algoritma	*Accuracy*	*Recall*	*F-Measure*	*AUC*
J48	89.53%	0.895	0.892	0.784
LR	88.40%	0.884	0.868	0.894
J48+Bagging	**89.68%**	**0.897**	**0.893**	**0.924**
LR+Bagging	88.50%	0.885	0.869	0.924

Pengujian dengan J48 + Bagging memiliki tingkat akurasi yang lebih baik jika dibandingkan dengan Logistic Regression. Nilai Recall, F-Measure, dan AUC untuk algoritma J48 juga lebih baik jika dibandingkan dengan Logistic Regression. Dalam hal ini penggunaan teknik

DOI: https://doi.org/10/25047/jtit.v7i1.123

bagging untuk algoritma Logistic Regression tidak ada peningkatan yang signifikan.

KESIMPULAN

Pengujian yang dilakukan untuk meningkatkan kinerja algoritma dengan teknik ensemble bagging terbukti dapat meningkatkan kinerja. Peningkatan terjadi untuk kedua algoritma J48 dan Logistic Regression. Kedepannya akan lebih baik jika digunakan perbandingan untuk algoritma-algoritma baru apakah teknik bagging masih meningkatkan performa dengan baik atau tidak.

REFERENSI

[1] A. De Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," *Eur. J. Oper. Res.*, vol. 269, no. 2, pp. 760–772, 2018.

[2] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks," *Neural Comput. Appl.*, vol. 31, no. 10, pp. 6893–6908, 2019.

[3] M. Cendana, S. Dian, and H. Permana, "Analisis Perbandingan Algoritma Naive Bayes , J48 , Dan Random Forest Tree Dalam Peningkatan Loyalitas Pelanggan Umkm Dengan Voucher Belanja," vol. 11, no. 2, pp. 140–145, 2019.

[4] O. J. Okesola, K. O. Okokpujie, A. A. Adewale, S. N. John, and O. Omoruyi, "An Improved Bank Credit Scoring Model: A Naïve Bayesian Approach," *Proc. - 2017 Int. Conf. Comput. Sci. Comput. Intell. CSCI 2017*, pp. 228–233, 2018.

[5] M. A. Muslim, A. Nurzahputra, and B. Prasetiyo, "Improving accuracy of C4.5 algorithm using split feature reduction model and bagging ensemble for credit card risk prediction," *2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018*, vol. 2018-Janua, no. 1996, pp. 141–145, 2018.

[6] U. Ahmed, A. Khan, S. H. Khan, A. Basit, I. U. Haq, and Y. S. Lee, "Transfer Learning and Meta Classification Based Deep Churn Prediction System for Telecom Industry," pp. 1–10.

[7] W. W. Y. Ng, J. Zhang, C. S. Lai, W. Pedrycz, L. L. Lai, and X. Wang, "Cost-sensitive weighting and imbalance-reversed bagging for streaming imbalanced and concept drifting in electricity pricing classification," *IEEE Trans. Ind. Informatics*, vol. 15, no. 3, pp. 1588–1597, 2019.

[8] S. Akila and U. S. Reddy, "Credit Card Fraud Detection Using Non-Overlapped Risk Based Bagging Ensemble (NRBE)," *2017 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2017*, no. November, pp. 1–4, 2018.

[9] A. Saifudin, F. Teknik, U. Pamulang, R. S. Wahono, F. I. Komputer, and U. D. Nuswantoro, "Penerapan Teknik Ensemble untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software," *J. Softw. Eng.*, vol. 1, no. 1, pp. 28–37, 2015.

[10] S. Aries and R. S. Wahono, "Pendekatan Level Data untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software," *J. Softw. Eng.*, vol. 1, no. 2, pp. 76–85, 2015.

[11] J. Dou *et al.*, "Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan," *Landslides*, no. October, 2019.

[12] D. Tien Bui, T. C. Ho, B. Pradhan, B. T. Pham, V. H. Nhu, and I. Revhaug, "GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks," *Environ. Earth Sci.*, vol. 75, no. 14, 2016.

[13] F. Schwenker, "Ensemble Methods: Foundations and Algorithms [Book Review]," *IEEE Comput. Intell. Mag.*, vol. 8, no. 1, pp. 77–79, 2013.

[14] H. Rianto and R. S. Wahono, "Resampling Logistic Regression untuk Penanganan Ketidakseimbangan Class pada Prediksi Cacat Software," *J. Softw. Eng.*, vol. 1, no. 1, pp. 46–53, 2015.

[15] G. Santosa, U. Kristen, D. Wacana, A. Rachmat, U. Kristen, and D. Wacana, "Perbandingan Akurasi Model Regresi Logistik umtuk Prediksi Kategori IP Mahasiswa Jalur Prestasi dengan Non Jalur Prestasi," *J. Tek. dan Ilmu Komput.*, no. May, 2018.

[16] P. Chandrasekar, K. Qian, H. Shahriar, and P. Bhattacharya, "Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing," *2017 IEEE 41st Annu. Comput. Softw. Appl. Conf.*, pp. 481–484, 2017.

[17] X. Zhang, D. Jiang, Q. Long, and T. Han, "Rotating machinery fault diagnosis for imbalanced data based on decision tree and fast clustering algorithm," *J. Vibroengineering*, vol. 19, no. 6, pp. 4247–4260, 2017.

[18] H. Kaur and A. Sharma, "Novel Email Spam Classification using Integrated Particle Swarm Optimization and J48," *Int. J. Comput. Appl.*, vol. 149, no. 7, pp. 23–27, 2016.

[19] N. S. anaN and V. G. thri, "Performance and Classification Evaluation of J48 Algorithm and Kendall's Based J48 Algorithm (KNJ48)," *Int. J. Comput. Trends Technol.*, vol. 59, no. 2, pp. 73–80, 2018.

DOI: https://doi.org/10/25047/jtit.v7i1.123