**JURNAL MEDIA INFORMATIKA BUDIDARMA**
**Volume 8, Nomor 1, Januari 2024, Page 612-623**
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7074

# Prediksi Retensi Pengguna Baru Shopee Menggunakan Machine Learning

**Wahyu Fajrin Mustafa[*], Syarif Hidayat, Dhomas Hatta Fudholi**

Fakultas Teknologi Industri, Magister Informatika, Universitas Islam Indonesia, Yogyakarta, Indonesia
Email: [1,*]19917037@students.uii.ac.id, [2]055230703@uii.ac.id, [3]085230103@uii.ac.id
Email Penulis Korespondensi: 19917037@students.uii.ac.id

**Abstrak**−Shopee telah berkembang menjadi salah satu platform e-commerce terkemuka yang menghubungkan penjual dengan konsumen. Namun, tantangan untuk menjaga pengguna tetap aktif dan terlibat dengan platform menjadi semakin kompleks. Retensi pengguna, yaitu kemampuan sebuah platform dalam mempertahankan dan meningkatkan kehadiran pengguna, serta faktor kunci dalam kesuksesan jangka panjang suatu platform e-commerce. Memahami faktor-faktor yang mempengaruhi keputusan pengguna untuk terus aktif atau berhenti berinteraksi dengan platform melibatkan analisis dari berbagai variabel, termasuk perilaku pengguna, preferensi, pengalaman berbelanja, dan interaksi dengan platform tersebut. Penelitian ini dirancang untuk mengembangkan model prediksi retensi pengguna yang efektif dengan menggunakan data dari pengguna baru di Shopee. Dengan menganalisis data tersebut dan menerapkan teknik pembelajaran mesin menggunakan metode Logistic Regression, Decision Tree, Gaussian Naive Bayes, Random Forest, KNN (K-Nearest Neighbors), MLP (Multi-Layer Perceptron), AdaBoost, dan XGBoost, penelitian ini memprediksi retensi pengguna dalam periode 14 hari setelah pendaftaran di Shopee. Hasil dari penelitian ini menunjukan bahwa model Random Forest memiliki kinerja terbaik dengan nilai Accuracy 0.733677, Precision 0.702161, Recall 0.811626, dan F1-Score 0.752936. Nilai cross-validation menunjukkan konsistensi model dengan Accuracy 0.727626, Precision 0.698143, Recall 0.801884, dan F1-Score 0.746328. Model Random Forest menjadi model dengan nilai recall tinggi, menunjukkan sensitivitas yang baik dalam mengidentifikasi pengguna yang bertahan.

Dengan demikian, hasil penelitian ini memberikan wawasan yang berharga bagi Shopee dalam mengembangkan strategi retensi pada pengguna baru, yang merupakan aspek penting dalam pertumbuhan dan keberlanjutan bisnis e-commerce.

**Kata Kunci:** Prediksi; Retensi Pengguna; Shopee; E-commerce; Pembelajaran Mesin

**Abstract**−Shopee has evolved into one of the leading e-commerce platforms connecting sellers with consumers. However, the challenge of keeping users active and engaged on the platform has become increasingly complex. User retention, the ability of a platform to sustain and enhance user presence, is a key factor in the long-term success of an e-commerce platform. Understanding the factors influencing users' decisions to remain active or cease interactions with the platform involves analyzing various variables, including user behavior, preferences, shopping experiences, and interactions with the platform. This research is designed to develop an effective user retention prediction model using data from new Shopee users. By analyzing the data and applying machine learning techniques using Logistic Regression, Decision Tree, Gaussian Naive Bayes, Random Forest, KNN (K-Nearest Neighbors), MLP (Multi-Layer Perceptron), AdaBoost, and XGBoost methods, this study predicts user retention within a 14-day period after registration on Shopee. The results of this research indicate that the Random Forest model performs the best with an Accuracy value of 0.733677, Precision of 0.702161, Recall of 0.811626, and F1-Score of 0.752936. Cross-validation values demonstrate the model's consistency with an Accuracy of 0.727626, Precision of 0.698143, Recall of 0.801884, and F1-Score of 0.746328. The Random Forest model becomes a model with a high recall value, indicating good sensitivity in identifying users who retain. Consequently, the results of this research provide valuable insights for Shopee in developing retention strategies for new users, which is an important aspect in the growth and sustainability of the e-commerce business.

**Keywords**: Prediction; User Retention; Shopee; E-commerce; Machine Learning

## 1. PENDAHULUAN

Perkembangan e-commerce telah mengalami pertumbuhan pesat dan menjadi bagian integral dari pola konsumsi masyarakat. Shopee, sebagai salah satu platform e-commerce terkemuka di Asia Tenggara, sebagai peran penting dalam menghubungkan penjual dengan konsumen. Namun, dengan persaingan yang semakin ketat dalam industri ini, menjaga pengguna agar tetap terlibat dan aktif di platform merupakan tantangan yang signifikan. Berdasarkan data yang dikumpulkan oleh iPrice, pada kuartal kedua tahun 2022, Shopee memiliki rata-rata 131,3 juta pengunjung website per bulan. Angka ini lebih rendah dibandingkan dengan Tokopedia, yang berhasil menarik 158,3 juta pengunjung per bulan pada periode yang sama. Sebelum pandemi, pada kuartal ketiga tahun 2019, Shopee mencatatkan 56 juta pengunjung per bulan. Jumlah ini terus bertambah selama pandemi, tetapi trennya menunjukkan penurunan selama dua kuartal pertama tahun 2022. Secara kumulatif, dari kuartal ketiga tahun 2019 hingga kuartal kedua tahun 2022, jumlah pengguna Shopee telah tumbuh sekitar 134% [1].

Retensi adalah suatu konsep yang mengacu pada kemampuan suatu platform atau layanan untuk mempertahankan dan memperpanjang kehadiran pengguna. Dalam pemasaran, retensi mencerminkan usaha untuk mempertahankan dan memperpanjang hubungan dengan pelanggan. Tingkat retensi yang tinggi menjadi indikator kunci keberhasilan dalam menjaga pelanggan yang sudah ada. Upaya retensi bertujuan untuk memastikan bahwa pelanggan tetap merasa puas dan terlibat dengan produk atau layanan yang disediakan. Retensi pengguna, yang merujuk pada kemampuan platform untuk mempertahankan dan memperpanjang kehadiran pengguna, menjadi faktor kunci dalam kesuksesan jangka panjang sebuah platform e-commerce. Pengguna yang tetap aktif cenderung

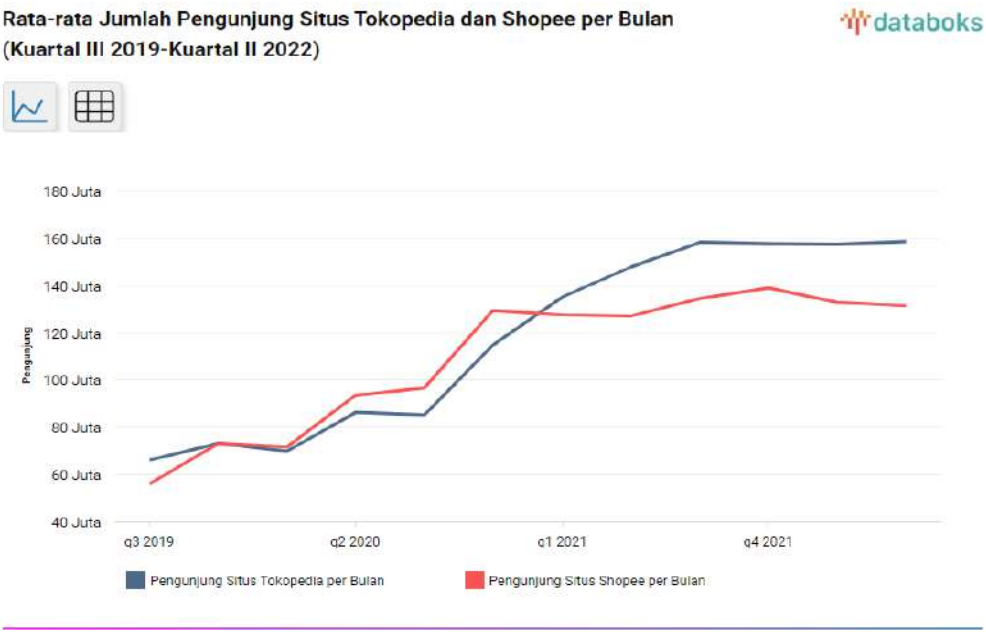

Submitted: **05/12/2023**; Accepted: **30/01/2024**; Published: **31/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
**Volume 8, Nomor 1, Januari 2024, Page 612-623**
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7074

memberikan kontribusi pada pendapatan yang stabil, membantu membangun dan menciptakan lingkungan yang sehat bagi penjual dan konsumen.

