Jurnal TRANSFORMATIKA
Vol. 22, No. 2, Januari 2025, pp. 81 - 89
P-ISSN: 1693-3656, E-ISSN: 2460-6731
https://journals.usm.ac.id/index.php/transformatika/

Boosting Performance Classification KNN Customer Loyalty with Chi-Square and Information Gain

Atika Mutiarachim[1*], Fari Katul Fikriah[2], Basirudin Ansor[3], Aditya Putra Ramdani[4]
[1]Universitas 17 Agustus 1945 Semarang
Jl. Pawiyatan Luhur I, Semarang, (024) 8441771, e-mail: atikamutiarachim@untagsmg.ac.id
[2]Universitas Widya Husada Semarang
Jl. Subali Raya No 12, Semarang, (024) 7612988, e-mail: farichatulfikriyah45@gmail.com
[3]Universitas Muhammadiyah Semarang
Jl Kedungmundu No 18, Semarang (024) 76740296, e-mail: basirudinansor@unimus.ac.id
[4]Universitas Muhammadiyah Semarang
Jl Kedungmundu No 18, Semarang (024) 76740296, e-mail: adityaputraramdani@unimus.ac.id

ARTICLE INFO

History of the article :
Received 11 December 2024
Received in revised form 28 December 2024
Accepted 15 January 2025
Available online 20 January 2025

Keywords:

Chi-Square; Customer Loyalty; Feature Selection; Information Gain; kNN

*** Correspondece:**

Telepon:
+6281391793989

E-mail:
atikamutiarachim@untagsmg.ac.id

ABSTRACT

Understanding customer purchasing behavior is essential for predicting customer loyalty, which directly impacts a company's long-term success. This research aims to determine the effect of chi-square and information gain feature selection in optimizing customer loyalty classification performance, compared to pure kNN. Using a public customer purchasing behavior dataset from Kaggle, containing 10,000 data, 12 attributes with loyalty_status as the label (Gold, Regular, Silver). Evaluating performance by accuracy, kappa, classification error, recall, precision, and RMSE. The highest accuracy 91.99% was obtained by kNN k=3 with information gain, kappa 0.844, precision 95.44%, recall 86.30%, with the lowest classification error 8.01% and the second lowest RMSE 0.245, after kNN k=3 with chi-square. Results show that feature selection has a positive impact on classification, increasing accuracy and reducing errors, with the combination of the kNN k=3 method and information gain proving successful in obtaining high accuracy in classifying customer loyalty.

1. INTRODUCTION

Perkembangan teknologi digitalisasi, perubahan preferensi konsumen yang pesat menciptakan persaingan bisnis yang semakin ketat, menuntut para pelaku bisnis memahami aspek-aspek yang dapat mempengaruhi loyalitas pelanggan, salah satunya terkait pola belanja pelanggan. Loyalitas pelanggan dipengaruhi penawaran produk dan layanan yang tepat [1][2]. Pelanggan yang

loyal cenderung tetap menggunakan suatu produk/jasa secara terus-menerus, bahkan merekomendasikannya kepada orang lain, hal tersebut berdampak signifikan pada *long-term loyalty* dan profitabilitas bisnis [3]. Strategi yang dapat dilakukan untuk menjaga eksistensi bisnis adalah mempertahankan dan meningkatkan loyalitas pelanggan, baik pelanggan baru maupun pelanggan lama, sehingga dibutuhkan kemampuan prediksi yang tepat terkait loyalitas pelanggan untuk memudahkan perusahaan menjalankan strategi bisnis tersebut.

Data pelanggan merupakan salah satu aset berharga yang perlu dikelola untuk memperoleh pengetahuan, mempertahankan loyalitas pelanggan dan mendukung kemajuan bisnis. Era digital memberi kemudahan pelaku bisnis untuk melakukan prediksi loyalitas pelanggan, dengan mengolah data pelanggan yang dimiliki perusahaan menggunnakan metode machine learning [4]. Tantangan utamanya adalah menemukan metode yang paling tepat dan akurat dalam mengklasifikasi loyalitas pelanggan.