Meskipun pentingnya retensi pengguna telah diakui secara luas, memahami faktor-faktor yang mempengaruhi keputusan pengguna untuk tetap aktif atau berhenti berinteraksi dengan platform e-commerce merupakan tugas yang kompleks. Ini melibatkan berbagai variabel, termasuk perilaku pengguna dalam berinteraksi dengan platform. Dalam konteks ini, penelitian ini bertujuan untuk membuat model prediksi retensi pengguna yang efektif berdasarkan data yang dikumpulkan dari pengguna baru Shopee. Melalui analisis data dan menerapkan metode machine learning, penelitian ini akan mencoba memprediksi kecenderungan retensi pengguna dalam jangka waktu 14 hari setelah mereka mendaftar di Shopee.

Penelitian ini memiliki relevansi yang signifikan dalam konteks bisnis e-commerce dan dapat membantu Shopee dalam mengoptimalkan strategi retensi pengguna. Machine learning menjadi solusi untuk memprediksi retensi pengguna dengan tingkat akurasi yang tinggi. Machine learning merupakan salah satu teknologi untuk mengolah data dan mempermudah analisis data dalam jumlah besar. Dalam beberapa tahun terakhir, meluasnya penggunaan pembelajaran mesin di berbagai sektor telah menyebabkan penerapannya secara luas. Gambar menunjukan grafik perbandingan pengunjung situs Tokopedia dan Shopee.




**Gambar 1.** Grafik Pengunjung Situs Tokopedia dan Shopee per Bulan

Studi yang dilakukan pada tahun 2023 fokus pada penggunaan algoritma machine learning untuk meningkatkan prediksi churn klien. Mereka mengevaluasi potensi jaringan saraf buatan dalam mengidentifikasi klien yang berisiko churn sembilan bulan sebelumnya untuk memungkinkan perusahaan merancang strategi retensi yang lebih efektif. Hasil studi menunjukkan bahwa algoritma Stochastic Gradient Booster memiliki akurasi tertinggi 83,9%, diikuti oleh KNN dengan 82,9%, Random Forest dengan 82,6%, dan Logistic Regression dengan 78,1% [2].

Menggunakan model Random Forest dan XGBoost dengan menggunakan teknik optimasi parameter GridSearchCV sebagai hyperparameter. Dataset yang digunakan yaitu data telekomunikasi pada salah satu perusahaan. Kemudian dari kedua metode tersebut dibandingkan untuk mengetahui metode mana yang memiliki akurasi paling baik dalam memprediksi Customer Churn. Hasil penelitian mendapat nilai akurasi Random Forest 93.5% dan XGBoost 95.6% [3].

Studi yang dilakukan pada tahun 2021 mengaplikasikan metode oversampling SMOTE dan teknik Boosting untuk memprediksi churn pelanggan. Teknik-teknik ini diintegrasikan dengan metode klasifikasi seperti Random Forest, Naïve Bayes, Decision Tree, K-Nearest Neighbor, dan Deep Learning. Hasil dari studi ini menunjukkan bahwa penggunaan SMOTE meningkatkan akurasi rata-rata sebesar 3% dan penggunaan AdaBoost meningkatkan akurasi rata-rata sebesar 8%. Dari semua metode yang diuji, Random Forest memberikan akurasi tertinggi yaitu 89,19% [4].

Dalam industri perbankan yang kompetitif, retensi nasabah adalah masalah yang krusial untuk menarik nasabah baru dan membangun kepercayaan serta rujukan dari nasabah yang ada. Peneliti menganalisis data bank untuk memprediksi nasabah yang berisiko tinggi berhenti menggunakan layanan bank dengan menggunakan algoritma seperti logistic regression, SVM, random forest, dan XGBoost. Hasil studi menunjukkan bahwa XGBoost memiliki akurasi tertinggi (83,9%), logistic regression memiliki sensitivitas tertinggi (71,4%), sementara random forest menonjol sebagai yang terbaik secara keseluruhan dengan akurasi 78,3% dan sensitivitas 69,3%, menunjukkan kemampuan baik dalam menangani data besar dan banyak fitur [5].


Submitted: **05/12/2023**; Accepted: **30/01/2024**; Published: **31/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 612-623
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7074

Dalam studi yang dilakukan pada tahun 2019, dibandingkan efektivitas antara Regresi Logistik dan XGBoost dalam memprediksi customer churn pada platform e-commerce menggunakan 25 indeks. Metode Regresi Logistik mencapai akurasi 75,9%, precision 75,7%, dan recall 83,6%. Sedangkan XGBoost mencapai akurasi yang sedikit lebih tinggi yaitu 76,6%, dengan precision 76,3%, dan recall 84,2%. Dari 25 indeks, 10 di antaranya diidentifikasi sebagai yang paling penting untuk memprediksi churn. Secara keseluruhan, XGBoost terbukti lebih akurat daripada Regresi Logistik [6].

Studi yang dilakukan dengan mengeksplorasi sektor Educational Technology (EdTech), khususnya Pembelajaran berbasis permainan digital (DGBL) untuk meningkatkan retensi pengguna. Prediksi churn dianggap penting untuk retensi pelanggan dan memungkinkan penciptaan strategi pemasaran yang lebih efektif. Mereka menggunakan model Decision tree, Random Forest, dan Logistic regression untuk prediksi churn. Hasil menunjukkan bahwa Logistic regression mengungguli model lain dengan Precision 0.9228, Recall 0.9185, F1-Score 0.9194, dan AUC 0.9225, menandakan tingkat presisi dan recall yang seimbang dan performa prediksi yang sangat baik dalam konteks game online dan pendidikan [7].

Studi yang dilakukan menggunakan data dari industri telekomunikasi Nepal, mencakup 52,332 catatan pelanggan dengan 46,204 non-churn dan 6,128 churn. Dengan menerapkan algoritma XGBoost, penelitian ini berhasil mencapai akurasi 97% dan skor F1 88% pada dataset asli. Selain itu, penelitian ini juga menguji pada dataset publik yang meliputi 3,333 pelanggan, hasilnya menunjukkan peningkatan akurasi dan skor F1 menjadi 96,25% dan 86,34%, bertujuan untuk membandingkan dengan studi-studi sebelumnya [8].

Studi yang mengembangkan model prediksi churn untuk pelanggan e-commerce menggunakan data dari e-commerce Brasil yang terdiri dari 11,224 sampel. Dari jumlah tersebut, 75% (8,418 sampel) digunakan untuk pelatihan dan sisanya 25% (2,806 sampel) untuk pengujian. Model dievaluasi berdasarkan akurasi, sensitivitas, spesifisitas, nilai positif sebenarnya, dan nilai negatif sebenarnya. Hasil evaluasi menunjukkan akurasi tinggi pada berbagai algoritma, dengan Neural Network mencapai 98,86%, SVM 97,57%, Naïve Bayes 94,76%, Random Forest 99,68%, dan Adam 96,08%. Random Forest, yang fiturnya dipilih melalui teknik pemilihan fitur NCA, mencapai akurasi dan kinerja paling tinggi dalam penelitian tersebut [9].

Penelitian menggunakan metode SMOTE untuk mengatasi data tidak seimbang dan dilanjutkan dengan klasifikasi Adaboost. Data yang digunakan adalah kualitas wine dengan proporsi kelas mayoritas dan minoritas sebesar 0,86:0,14. Rasio pembagian data training dan testing adalah 70%:30 %. Hasil analisis yang didapatkan nilai AUC untuk model SMOTE di klasifikasi menggunakan algoritma Adaboost yaitu 0,784 lebih baik dibandingkan dengan model Adaboost yaitu 0,664 [10].

Membandingkan tiga metode yaitu K-NN, Random Forest, dan XGBoost untuk memprediksi customer churn pada perusahaan telekomunikasi. Hasil penelitian dari ketiga metode klasifikasi didapatkan tingkat akurasi untuk K-NN senilai 75,4%, metode Random Forest dengan tingkat akurasi senilai 77,5%, dan metode XGBoost dengan tingkat akurasi 79,8%. Hasil penelitian yang didapatkan dari ujicoba adalah metode XGBoost yang lebih unggul [11].