Seleksi fitur mampu meningkatkan model generalisasi, menurunkan penggunaan memori, meningkatkan efektivitas komputasi sekaligus meningkatkan efisiensi proses *learning* [5]. Penelitian terdahulu menunjukkan seleksi fitur dapat meningkatkan akurasi *performance* algoritma-algoritma klasifikasi pada berbagai dataset. Implementasi ini masih jarang diterapkan pada loyalitas pelanggan, hal tersebut merupakan state of the art pada penelitian ini terkait penerapan k-Nearest Neighbor (kNN) dipadukan dengan seleksi fitur *chi-square* dan *information gain* pada klasifikasi loyalitas pelanggan. Tujuan penelitian ini adalah mengkaji lebih lanjut penerapan teknik klasifikasi KNN, dipadukan dengan metode seleksi fitur dalam konteks loyalitas pelanggan. Penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan teknik klasifikasi loyalitas pelanggan yang lebih adaptif dan responsif.

Penelitian terdahulu menunjukkan pengaruh positif seleksi fitur terhadap performa klasifikasi. Klasifikasi data pelanggan dengan membandingkan metode klasifikasi kNN, Naive Bayes, SVM, Random Forest, SGD, ANN, AdaBoost, XgBoost dan algoritma *hybrid stacking* seperti SVM Adaboost, Random Forest Adaboost dan kNNSGD, akurasi tertinggi diperoleh kombinasi algoritma Knnsgd 92,42% dimana kNN mampu meminimalkan terjadinya eror pada klasifikasi [6].

Penelitian [7] menggunakan data *customer demographic and behavioral information* dengan atribut *age, gender, annual income, number of purchases, product category, time spent on website, loyalty program, discounts availed,* dan *label purchase status* dengan 1 untuk *yes* dan 0 untuk *no*. *Preprocessing* dengan menghilangkan *missing value, convert* variabel kategorikal dan *feature scalling*. Akurasi klasifikasi tertinggi 0.94 diperoleh oleh Random Forest, Gradient Boosting, XGBoost, kemudian 0.92 pada SVM, 0.89 kNN dan 0.82 Logistic Regression.

Penelitian klasifikasi untuk mengetahui pola belanja konsumen dengan dataset publik *Social Network Ads Kaggle* [8], dengan atribut *user id, gender, age, estimated salary* dan *purchase*, pembagian data 10-*fold cross validation*, akurasi yang diperoleh kNN 87% dan Decision Tree 95%.

Seleksi fitur *chi-square* diterapkan pada berbagai dataset, mampu meningkatkan *performance* klasifikasi. Penelitian [9] melakukan klasifikasi dataset Pima India dengan 768 data diabetes, 268 positif dan 500 negatif, seleksi fitur *chi-square* dan *information gain*, klasifikasi kNN dan SVM. Akurasi dari *chi-square* kNN 84% waktu eksekusi 0.03 detik, SVM 88% waktu eksekusi 0.02 detik, information gain kNN 82% waktu eksekusi 0.02 detik, SVM 87% waktu eksekusi 0.02 detik.

Penelitian [10] dengan tiga dataset PICK108, RGP104, CF15, pembagian 10-*fold cross validation*. Hasil menunjukkan *chi-square* memperoleh performa terbaik dibandingkan tujuh metode seleksi fitur lain. Klasifikasi dengan sepuluh algoritma lain yaitu Random Forest, kNN, AB, GB, Decision Tree, Bagging, SGD, SVM, MLP, ET, hasil akurasi diperoleh algoritma Random Forest yaitu 95% pada ketiga dataset.

Penelitian [11] memprediksi dataset publik *student performance* dari UCI, sejumlah 650 data terbagi menjadi 450 siswa berhasil, 161 siswa gagal dalam studi, dengan 32 atribut. Seleksi fitur *chi-square, split data* 70:30, klasifikasi Decision Tree, Random Forest, SVM, kNN dan XGBoost, akurasi yang dihasilkan masing-masing algoritma 87.67, 95.38, 93.07, 90.76, 91.53, 86.87.

Penelitian [12] melakukan klasifikasi data jurnal dengan metode kNN membandingkan performa seleksi fitur *gini index* dan *chi-square*, hasil menunjukkan chi-square menghasilkan performa akurasi, *precision, recall* dan f1-*score* lebih unggul. *Gini index* menghasilkan akurasi tertinggi 81.2% pada k=4, dan *chi-square* menghasilkan akurasi 85% pada k=6, sehingga terpilih menjadi model terbaik.

Seleksi fitur *information gain* banyak diterapkan pada berbagai dataset dan mampu meningkatkan *performance* klasifikasi, beberapa penelitian tersebut adalah klasifikasi penyakit jantung [13] dengan kNN dan naive bayes menggunakan seleksi fitur *information gain*. Hasil pengujian menunjukkan kombinasi information gain dan kNN memperoleh akurasi terbaik 92.31% dengan enam fitur terpilih dan nilai k=25.

Penelitian [11] melakukan klasifikasi *instrusion detection system* dengan dataset KDD CUP 99 dengan seleksi fitur *information gain* dan *correlation*, pembagian data 10-*fold cross validation*, akurasi tertinggi diperoleh kNN dengan seleksi fitur pada k=5 yaitu 99.61% dibandingkan tanpa seleksi fitur 99.59%. Klasfikasi penyakit diabetes [14] dengan metode kNN dan seleksi fitur *information gain*, dengan *split* data 90:10, 80:20, 70:30, 60:40, akurasi teringgi pada nilai k=17 sebesar 70.96%, lebih tinggi dibandingkan kNN murni 69,11%.

Penelitian terdahulu menunjukkan terdapat peningkatan *performance* klasifikasi dengan seleksi fitur pada berbagai dataset. Penelitian ini menggunakan klasifikasi kNN karena KNN memperoleh akurasi tinggi pada penelitian terdahulu, selain itu kNN tahan terhadap data yang mengandung *noise*, kNN merupakan algoritma dengan perhitungan yang sederhana dan banyak diterapkan dalam klasifikasi di berbagai bidang [15], memiliki kemampuan *learning* yang cepat sehingga cocok diterapkan pada dataset dengan jumlah yang besar [16], dalam hal ini sesuai dengan kebutuhan kapasitas dataset pelanggan yang terus bertambah. Prediksi penyakit stroke dengan metode kNN dan seleksi fitur *information gain* memperoleh akurasi tertinggi 99.85% dengan 5 fitur terpilih dan k=25 [17].

Seleksi fitur membantu mengurangi atribut yang tidak diperlukan, memilih atribut yang paling berpengaruh [18] sehingga hasil akurasi meningkat dan waktu eksekusi lebih efisien. Penelitian terdahulu menunjukkan *chi-square* dan *information gain* terbukti mampu meningkatkan performance akurasi secara signifikan pada berbagai dataset dan beragam metode klasifikasi, sehingga diterapkan pada penelitian ini.

RESEARCH METHODS

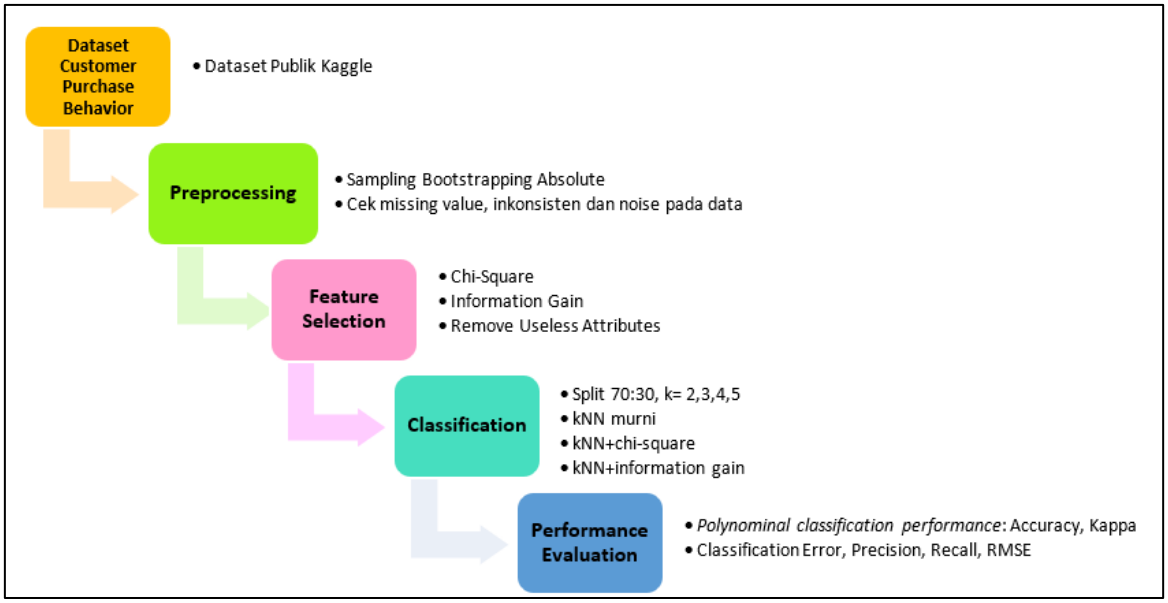
Penelitian menggunakan dataset publik *Customer Purchases Behavior* dari Kaggle, dengan 12 atribut. Atribut id tidak diolah, atribut loyalty_status merupakan variabel dependen/*class/label*, sehingga total variable independen menjadi 10.