Penelitian dilakukan dengan membandingkan metode XGBoost dengan Regresi Logistik dalam prediksi pelanggan, menggunakan data dari Oktober 2017 hingga Maret 2018 dan melibatkan tuning hyperparameter. Hasilnya, XGBoost lebih akurat (97,8%) dan memiliki nilai ROC-AUC yang lebih tinggi (0,99) dibandingkan Regresi Logistik (akurasi 90,7% dan ROC-AUC 0,81), menunjukkan keunggulan XGBoost dalam menangani data yang tidak seimbang [12].

Penelitian yang dilakukan dengan membandingkan kinerja XGBoost dan Random Forest dalam mengklasifikasikan data sekuens DNA, dengan melakukan tuning hyperparameter melalui grid search. Mereka menemukan bahwa keduanya adalah metode klasifikasi yang efektif, dan melalui optimasi hyperparameter, kedua metode tersebut meningkatkan akurasi klasifikasi. XGBoost, khususnya, menunjukkan tingkat akurasi yang lebih tinggi daripada Random Forest, yaitu 95,6% sebelum tuning dan 96,2% setelah tuning hyperparameter [13].

Penelitian yang dilakukan dengan membandingkan berbagai teknik machine learning termasuk LDA, Decision Tree, KNN, SVM, Logistic Regression, dan teknik pembelajaran berbasis ensemble seperti Random Forest, AdaBoost, dan Stochastic Gradient Boosting, serta Naïve Bayesian, dan MLP. Model-model ini diterapkan pada dataset telekomunikasi dengan 3,333 catatan. Hasilnya menunjukkan bahwa Random Forest dan AdaBoost memberikan kinerja terbaik dengan akurasi 96%. MLP dan SVM mencatat akurasi 94%, Decision Tree 90%, Naïve Bayesian 88%, dan Logistic Regression serta LDA mendapat 86,7% [14].

Berdasarkan hasil penelitian sebelumnya, penggunaan machine learning untuk prediksi retensi telah banyak dilakukan. Namun untuk prediksi retensi pengguna baru di Shopee, belum ada penelitian yang dilakukan. Penelitian ini memanfaatkan model algoritma machine learning yang telah teruji mencapai tingkat akurasi tinggi dalam penelitian-penelitian sebelumnya seperti Logistic Regression, Decision Tree, Naïve Bayes, Random Forest, KNN, MLP, AdaBoost, dan XGBoost, bertujuan untuk mengadaptasi dan mengoptimalkan penggunaan model untuk retensi pengguna baru Shopee, guna menghasilkan prediksi retensi pengguna yang lebih akurat. Peneliti melakukan analisis untuk membuat model yang dapat memprediksi dan menganalisis faktor-faktor yang mempengaruhi retensi pengguna baru pada platform e-commerce Shopee. Model yang digunakan dalam analisis penelitian ini menerapkan teknik pembelajaran mesin menggunakan metode Logistic Regression, Decision Tree, Gaussian Naive Bayes, Random Forest, KNN (K-Nearest Neighbors), MLP (Multi-Layer Perceptron), AdaBoost, dan XGBoost. Model-model ini telah menunjukkan kinerja yang baik pada penelitian sebelumnya dan mampu

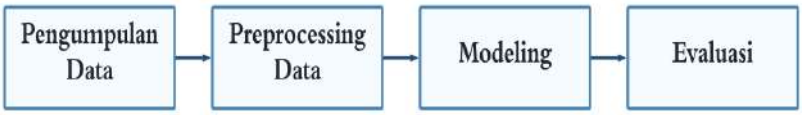

Submitted: **05/12/2023**; Accepted: **30/01/2024**; Published: **31/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 612-623
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7074

merepresentasikan informasi dengan baik. Dengan demikian, penelitian ini dapat memberikan pemahaman yang lebih mendalam tentang retensi pengguna dalam industri e-commerce. Model prediksi yang dihasilkan dapat memberikan wawasan yang bermanfaat bagi platform e-commerce Shopee untuk mengambil langkah strategis yang efektif dalam meningkatkan retensi pada pengguna baru.

## 2. METODOLOGI PENELITIAN

Penelitian ini dilakukan menggunakan bahasa pemrograman Python, yang dipilih karena kemampuannya yang unggul dalam analisis data dan machine learning. Berbagai langkah dilakukan, mulai dari analisis deskriptif, pra-pemrosesan data, hingga penggunaan library Scikit-learn untuk menerapkan berbagai metode klasifikasi machine learning. Metode tersebut termasuk Logistic Regression, Decision Tree, Gaussian Naive Bayes, Random Forest, Extra Trees, KNN, Multi-Layer Perceptron, AdaBoost, dan XGBoost. Selain itu, imbalanced-learn untuk penerapan teknik SMOTE juga digunakan untuk oversampling, yang bertujuan mengatasi ketidakseimbangan data. Proses ini dilanjutkan dengan evaluasi kinerja dari metode klasifikasi yang telah digunakan. Gambar 2 menunjukan tahapan penelitian yang dilakukan.




**Gambar 2.** Tahapan Penelitian

### 2.1 Pengumpulan Data

Penelitian ini menggunakan data sekunder dari pengguna baru Shopee yang diperoleh dari periode 1 Januari hingga 31 Mei 2021. Dataset yang digunakan, berjudul 'Shopee New User Behavior', tersedia di platform Kaggle [15]. Dataset ini terdiri dari 222,378 entri data yang mencakup 38 variabel. Variabel-variabel ini menggambarkan berbagai karakteristik dari pengguna baru di platform Shopee.

### 2.2 Preprocessing Data

Preprocessing data adalah proses pengolahan data yang dilakukan pada data mentah sebelum digunakan dalam analisis dan pemodelan. Tujuannya adalah untuk mempersiapkan data agar lebih terstruktur, bersih, dan sesuai untuk diolah lebih lanjut. Dalam preprocessing, data yang awalnya mungkin tidak konsisten, berisi kecacatan, atau memiliki format yang tidak sesuai, diubah menjadi format yang lebih standar dan efisien untuk proses analisis. Proses ini melibatkan langkah-langkah seperti pembersihan data, transformasi fitur, normalisasi, dan penanganan masalah umum seperti nilai yang hilang, data terduplikasi, outliers, dan ketidakseimbangan kelas. Melalui preprocessing, data menjadi lebih akurat dan representatif, sehingga meningkatkan keandalan dan efektivitas model machine learning yang akan dibangun. Tahapan preprocessing yang dilakukan pada penelitian ini sebagai berikut:

#### 2.2.1 Handling Missing Values

Proses ini melibatkan identifikasi dan pengelolaan nilai-nilai yang kosong atau hilang dalam dataset. Penanganan nilai yang hilang dapat dilakukan dengan beberapa metode, termasuk penghapusan baris atau kolom yang mengandung nilai hilang atau pengisian nilai kosong dengan teknik imputasi. Teknik imputasi yang umum digunakan adalah imputasi rata-rata, median, modus, atau dengan nilai konstan. Berdasarkan analisis yang dilakukan dalam penelitian ini, terdapat variabel dalam dataset yang menunjukkan persentase nilai yang hilang dalam jumlah signifikan dapat dilihat pada tabel 1. Variabel top_up_14d memiliki jumlah data yang kosong paling banyak, dengan 204,966 entri atau 92.17% dari keseluruhan data. Ini diikuti oleh total_voucher_claim_14d yang memiliki 97,907 entri data yang kosong, yang merupakan 44.03% dari data. Variabel lainnya shop_flash_sale, shop_normal_shop, shop_sbs, shop_cb, shop_ss, shop_ss_plus, shop_mall, use_hemat, use_regular, use_nextday, use_instant, use_cc_debit, use_va_bt, use_cod, use_shopeepaylater, use_shopeepay, total_order_14d, gmv_14d, dan use_sameday memiliki jumlah data yang kosong, yaitu 67,577 atau 30.39% dari total data. Sementara itu, shop_views_14d memiliki 60,857 atau 27.37% data yang hilang, yang menunjukkan tingkat interaksi pengguna dengan toko dalam 14 hari terakhir. Di sisi lain, pdp_views_14d yang mengukur tampilan halaman produk hanya memiliki 8355 atau 3.76% data yang hilang. Variabel avg_time_per_session_14d memiliki nilai yang hilang sebanyak 872. time_spent_platform_14d dan total_login_sessions_l14d yang mengukur rata-rata waktu per sesi dan total waktu yang dihabiskan di platform selama 14 hari, memiliki kurang dari 1% data yang hilang, dengan 'new_buyer_initiative' kehilangan 378 nilai. Dalam penelitian ini, semua nilai yang hilang diatasi dengan mengisinya dengan nilai konstan yaitu 0. Pendekatan ini dipilih untuk memberikan nilai yang netral dan tidak memberikan dampak tambahan terhadap analisis statistik atau pemodelan prediktif. Pilihan ini berdasarkan asumsi bahwa nilai nol tidak akan mengganggu distribusi dan pola data yang ada, memungkinkan analisis yang lebih objektif dan hasil yang tidak bias oleh nilai-nilai yang mungkin berlebihan atau tidak representatif.