Tabel 1. Atribut

No	Atribut	Tipe Data	Keterangan
1	id	id	id
2	age	*Integer*	Usia
3	gender	*Binominal*	Jenis kelamin (*Male, Female*)
4	income	*Integer*	Pendapatan tahunan pelanggan
5	education	*Binominal*	Pendidikan terakhir pelanggan (*Bachelor, College, Masters, Highschool*)

DOI : https:// doi.org/10.26623/6wgy1097

6	region	*Polynominal*	Wilayah asal pelanggan (*East, North, South, West*)
7	loyalty_status	*Polynominal/Label*	Tingkat loyalitas pelanggan (*Gold, Regular, Silver*)
8	purchase_frequency	*Polynominal*	Frekuensi pembelian (*Frequent, Occasional, Rare*)
9	purchase_amount	*Integer*	Jumlah total pembelian yang dilakukan pelanggan (USD)
10	product_category	*Polynominal*	Kategori produk yang dibeli *(Beauty, Books, Clothing, Electronics, Food, Health, Home*)
11	promotion_usage	*Binominal*	Pelanggan menggunakan promos (0 untuk Tidak, 1 untuk Ya).
12	satisfaction_score	*Integer*	Skor kepuasan pelanggan



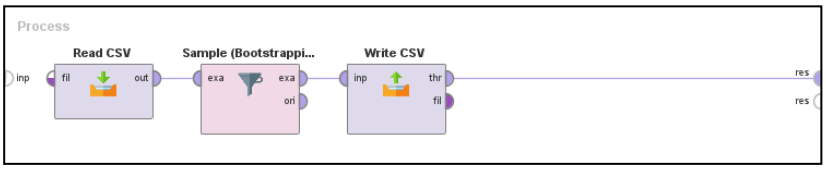
Gambar 1. Alur Penelitian

Metode *split data* 70:30 dipilih karena memberikan keseimbangan optimal, fleksibel, memberikan confidence level yang baik dalam evaluasi. Preprocessing dilakukan guna memastikan data yang diolah lengkap dan konsisten. Klasifikasi kNN berdasarkan perhitungan *jarak euclidean distance*, dengan nilai k 2,3,4 dan 5. Evaluasi performance paling utama dilihat dari akurasi dan nilai kappa karena menggunakan operator *performance polynominal classification*, namun pada penelitian ini dilengkapi sebagai pertimbangan tambahan dalam melakukan perbandingan *performance*.

1. ***Preprocessing***

   Preprocessing dilakukan dengan *sampling*, pengecekan *missing value* dan seleksi fitur. Batasan teknologi dan sumber daya yang tersedia mengharuskan peneliti mengurangi jumlah dataset, sehingga dilakukan *sampling bootstrapping absolute*. Metode *sample bootstrapping* mampu menghasilkan data yang lebih ringkas namun tetap representative, untuk menghindari

bias dan *overfitting* sekaligus meningkatkan stabilitas algoritma [19]. Hasil *sampling* di ekspor menjadi dataset baru berjumlah 10.000 data.



Gambar 2. *Sampling Bootstrapping Absolute*

Seleksi atribut dilakukan dengan weighting by *chi-square* dan *information gain*, kemudian hapus atribut dengan minimal deviation 0.05.

AttributeWeights (Weight by Chi Squared Statistic)

attribute	weight
promotion_usage	0
gender	0.041
purchase_frequency	0.083
education	0.442
age	0.625
region	0.710
income	0.779
product_category	0.824
purchase_amount	0.844
satisfaction_score	1

AttributeWeights (Weight by Information Gain)

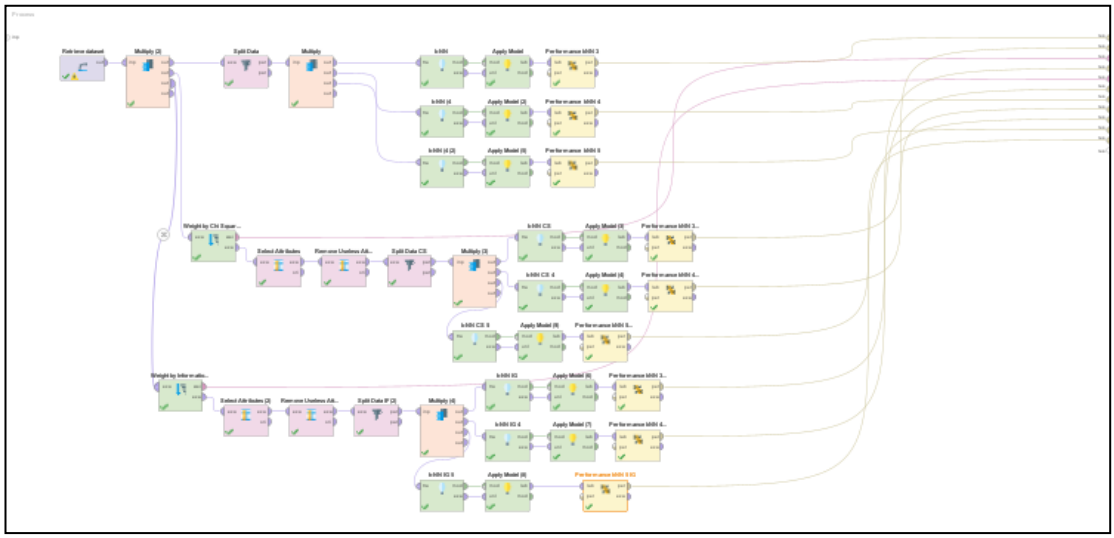
attribute	weight
promotion_usage	0
gender	0.049
purchase_frequency	0.101
age	0.248
satisfaction_score	0.347
education	0.539
income	0.680
purchase_amount	0.716
region	0.882
product_category	1

Gambar 3. *Attribute Weights*

2. Klasifikasi kNN

Data dibagi dengan metode *split* 70:30, klasifikasi *polynominal* kNN dengan perhitungan jarak *euclidean distance*. Evaluasi *performance* dilihat dari perolehan nilai *accuracy* dan kappa [10], dilengkapi nilai *classification error*, *recall*, *precision* dan RMSE sebagai pertimbangan tambahan dalam melakukan perbandingan hasil. Akurasi 100% diperoleh kNN k=2, namun akurasi menurun ketika data testing dihubungkan langsung dari operator *split data*, sehingga *overfitting* dan kNN k=2 dihapus. Perbandingan hasil pengujian kNN dilakukan untuk nilai k=3, 4 dan 5.

DOI : https:// doi.org/10.26623/6wgy1097



Gambar 4. Klasifikasi kNN

RESULTS AND DISCUSSION

Hasil menunjukkan pada seluruh nilai k 3,4 dan 5, terdapat peningkatan nilai akurasi serta penurunan nilai eror setelah dilakukan seleksi fitur dengan *chi-square* dan *information gain*, dibandingkan hanya menggunakan metode kNN murni tanpa seleksi fitur, menunjukkan penerapan seleksi fitur pada penelitian ini mendukung performa klasifikasi yang dihasilkan semakin baik.