Submitted: **05/12/2023**; Accepted: **30/01/2024**; Published: **31/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 612-623
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7074

**Tabel 1.** Missing Values Dataset

| No | Column | Total Null Values | Percentage |
|---|---|---|---|
| 1. | top_up_14d | 204966 | 92.170089 |
| 2. | total_voucher_claim_14d | 97907 | 44.027287 |
| 3. | shop_flash_sale | 67577 | 30.388348 |
| 4. | shop_normal_shop | 67577 | 30.388348 |
| 5. | shop_sbs | 67577 | 30.388348 |
| 6. | shop_cb | 67577 | 30.388348 |
| 7. | shop_ss | 67577 | 30.388348 |
| 8. | shop_ss_plus | 67577 | 30.388348 |
| 9. | shop_mall | 67577 | 30.388348 |
| 10. | use_hemat | 67577 | 30.388348 |
| 11. | use_regular | 67577 | 30.388348 |
| 12. | use_nextday | 67577 | 30.388348 |
| 13. | use_instant | 67577 | 30.388348 |
| 14. | use_cc_debit | 67577 | 30.388348 |
| 15. | use_va_bt | 67577 | 30.388348 |
| 16. | use_cod | 67577 | 30.388348 |
| 17. | use_shopeepaylater | 67577 | 30.388348 |
| 18. | use_shopeepay | 67577 | 30.388348 |
| 19. | total_order_14d | 67577 | 30.388348 |
| 20. | gmv_14d | 67577 | 30.388348 |
| 21. | use_sameday | 67577 | 30.388348 |
| 22. | shop_views_14d | 60857 | 27.366466 |
| 23. | pdp_views_14d | 8355 | 3.757116 |
| 24. | avg_time_per_session_14d | 872 | 0.392125 |
| 25. | time_spent_platform_14d | 790 | 0.355251 |
| 26. | total_login_sessions_l14d | 790 | 0.355251 |
| 27. | new_buyer_initiative | 378 | 0.169981 |

### 2.2.2 Handling Duplicated Data

Proses identifikasi dan pengelolaan data terduplikasi melibatkan penemuan dan penanganan baris data yang esensialnya berisi informasi identik. Duplikasi ini sering terjadi akibat kesalahan entri data atau saat penggabungan beberapa dataset. Untuk menjaga kualitas analisis, penanganan data terduplikasi umumnya meliputi penghapusan baris duplikat agar tidak menyebabkan bias atau redundansi informasi. Berdasarkan analisis yang dilakukan dalam penelitian ini, tidak ditemukan adanya kasus data terduplikasi dalam dataset.

### 2.2.3 Handling Outliers

Proses identifikasi dan pengelolaan nilai-nilai ekstrem dalam dataset merupakan langkah krusial untuk menghindari distorsi dalam hasil analisis. Outliers, atau nilai yang signifikan berbeda dari mayoritas data, seringkali muncul karena variasi acak atau kesalahan pengukuran. Dalam penelitian ini, outliers diidentifikasi dan dielola melalui metode Winsorization, sebuah teknik statistik yang menggantikan nilai ekstrem dalam dataset. Metode ini menggantikan semua observasi di atas persentil ke-90 dan di bawah persentil ke-10 dengan nilai pada persentil tersebut. Hal ini dilakukan untuk memastikan bahwa outliers tidak memberikan pengaruh berlebihan yang dapat mengganggu analisis serta pemodelan data, menjadikan hasil yang diperoleh lebih akurat dan representatif. Jumlah outliers dalam dataset ditampilkan pada Tabel 2. Grafik yang menggambarkan distribusi data setelah Winsorization terdapat pada Gambar 5. Grafik ini menunjukkan titik-titik yang lebih terintegrasi dengan kumpulan data utama, mengindikasikan pengelolaan outliers yang efektif. Kolom-kolom seperti total_voucher_claim_14d, total_order_14d, gmv_14d, dan lainnya menunjukkan variasi nilai yang lebih terkontrol setelah proses Winsorization. Sebelum penanganan, dataset mengandung total 222,378 baris. Setelah penerapan Winsorization, yang menggantikan nilai-nilai ekstrem tanpa menghapus baris data, jumlah baris tetap 222,378. Ini menunjukkan bahwa tidak ada baris data yang dieliminasi, namun nilai-nilai ekstrem telah diubah untuk meningkatkan akurasi model prediktif dengan mengurangi potensi bias yang disebabkan oleh data ekstrem atau tidak biasa yang tidak merepresentasikan tren umum dalam dataset.

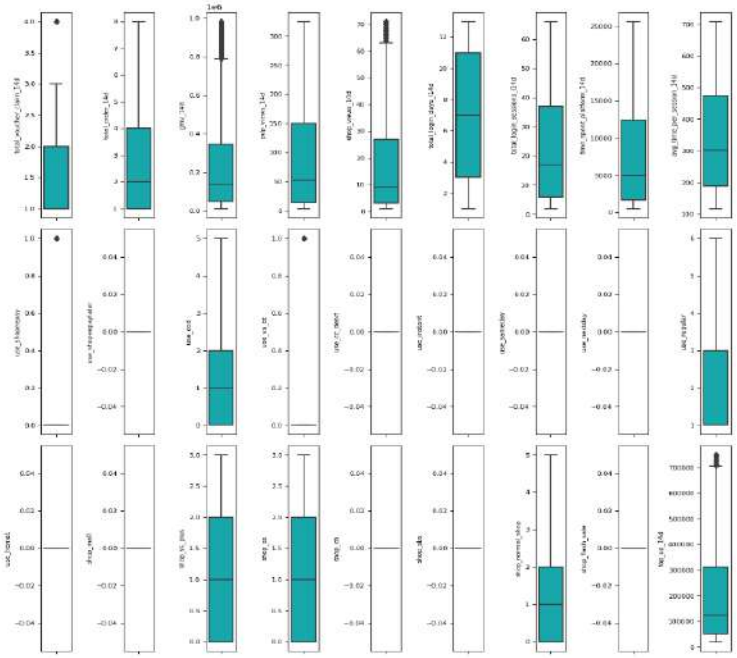
**Tabel 2.** Outliers Dataset

| No | Column Name | is Outlier | Outlier | No Outlier |
|---|---|---|---|---|
| 1. | total_voucher_claim_14d | True | 14983 | 207395 |
| 2. | total_order_14d | True | 13938 | 208440 |


Submitted: **05/12/2023**; Accepted: **30/01/2024**; Published: **31/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 612-623
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7074

| No | Column Name | is Outlier | Outlier | No Outlier |
|---|---|---|---|---|
| 3. | gmv_14d | True | 18942 | 203436 |
| 4. | pdp_views_14d | True | 18845 | 203533 |
| 5. | shop_views_14d | True | 18190 | 204188 |
| 6. | total_login_days_l14d | False | 0 | 222378 |
| 7. | total_login_sessions_l14d | True | 13850 | 208528 |
| 8. | time_spent_platform_14d | True | 18662 | 203716 |
| 9. | avg_time_per_session_14d | True | 11479 | 210899 |
| 10. | use_shopeepay | True | 20902 | 201476 |
| 11. | use_shopeepaylater | False | 0 | 222378 |
| 12. | use_cod | True | 15455 | 206923 |
| 13. | use_va_bt | True | 19782 | 202596 |
| 14. | use_cc_debit | True | 341 | 222037 |
| 15. | use_instant | True | 496 | 221882 |
| 16. | use_sameday | True | 705 | 221673 |
| 17. | use_nextday | True | 83 | 222295 |
| 18. | use_regular | True | 14409 | 207969 |
| 19. | use_hemat | True | 9535 | 212843 |
| 20. | shop_mall | True | 13875 | 208503 |
| 21. | shop_ss_plus | True | 6705 | 215673 |
| 22. | shop_ss | True | 6705 | 215673 |
| 23. | shop_cb | True | 7089 | 215289 |
| 24. | shop_sbs | True | 1867 | 220511 |
| 25. | shop_normal_shop | True | 12708 | 209670 |
| 26. | shop_flash_sale | True | 3507 | 218871 |



**Gambar 3.** Hasil Penanganan Outliers Dataset

### 2.2.4 Feature Encoding

Teknik ini mengubah data kategorikal menjadi format numerik, memungkinkan algoritma pembelajaran mesin mengolahnya lebih efektif. Biasanya, dataset mengandung variabel kategorikal seperti nama, label, atau kode, yang harus diubah ke format yang sesuai untuk model berbasis kalkulasi matematis. Dalam penelitian ini, menggunakan One-Hot Encoding untuk mengonversi data kategorikal menjadi format numerik. Metode ini khususnya diterapkan pada kolom seperti 'gender', 'age_group', 'region', 'area', dan 'new_buyer_initiative'. One-Hot Encoding menciptakan kolom baru untuk setiap kategori unik dalam variabel tersebut, dengan nilai biner 0 dan 1 yang menandakan kategori tertentu dalam sebuah baris. Kolom asli dihapus dari dataset, dan kolom baru yang dihasilkan dari One-Hot Encoding diintegrasikan ke dalamnya. Dengan demikian, dataset telah diubah menjadi format yang sepenuhnya numerik, yang memudahkan proses pembelajaran algoritma dan meningkatkan kualitas analisis yang akan dilakukan. Dataset yang telah di-encode akan digunakan dalam model pembelajaran mesin untuk proses pelatihan dan evaluasi. Gambar 6 merupakan hasil feature encoding pada dataset.


Submitted: **05/12/2023**; Accepted: **30/01/2024**; Published: **31/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 612-623
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7074

| gender_Female | gender_Male | gender_Not_filled | age_group_19-24 | age_group_25-30 | age_group_30-35 | age_group_Above_35 | age_group_Not_filled | age_group_Under_19 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

**Gambar 4.** Hasil Feature Encoding

### 2.2.5 Feature Selection

Proses Feature Selection melibatkan pemilihan subset fitur yang paling relevan dari dataset untuk digunakan dalam pembuatan model. Tujuannya adalah meningkatkan efisiensi komputasi, mengurangi kompleksitas model, dan memperbaiki kinerja model dengan mengeliminasi fitur-fitur yang tidak esensial atau redundan. Dalam machine learning, Feature Selection penting untuk mencegah overfitting, di mana model menjadi terlalu spesifik untuk data latih dan kurang efektif dalam memprediksi data baru. Proses ini biasanya menggunakan teknik berbasis statistik, model, atau metode iteratif. Dalam penelitian ini, Feature Selection diterapkan untuk mengidentifikasi dan mempertahankan fitur-fitur yang berkontribusi signifikan terhadap model. Pendekatan ini penting untuk memastikan model memiliki kemampuan generalisasi yang baik pada data baru, yang krusial untuk akurasi prediksi. Dalam penelitian ini, fitur-fitur seperti 'user_id', 'regist_date', 'use_shopeepaylater', 'top_up_14d', 'use_cc_debit', 'use_instant', 'use_sameday', 'use_nextday', 'use_hemat', 'shop_mall', 'shop_cb','shop_sbs', 'shop_flash_sale' dieliminasi karena tidak berkorelasi atau tidak memberikan kontribusi signifikan terhadap hasil model.

### 2.2.6 Feature Extraction

Proses kunci dalam pengurangan dimensi data yang bertujuan untuk mengidentifikasi dan memilih fitur yang paling relevan dan informatif dari sebuah dataset. Proses ini dirancang untuk mempermudah analisis dan pemodelan dengan mengurangi kompleksitas data, namun tetap mempertahankan informasi penting. Dalam konteks Feature Extraction, data yang semula besar dan kompleks disederhanakan menjadi format yang lebih ringkas dan efisien. Dalam penelitian ini, Feature Extraction diimplementasikan untuk mengekstraksi elemen informasi utama dari berbagai variabel. Hal ini memudahkan pemrosesan data lebih lanjut dan meningkatkan efisiensi komputasi dalam pemodelan machine learning. Pendekatan ini sangat penting dalam mengurangi dimensi data tanpa mengorbankan integritas informasi yang diperlukan untuk analisis yang akurat. Penelitian ini mengembangkan fitur baru 'total_shopping_activity', yang merupakan kombinasi dari pdp_views_14d dan shop_views_14d. Tujuan pembuatan fitur adalah untuk menyediakan variabel yang lebih efektif dan efisien dalam menganalisis perilaku pengguna. Dengan menggabungkan dua aspek interaksi pengguna dengan platform dapat meningkatkan pemahaman dan prediksi terhadap perilaku pembelian dan retensi.

### 2.2.7 Feature Transformation

Proses Feature Transformation dalam penelitian ini melibatkan modifikasi atau transformasi fitur tertentu dalam dataset untuk meningkatkan kualitas dan efektivitasnya dalam pemodelan. Tujuan utamanya adalah untuk menyusun ulang data ke dalam format yang lebih homogen dan normal, sering melalui teknik standarisasi. Transformasi ini esensial dalam machine learning untuk memastikan semua fitur diproses dalam skala yang seragam, memungkinkan model untuk belajar dan menginterpretasikan pola data secara efisien. Dalam penelitian ini, standarisasi diterapkan pada fitur 'total_voucher_claim_14d', 'total_order_14d', 'gmv_14d', 'pdp_views_14d', 'shop_views_14d', 'total_login_sessions_l14d', 'time_spent_platform_14d', 'avg_time_per_session_14d', 'use_shopeepay', 'use_cod', 'use_va_bt', 'use_regular', 'shop_ss_plus', 'shop_ss', 'shop_normal_shop', 'total_shopping_activity' menggunakan StandardScaler. StandardScaler mengubah data sehingga memiliki rata-rata nol dan deviasi standar satu, membantu dalam mengurangi bias yang mungkin disebabkan oleh fitur dengan skala besar dan memastikan setiap fitur memberikan kontribusi yang seimbang ke dalam model. Hal ini penting untuk meningkatkan akurasi dan konsistensi prediksi model, khususnya dalam algoritma yang sensitif terhadap skala fitur. Pendekatan ini membantu model mengenali dan memanfaatkan pola penting dalam data secara lebih efektif.

### 2.2.8 Class Imbalance

Class Imbalance adalah masalah umum dalam pengolahan data untuk machine learning, terutama dalam tugas klasifikasi. Terjadi ketika jumlah sampel untuk setiap kelas dalam dataset tidak seimbang, biasanya dengan satu kelas yang jauh lebih banyak dibandingkan kelas lainnya [16]. Masalah ini dapat mengakibatkan model menjadi bias, di mana cenderung memprediksi kelas mayoritas dan mengabaikan kelas minoritas [17]. Dalam konteks analisis data, Class Imbalance dapat menyebabkan hasil yang tidak akurat dan mengurangi kemampuan model untuk mempelajari karakteristik kelas minoritas. Dalam penelitian ini, teknik SMOTE (Synthetic Minority Over-sampling Technique) digunakan untuk menyeimbangkan proporsi kelas dengan menciptakan sampel yang memungkinkan distribusi kelas menjadi lebih merata [18]. Pendekatan ini memfasilitasi model untuk belajar secara


Submitted: **05/12/2023**; Accepted: **30/01/2024**; Published: **31/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 612-623
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7074

lebih komprehensif dari kedua kelas, yang pada gilirannya meningkatkan keakuratan prediksi model pada data baru dan menjamin evaluasi yang lebih objektif terhadap performa model. Dataset yang dianalisis menunjukkan adanya ketidakseimbangan pada kelas 'retained', dengan 77,500 sampel untuk kelas tidak menggunakan shopee (retained = 0) dan 54,319 sampel untuk kelas yang masih menggunakan shopee (retained = 1) sebelum penerapan SMOTE. Dengan implementasi SMOTE, kedua kelas tersebut sekarang memiliki jumlah sampel yang sama, yaitu 77,500, yang mencerminkan distribusi kelas yang lebih seimbang. Hal ini tergambar pada tabel 2, yang menampilkan kesetaraan jumlah sampel antara kelas yang masih menggunakan shopee dan tidak menggunakan shopee, memastikan bahwa kedua kelas tersebut memiliki representasi yang sama dalam model pembelajaran mesin. Dengan demikian, SMOTE membantu dalam memperbaiki bias yang mungkin ditimbulkan oleh Class Imbalance dan memungkinkan model untuk mempelajari karakteristik dari kedua kelas secara lebih efektif. Ini penting untuk meningkatkan akurasi prediksi model pada data yang belum dilihat sebelumnya dan menghasilkan evaluasi yang lebih adil dan representatif terhadap performa model.