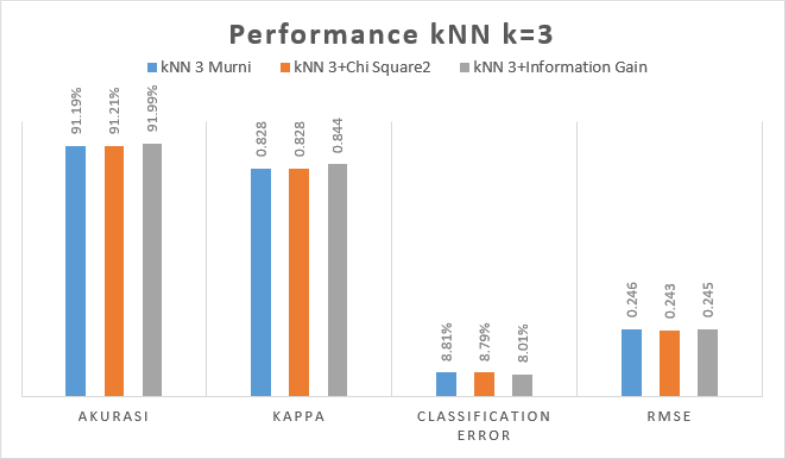
Tabel 2. *Performance*

***Split* 70:30**	**Akurasi**	**Kappa**	***Classification Error***	***Recall***	***Precision***	**RMSE**
kNN k=3 Murni	91.19%	0.828	8.81%	83.85%	94.98%	0.246
kNN k=4 Murni	84.23%	0.698	15.77%	76.07%	81.84%	0.341
kNN k=5 Murni	77.51%	0.555	22.49%	65.82%	75.71%	0.800
kNN k=3 + CS	91.21%	0.828	8.79%	84.13%	95.07%	**0.243**
kNN k=4 + CS	85.04%	0.714	14.96%	77.46%	83.54%	0.338
kNN k=5 + CS	77.63%	0.558	22.37%	66.91%	77.35%	0.395
kNN k=3 + IG	**91.99%**	**0.844**	**8.01%**	**86.30%**	**95.44%**	0.245
kNN k=4 + IG	85.11%	0.716	14.89%	77.79%	82.85%	0.338
kNN k=5 + IG	78.11%	0.570	21.89%	67.96%	77.10%	0.394

Evaluasi dilihat dari nilai *performance*, metode yang paling baik adalah yang memperoleh akurasi, *recall* dan precision paling tinggi, dengan nilai *classification error* serta RMSE yang rendah menunjukkan tingkat kesalahan klasifikasi terjadi dalam jumlah yang sedikit. Akurasi merupakan. *Recall* menunjukkan kemampuan klasifikasi dalam mengidentifikasi selutruh data positif secara benar (*true positif*). *Precision* mengukur kemampuan model klasifikasi dalam mengidentifikasi contoh positif dengan benar dari semua yang diidentifikasi sebagai positif. Nilai kappa 1 artinya sempurna, sehingga semakin mendekati 1 menujukkan kinerja klasifikasi yang semakin baik. Akurasi tertinggi diperoleh kNN k=3+IG yaitu 91.99% dengan nilai kappa tertinggi 0.844 menunjukkan kesesuaian hasil klasifikasi yang sangat baik, perolehan nilai *precision* dan *recall* yaitu 95.44% dan 86.30%. *Classification error* terendah dibandingkan metode lain yaitu 8.01%, RMSE 0.245, terendah kedua setelah kNN k=3+CS.

Performance terbaik pada urutan kedua diperoleh kNN k=3+CS dengan akurasi 91.21%, nilai kappa 0.828, precision 95.07%, recall 84.13%, classification error 8.79% dan RMSE terendah dan terbaik 0.243. Dibandingkan dengan kNN k=3+IG, kNN k=3+CS hanya unggul pada nilai RMSE sehingga secara keseluruhan perolehan *performance* terbaik adalah dari kNN k=3+IG.

Pada urutan ketiga, kNN k=3 Murni memperoleh *performance* lebih baik dibandingkan kNN k=4 dan 5 murni maupun dengan seleksi fitur. Hal tersebut menunjukkan adanya pengaruh pemilihan nilai k dari klasifikasi kNN [20] dimana k=3 memperoleh hasil paling baik dibandingkan k=4 dan 5.



Gambar 5. Perbandingan *Performance* kNN k=3

Tabel 3, 4 dan 5 merupakan *confusion matrix* dari metode kNN k=3 Murni, kNN k=3+IG dan kNN k=3+CS untuk menampilkan jumlah data yang diklasifikasikan dengan benar dan salah, memberikan gambaran menyeluruh hasil klasifikasi, membantu mengidentifikasi *class/label* yang paling sering salah saat diklasifikasi [20].

Tabel 3. *Confusion Matrix* kNN k=3 Murni

kNN Murni	***True Regular***	***True Silver***	***True Gold***
Pred Regular	4193	414	146
Pred Silver	0	1687	57
Pred Gold	0	0	503

Tabel 4. *Confusion Matrix* kNN k=3+IG

kNN+IG	***True Regular***	***True Silver***	***True Gold***
Pred Regular	4193	408	104
Pred Silver	0	1693	49
Pred Gold	0	0	553

Tabel 5. *Confusion Matrix* kNN k=3+CS

kNN+CS	***True Regular***	***True Silver***	***True Gold***
Pred Regular	4193	420	144
Pred Silver	0	1681	51
Pred Gold	0	0	511

DOI : https:// doi.org/10.26623/6wgy1097

KESIMPULAN DAN REKOMENDASI

Hasil penelitian menunjukkan seleksi fitur terbukti meningkatkan *performance* klasifikasi kNN dari peningkatan nilai akurasi dan rendahnya nilai classification error, RMSE. Performance terbaik diperoleh klasifikasi kNN dengan k=3, dimana seleksi fitur information gain memberikan hasil paling optimal dibandingkan kNN murni dan kNN dengan *chi-square*. Metode kNN k=3+IG memperoleh akurasi 91.99% dengan nilai kappa 0.844, nilai precision dan recall 95.44% dan 86.30%, merupakan perolehan tertinggi dibandingkan metode lainnya. *Classification error* terendah dibandingkan metode lain yaitu 8.01%, dan RMSE terendah kedua setelah kNN k=3+CS dengan nilai RMSE 0.245 menunjukkan nilai eror/kesalahan yang rendah dalam melakukan klasifikasi.

Perolehan akurasi yang tinggi dan nilai *classification error* yang rendah, membuktikan bahwa *state of the art* penelitian ini terkait penerapan kNN dipadukan dengan seleksi fitur *chi-square* dan *information gain* mampu menghasilkan *performance* yang baik, sehingga sangat tepat diterapkan dalam klasifikasi loyalitas pelanggan, terlebih kNN memiliki kemampuan *learning* yang cepat sehingga sangat sesuai digunakan pada dataset dengan jumlah yang besar, dimana data pelanggan pasti terus bertambah seiring berjalannya bisnis. Penelitian ini mengusulkan metode kNN k=3+IG sebagai metode dengan kinerja terbaik.

REFERENCES

[1] W. N. Wassouf, R. Alkhatib, K. Salloum, and S. Balloul, "Predictive Analytics using Big Data for Increased Customer Loyalty: Syriatel Telecom Company Case Study," *J Big Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00290-0.

[2] K. Tarnowska, Z. W. Ras, and L. Daniel, "Recommender System for Improving Customer Loyalty," in *Studies in Big Data* , vol. 55, 2020.

[3] A. Mutiarachim and N. A. Yuniarti, "Jurnal Sistem Informasi, Manajemen, dan Akuntansi (SIMAK) The Role of Driver Services and Application Quality in Enhancing Gojek Customer Loyalty Through Satisfaction".

[4] M. Kimura, "Customer Segment Transition Through the Customer Loyalty Program," *Asia Pacific Journal of Marketing and Logistics*, vol. 34, no. 3, pp. 611–626, Feb. 2022, doi: 10.1108/APJML-09-2020-0630.

[5] Z. Deng, Z. Zheng, D. Deng, T. Wang, Y. He, and D. Zhang, "Feature Selection for Multi-Label Learning Based on F-Neighborhood Rough Sets," *IEEE Access*, vol. 8, pp. 39678–39688, 2020, doi: 10.1109/ACCESS.2020.2976162.