**Tabel 3.** Class Imbalance Dataset

| | Distribusi Dataset sebelum SMOTE | Distribusi Dataset setelah SMOTE |
|---|---|---|
| Not Retained | 54319 | 77500 |
| Retained | 77500 | 77500 |

### 2.3 Modeling

Proses pembuatan dan pelatihan model prediktif dalam penelitian ini menggunakan berbagai algoritma machine learning. Data yang telah diolah dan disiapkan melalui tahapan preprocessing kemudian digunakan untuk mengembangkan model yang mampu memprediksi atau mengklasifikasikan data baru, berdasarkan pola dan hubungan yang ditemukan dalam data latih. Modeling dalam penelitian ini melibatkan penerapan berbagai algoritma klasifikasi yaitu Logistic Regression, Decision Tree, Gaussian Naive Bayes, Random Forest, KNN, MLP, Adaboost, dan XGBoost [19] [20] [21]. Setiap model diajarkan untuk mengenali pola dalam data dan membuat prediksi berdasarkan pola tersebut. Proses modeling juga mencakup validasi model untuk menguji kinerjanya pada data tes yang belum pernah dilihat sebelumnya, serta tuning hyperparameter guna meningkatkan akurasi dan efektivitas prediksi.

### 2.4 Evaluasi Model

Evaluasi adalah proses untuk menilai kinerja dan keefektifan model klasifikasi yang telah dikembangkan [22]. Dalam penelitian ini, evaluasi dilakukan dengan menggunakan Confusion Matrix. Confusion Matrix mengkategorikan hasil prediksi ke dalam empat bagian utama: True Positive (TP) untuk prediksi positif yang benar, True Negative (TN) untuk prediksi negatif yang benar, False Positive (FP) untuk prediksi positif yang salah, dan False Negative (FN) untuk prediksi negatif yang salah [23]. Kategorisasi ini membantu memahami performa model dalam membedakan antara kelas yang berbeda. Metrik evaluasi lain juga diterapkan untuk mengukur keakuratan dan efektivitas model, yaitu Accuracy mengukur frekuensi model dalam membuat prediksi yang tepat, dihitung dengan rumus persamaan (1). Precision mengukur proporsi prediksi positif yang tepat, penting dalam kasus di mana biaya False Positive tinggi, dihitung dengan rumus persamaan (2). Recall mengukur kemampuan model dalam mengidentifikasi semua kasus positif sebenarnya, dihitung dengan rumus persamaan (3). F1-Score mengukur rata-rata harmonis dari presisi dan recall, memberikan keseimbangan antara kedua metrik ini, dihitung dengan rumus persamaan (4). Penerapan metrik-metrik ini memberikan penilaian yang menyeluruh mengenai keakuratan prediksi model serta kemampuannya dalam membedakan antar kelas yang berbeda. Hal ini penting untuk memastikan efektivitas model dalam kondisi data yang beragam dan kompleks.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1 Score} = 2\ X\ \frac{\text{Precision x Recall}}{\text{Precision+Recall}} \quad (4)$$

## 3. HASIL DAN PEMBAHASAN

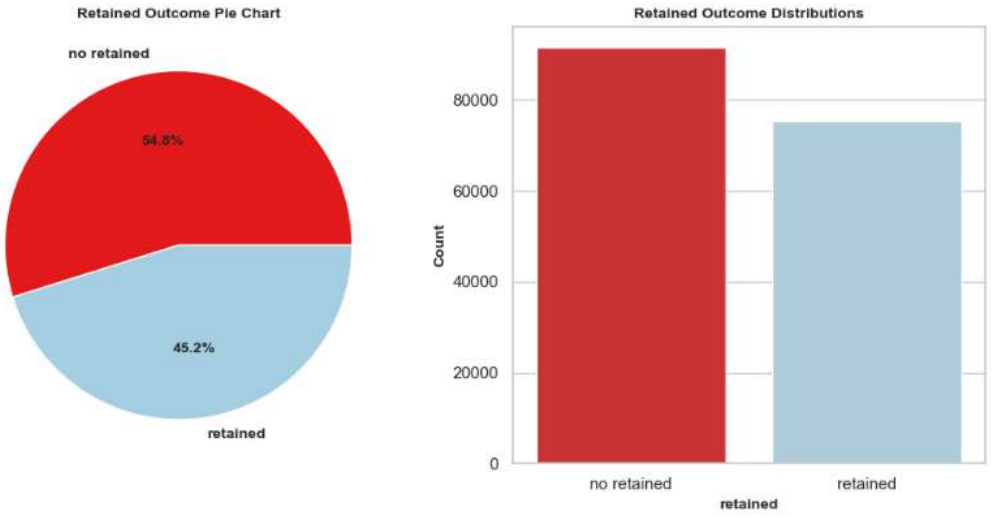
### 3.1 Analisis Deskriptif

Penelitian ini menerapkan analisis deskriptif untuk memahami karakteristik 222,378 data pengguna Shopee. Dalam data tersebut, variabel 'retained' dengan nilai 1 menandakan pengguna yang terus menggunakan Shopee, sedangkan nilai 0 menunjukkan pengguna yang berhenti menggunakan layanan tersebut. Gambar 4 menampilkan pie chart yang menggambarkan persentase pengguna Shopee yang bertahan (retained) dan yang tidak (no retained),


Submitted: **05/12/2023**; Accepted: **30/01/2024**; Published: **31/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 612-623
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7074

dengan 54.8% tidak lagi menggunakan shopee dan 45.2% tetap menggunakan Shopee. Adanya perbedaan signifikan antara kedua kelompok menandakan masalah kualitas layanan dari Shopee, dengan lebih banyak pengguna yang berhenti menggunakan layanan daripada yang bertahan. Penelitian ini bertujuan untuk mengidentifikasi faktor-faktor yang berpengaruh terhadap keputusan pengguna untuk tetap menggunakan layanan agar tidak menambah jumlah pengguna yang berhenti. Untuk itu, akan dianalisis faktor-faktor yang dapat menyebabkan pengguna berhenti menggunakan layanan Shopee.




**Gambar 5.** Pie Chart dan Barplot Customer Retained Shopee

### 3.2 Hasil Model Evaluasi Dataset

Dalam tahap pemodelan, penelitian ini menerapkan beragam algoritma machine learning untuk mengembangkan model prediktif. Kinerja setiap model dinilai berdasarkan kemampuannya dalam mengklasifikasikan data baru, menggunakan metrik akurasi, presisi, recall, dan F1-Score pada data pelatihan dan pengujian. Selain itu, evaluasi juga melibatkan nilai cross-validation (CV) untuk mengukur kestabilan dan konsistensi performa model di berbagai subset data. Hasil yang diperoleh menampilkan variasi kinerja yang signifikan di antara model-model klasifikasi yang diuji yang dapat di lihat pada tabel 3. Logistic Regression menunjukkan kinerja yang stabil dengan akurasi 0.727548, presisi 0.705894, recall 0.680614, dan F1 Score 0.693023. Skor CV akurasi 0.727105, presisi 0.705015, recall 0.681026, dan F1 Score 0.692805. Decision Tree akurasi 0.738160, presisi 0.707348, recall 0.717294, dan F1 Score 0.712286. Skor CV akurasi 0.724342, presisi 0.692051, recall 0.702679, dan F1 Score 0.697277. Naive Bayes akurasi 0.542174, presisi 0.496612, recall 0.967898, dan F1 Score 0.656424. Skor CV akurasi 0.541744, presisi 0.496376, recall 0.967102, dan F1 Score 0.656030. Random Forest akurasi 0.738507, presisi 0.733445, recall 0.661819, dan F1 Score 0.695793. Skor CV akurasi 0.730631, presisi 0.722731, recall 0.655336, dan F1 Score 0.687364. K-Nearest Neighbors (KNN) akurasi 0.792682, presisi 0.782029, recall 0.750320, dan F1 Score 0.765846. Skor CV akurasi 0.690314, presisi 0.665551, recall 0.632503, dan F1 Score 0.648597. MLPClassifier menunjukkan konsistensi yang baik dengan akurasi 0.751876, presisi 0.707917, recall 0.767579, dan F1 Score 0.736542. Skor CV akurasi 0.731691, presisi 0.704095, recall 0.701471, dan F1 Score 0.702563. AdaBoost akurasi 0.728749, presisi 0.709132, recall 0.677656, dan F1 Score 0.693037. Skor CV akurasi 0.728216, presisi 0.708322, recall 0.677543, dan F1 Score 0.692582. XGBoost menunjukkan kinerja yang kuat dengan akurasi pelatihan 0.790870, presisi 0.776223, recall 0.754770, dan F1 Score 0.765346. Skor CV akurasi 0.734730, presisi 0.710338, recall 0.697290, dan F1 Score 0.703745. Metrik Cross-Validation menunjukan konsistensi model di berbagai subset data. Beberapa model algoritma machine learning menunjukkan konsistensi yang baik untuk memastikan model yang dikembangkan tidak hanya berperforma baik pada data pelatihan, tetapi juga mampu menggeneralisasi dengan baik pada data yang belum dilihat sebelumnya, yang merupakan indikator penting dari model.