[6] G. Chaubey, P. R. Gavhane, D. Bisen, and S. K. Arjaria, "Customer Purchasing Behavior Prediction using Machine Learning Classification Techniques," *J Ambient Intell Humaniz Comput*, vol. 14, no. 12, pp. 16133–16157, Dec. 2023, doi: 10.1007/s12652-022-03837-6.

[7] E. Deniz and S. Ç. Bülbül, "Predicting Customer Purchase Behavior Using Machine Learning Models," *Information Technology in Economics and Business*, Jul. 2024, doi: 10.69882/adba.iteb.2024071.

[8] V. Umarani, "Investigation of KNN and Decision Tree Induction Modelin Predicting Customer Buying Pattern," European Alliance for Innovation n.o., Jan. 2022. doi: 10.4108/eai.7-12-2021.2314593.

[9] A. S. Jaddoa and Z. T. M. Al-Ta'i, "Diagnosis of Diabetes Mellitus using (chi square-information gain) selectors and (SVM and KNN) Classifiers," in *AIP Conference Proceedings*, American Institute of Physics Inc., Mar. 2023. doi: 10.1063/5.0102761.

[10] M. Onesime, Z. Yang, and Q. Dai, "Genomic Island Prediction via Chi-Square Test and Random Forest Algorithm," *Comput Math Methods Med*, vol. 2021, 2021, doi: 10.1155/2021/9969751.

[11] H. Bhoria, A. Dhankhar, and K. Solanki, "INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY Chi-Square Feature Selection Technique for Student's performance prediction," */ Indian Journal of Science and Technology*, vol. 16, no. 38, pp. 3250–3257, 2023, doi: 10.17485/IJST/v16i38.921.

[12] F. Istighfarizky, N. A. S. ER, I. M. Widiartha, L. G. Astuti, I. G. N. A. C. Putra, and I. K. G. Suhartana, "Klasifikasi Jurnal menggunakan Metode KNN dengan Mengimplementasikan Perbandingan Seleksi Fitur," *Jurnal Elektronik Ilmu Komputer Udayana*, vol. 11, no. 1, pp. 167–176, Aug. 2022, [Online]. Available: https://scholar.google.com

[13] S. Hidayatul, A. Aini, Y. A. Sari, and A. Arwan, "Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes," 2018. [Online]. Available: http://j-ptiik.ub.ac.id

[14] N. Devian *et al*., "Prediksi Penyakit Diabetes dengan Metode K-Nearest Neighbor (kNN) dan Seleksi Fitur Information Gain," 2024.

[15] W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," *IEEE Access*, vol. 8, pp. 28808–28819, 2020, doi: 10.1109/ACCESS.2019.2955754.

[16] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning k for kNN Classification," *ACM Trans Intell Syst Technol*, vol. 8, no. 3, Jan. 2017, doi: 10.1145/2990508.

[17] F. Nabiel Syahreza, P. Nurul Sabrina, E. Ramadhan Teknik Informatika, U. Jendral Achmad Yani Jl Terusan Jend Sudirman, J. Barat, and K. Cimahi, "Prediksi Penyakit Stroke Menggunakan Metode K-Nearest Neigbors dan Information Gain," *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 6, Dec. 2024.

[18] Y. Wang and C. Zhou, "Feature Selection Method Based on Chi-Square Test and Minimum Redundancy," in *Emerging Trends in Intelligent and Interactive Systems and Applications (IISA 2020)*, M. Tavana, N. Nedjah, and R. Alhajj, Eds., Emerging Trends in Intelligent and Interactive Systems and Applications (IISA 2020), Dec. 2020, pp. 171–178.

[19] K. Jain and R. Jindal, "Sampling and noise filtering methods for recommender systems: A literature review," *Eng Appl Artif Intell*, vol. 122, p. 106129, Jun. 2023, doi: 10.1016/J.ENGAPPAI.2023.106129.

[20] R. B. Widodo, *Machine Learning Metode k-Nearest Neighbors Klasifikasi Angka Bahasa Isyarat*. Malang: Media Nusa Creative, 2022.

DOI : https:// doi.org/10.26623/6wgy1097