**Tabel 4.** Hasil Model Evaluasi Dataset

| Model | Accuracy | Precision | Recall | F1 Score | CV | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Accuracy | Precision | Recall | F1 Score |
| Logistic Regression | 0.727548 | 0.705894 | 0.680614 | 0.693023 | 0.727105 | 0.705015 | 0.681026 | 0.692805 |
| Decision Tree | 0.738160 | 0.707348 | 0.717294 | 0.712286 | 0.724342 | 0.692051 | 0.702679 | 0.697277 |
| Naive Bayes | 0.542174 | 0.496612 | 0.967898 | 0.656424 | 0.541744 | 0.496376 | 0.967102 | 0.656030 |
| Random Forest | 0.738507 | 0.733445 | 0.661819 | 0.695793 | 0.730631 | 0.722731 | 0.655336 | 0.687364 |
| KNNeighbors | 0.792682 | 0.782029 | 0.750320 | 0.765846 | 0.690314 | 0.665551 | 0.632503 | 0.648597 |


Submitted: **05/12/2023**; Accepted: **30/01/2024**; Published: **31/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 612-623
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7074

| Model | Accuracy | Precision | Recall | F1 Score | CV | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Accuracy | Precision | Recall | F1 Score |
| MLPClassifier | 0.751876 | 0.707917 | 0.767579 | 0.736542 | 0.731691 | 0.704095 | 0.701471 | 0.702563 |
| AdaBoost | 0.728749 | 0.709132 | 0.677656 | 0.693037 | 0.728216 | 0.708322 | 0.677543 | 0.692582 |
| XGBoost | 0.790870 | 0.776223 | 0.754770 | 0.765346 | 0.734730 | 0.710338 | 0.697290 | 0.703745 |

### 3.3 Hasil Model Evaluasi Dataset Menggunakan SMOTE

Evaluasi model dengan menerapkan teknik SMOTE telah berhasil mengatasi masalah ketidakseimbangan kelas dalam dataset penelitian ini. Hasil yang diperoleh menunjukkan perubahan yang signifikan dalam kinerja model setelah penerapan SMOTE, yang dapat dilihat pada tabel 4. Model Logistic Regression menunjukkan peningkatan kinerja dengan akurasi 0.731262, presisi 0.722329, recall 0.728371, serta F1-Score 0.725337. Cross-Validation menunjukkan konsistensi yang baik, dengan akurasi 0.728275, presisi 0.720768, recall 0.721894, dan F1-Score 0.721249. Decision Tree akurasi 0.736827, presisi 0.710835, serta recall 0.775114, dan F1-Score 0.741584. Cross-Validation menunjukkan akurasi 0.725883, presisi 0.705885, recall 0.750046, dan F1-Score 0.727224. Naive Bayes menunjukkan akurasi 0.568230, presisi 0.531256, serta recall 0.966555, dan F1-Score 0.685651. Cross-Validation menunjukkan akurasi 0.567569, presisi 0.530855, recall 0.966333, dan F1-Score 0.685255. Random Forest memiliki akurasi 0.740626, presisi 0.731310, recall 0.739178, serta F1-Score 0.735223. Cross-Validation menunjukkan akurasi 0.731136, presisi 0.721340, recall 0.730332, dan F1-Score 0.725782. K-Nearest Neighbors (KNN) menunjukkan peningkatan dengan akurasi 0.806707, presisi 0.790492, recall 0.820773, serta F1-Score 0.805348. Cross-Validation menunjukkan akurasi 0.706344, presisi 0.690795, recall 0.717712, dan F1-Score 0.703634. MLPClassifier menunjukkan akurasi 0.757695, presisi 0.738199, recall 0.778852, serta F1-Score 0.757981. Cross-Validation menunjukkan akurasi 0.737410, presisi 0.729222, recall 0.734231, dan F1-Score 0.730962. AdaBoost menunjukkan akurasi 0.732404, presisi 0.724883, recall 0.726421, serta F1-Score 0.725652. Cross-Validation menunjukkan akurasi 0.730649, presisi 0.723382, recall 0.723966, dan F1-Score 0.723625. XGBoost menunjukkan akurasi 0.793076, presisi 0.780006, recall 0.801244, serta F1-Score 0.790482. Cross-Validation menunjukkan akurasi 0.741792, presisi 0.731078, recall 0.743286, dan F1-Score 0.736888. Setelah penerapan SMOTE, menandakan efektivitasnya dalam mengatasi ketidakseimbangan kelas dan kemampuannya dalam mengidentifikasi kasus positif. Penerapan SMOTE telah meningkatkan kinerja model dalam mengklasifikasikan data yang lebih seimbang. Hal ini terlihat dari peningkatan akurasi, presisi, recall, dan F1-Score di sebagian besar model. Metrik Cross-Validation juga memberikan gambaran yang baik tentang konsistensi dan keandalan model dalam kondisi data yang telah diimbangi melalui SMOTE. Ini menunjukkan pentingnya teknik pengimbangan kelas dalam meningkatkan efektivitas model klasifikasi, terutama dalam konteks data yang memiliki distribusi kelas yang tidak seimbang.

**Tabel 5.** Hasil Model Evaluasi Dataset Menggunakan SMOTE

| Model | Accuracy (SMOTE) | Precision (SMOTE) | Recall (SMOTE) | F1 (SMOTE) | CV | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Accuracy (SMOTE) | Precision (SMOTE) | Recall (SMOTE) | F1 (SMOTE) |
| Logistic Regression | 0.731262 | 0.722329 | 0.728371 | 0.725337 | 0.728275 | 0.720768 | 0.721894 | 0.721249 |
| Decision Tree | 0.736827 | 0.710835 | 0.775114 | 0.741584 | 0.725883 | 0.705885 | 0.750046 | 0.727224 |
| Naive Bayes | 0.568230 | 0.531256 | 0.966555 | 0.685651 | 0.567569 | 0.530855 | 0.966333 | 0.685255 |
| Random Forest | 0.740626 | 0.731310 | 0.739178 | 0.735223 | 0.731136 | 0.721340 | 0.730332 | 0.725782 |
| KNNeighbors | 0.806707 | 0.790492 | 0.820773 | 0.805348 | 0.706344 | 0.690795 | 0.717712 | 0.703634 |
| MLPClassifier | 0.757695 | 0.738199 | 0.778852 | 0.757981 | 0.737410 | 0.729222 | 0.734231 | 0.730962 |
| AdaBoost | 0.732404 | 0.724883 | 0.726421 | 0.725652 | 0.730649 | 0.723382 | 0.723966 | 0.723625 |
| XGBoost | 0.793076 | 0.780006 | 0.801244 | 0.790482 | 0.741792 | 0.731078 | 0.743286 | 0.736888 |

### 3.4 Hasil Hyperparameter Tuning

Proses tuning hyperparameter dilakukan untuk mengoptimalkan model dengan menggunakan teknik RandomizedSearchCV. Setiap model disesuaikan dengan rangkaian hyperparameter yang spesifik, yang ditentukan dalam grid_parameters. Proses tuning ini mencakup pencarian kombinasi hyperparameter yang optimal berdasarkan skor recall. Hasil tuning hyperparameter menunjukkan peningkatan signifikan dalam kinerja model yang dapat dilihat pada tabel 5. Model Logistic Regression menunjukan akurasi 0.733498, presisi 0.725288, recall 0.729136, serta F1 Score 0.727207. Cross-Validation akurasi 0.731460, presisi 0.723884, recall 0.725262, dan F1 Score 0.724374. Decision Tree menunjukan akurasi 0.730018, presisi 0.705933, recall 0.764135, serta nilai F1 Score 0.733882. Nilai Cross-Validation akurasi 0.724386, presisi 0.704384, recall 0.748628, dan F1 Score 0.725766. Naive Bayes dengan akurasi 0.569306, presisi 0.531945, recall 0.965309, serta F1 Score 0.685911. Cross-Validation akurasi 0.569751, presisi 0.532236, recall 0.964520, dan F1 Score 0.685952. Random Forest


Submitted: **05/12/2023**; Accepted: **30/01/2024**; Published: **31/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 612-623
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7074

menunjukan akurasi 0.839445, presisi 0.829968, recall 0.843176, serta F1 Score 0.836520. Cross-Validation akurasi 0.740500, presisi 0.725775, recall 0.751169, dan F1 Score 0.738148. K-Nearest Neighbors (KNN) menunjukan akurasi 0.726935, presisi 0.718829, recall 0.721845, serta F1 Score 0.720334. Cross-Validation menunjukkan akurasi 0.714163, presisi 0.706287, recall 0.707571, dan F1 Score 0.706910. MLPClassifier menunjukan akurasi 0.746089, presisi 0.743720, recall 0.730554, serta F1 Score 0.737079. Cross-Validation menunjukkan akurasi 0.739514, presisi 0.720401, recall 0.762594, dan F1 Score 0.740210. AdaBoost menunjukan akurasi 0.732404, presisi 0.724883, recall 0.726421, serta F1 Score 0.725652. Cross-Validation menunjukkan akurasi 0.730649, presisi 0.723382, recall 0.723966, dan F1 Score 0.723625. XGBoost menunjukan akurasi 0.738456, presisi 0.716795, recall 0.765652, serta F1 Score 0.740419. Cross-Validation menunjukkan akurasi 0.733750, presisi 0.714195, recall 0.756054, dan F1 Score 0.734514. Hasil tuning hyperparameter menunjukkan bahwa setiap model memiliki kemampuan untuk memprediksi dengan akurasi yang baik, mengindikasikan keberhasilan dalam optimasi model. Recall yang tinggi pada setiap model menunjukkan efektivitas model dalam mengidentifikasi kasus positif, yang sangat penting dalam konteks dataset dengan ketidakseimbangan kelas. Konsistensi hasil cross-validation juga menandakan bahwa model yang dioptimalkan mampu menggeneralisasi dengan baik pada data yang beragam. Model Random Forest yang telah dioptimalkan menunjukkan kinerja sangat baik dalam mendeteksi kasus positif, dengan nilai recall yang tinggi dalam mengidentifikasi pengguna yang terus menggunakan Shopee. Model ini menunjukkan sensitivitas yang tinggi terhadap kasus positif, kinerja yang konsisten, dan kemampuan menggeneralisasi yang baik, terlihat dari akurasi dan hasil cross-validation yang stabil. Keunggulannya dalam mengurangi overfitting menjadikannya pilihan efektif di antara model yang diuji dalam penelitian ini.

**Tabel 6.** Hasil Hyperparameter Tuning

| Model | Accuracy | Precision | Recall | F1 | CV | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Accuracy | Precision | Recall | F1 |
| Logistic Regression | 0.733498 | 0.725288 | 0.729136 | 0.727207 | 0.731460 | 0.723884 | 0.725262 | 0.724374 |
| Decision Tree | 0.730018 | 0.705933 | 0.764135 | 0.733882 | 0.724386 | 0.704384 | 0.748628 | 0.725766 |
| Naive Bayes | 0.569306 | 0.531945 | 0.965309 | 0.685911 | 0.569751 | 0.532236 | 0.964520 | 0.685952 |
| Random Forest | 0.839445 | 0.829968 | 0.843176 | 0.836520 | 0.740500 | 0.725775 | 0.751169 | 0.738148 |
| KNNeighbors | 0.726935 | 0.718829 | 0.721845 | 0.720334 | 0.714163 | 0.706287 | 0.707571 | 0.706910 |
| MLPClassifier | 0.746089 | 0.743720 | 0.730554 | 0.737079 | 0.739514 | 0.720401 | 0.762594 | 0.740210 |
| AdaBoost | 0.732404 | 0.724883 | 0.726421 | 0.725652 | 0.730649 | 0.723382 | 0.723966 | 0.723625 |
| XGBoost | 0.738456 | 0.716795 | 0.765652 | 0.740419 | 0.733750 | 0.714195 | 0.756054 | 0.734514 |

## 4. KESIMPULAN

Dengan melakukan analisis deskriptif, hasil data pengguna Shopee menunjukkan perbandingan antara pengguna yang bertahan dan yang tidak. Ditemukan bahwa lebih banyak pengguna yang berhenti menggunakan Shopee (54.8%) daripada yang tetap menggunakan (45.2%). Hal ini menunjukkan adanya masalah dalam kualitas layanan Shopee, yang menjadi fokus penelitian untuk mengidentifikasi faktor-faktor yang memengaruhi keputusan pengguna. Penerapan berbagai model machine learning menunjukkan variasi kinerja yang baik. Proses evaluasi model dengan dan tanpa penerapan teknik SMOTE menggambarkan perubahan signifikan. Penerapan SMOTE berhasil mengatasi ketidakseimbangan kelas, meningkatkan kinerja setiap model. Meskipun beberapa model mengalami penurunan performa di Cross-Validation setelah penerapan SMOTE, kebanyakan model menunjukkan peningkatan konsistensi dan keandalan dalam mengklasifikasikan data yang seimbang. Proses hyperparameter tuning yang dilakukan pada model Logistic Regression, Decision Tree, Naive Bayes, Random Forest, K-Nearest Neighbors (KNN), MLPClassifier, AdaBoost, dan XGBoost menghasilkan model yang dioptimalkan dengan akurasi yang baik. Model Random Forest memiliki kinerja terbaik dengan nilai Accuracy 0.839445, Precision 0.829968, Recall 0.843176, dan F1-Score 0.836520. Nilai cross-validation menunjukkan konsistensi model dengan Accuracy 0.740500, Precision 0.725775, Recall 0.751169, dan F1-Score 0.738148. Model Random Forest menjadi model dengan nilai recall tinggi, menunjukkan sensitivitas yang baik dalam mengidentifikasi pengguna yang bertahan. Keunggulan model ini terletak pada kemampuannya mengatasi overfitting dan konsistensi performa yang tinggi. Keterbatasan penelitian ini melibatkan ketidakpastian interpretasi model dan ketergantungan pada data historis. Perbaikan dapat dilakukan dengan memperluas dataset dan mempertimbangkan faktor-faktor eksternal. Kesimpulan yang dapat diambil dari penelitian ini memberikan wawasan mendalam tentang faktor-faktor yang memengaruhi retensi pengguna Shopee yang baru. Hasilnya dapat menjadi landasan untuk pengembangan strategi yang lebih efektif dalam mempertahankan pengguna baru Shopee dengan menyesuaikan layanan berdasarkan temuan dari model machine learning dan analisis deskriptif.

## REFERENCES

[1] Databoks, “Pertumbuhan Pengunjung Shopee sampai Kuartal II 2022,” Databoks. [Online]. Available:


Submitted: **05/12/2023**; Accepted: **30/01/2024**; Published: **31/01/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 1, Januari 2024, Page 612-623
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i1.7074

https://databoks.katadata.co.id/datapublish/2022/11/21/tokopedia-masih-ungguli-shopee-sampai-kuartal-ii-2022
[2] B. Prabadevi, R. Shalini, and B. R. Kavitha, "Customer churning analysis using machine learning algorithms," Int. J. Intell. Networks, vol. 4, no. September 2022, pp. 145–154, 2023, doi: 10.1016/j.ijin.2023.05.005.
[3] A. N. Rachmi, "Implementasi Metode Random Forest Dan Xgboost Pada Klasifikasi Customer Churn," pp. 1–101, 2020, [Online]. Available: https://dspace.uii.ac.id/handle/123456789/30082
[4] N. Suryana, Pratiwi, and R. Tri Prasetio, "Penanganan Ketidakseimbangan Data pada Prediksi Customer Churn Menggunakan Kombinasi SMOTE dan Boosting," IJCIT (Indonesian J. Comput. Inf. Technol., vol. 6, no. 1, pp. 31–37, 2021, [Online]. Available: https://creativecommons.org/licenses/by-sa/4.0/
[5] P. P. Singh, F. I. Anik, R. Senapati, A. Sinha, N. Sakib, and E. Hossain, "Investigating customer churn in banking: A machine learning approach and visualization app for data science and management," Data Sci. Manag., 2023, doi: 10.1016/j.dsm.2023.09.002.
[6] X. Li and Z. Li, "A hybrid prediction model for e-commerce customer churn based on logistic regression and extreme gradient boosting algorithm," Ing. des Syst. d'Information, vol. 24, no. 5, pp. 525–530, 2019, doi: 10.18280/isi.240510.
[7] M. Kiguchi, W. Saeed, and I. Medi, "Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest," Appl. Soft Comput., vol. 118, p. 108491, 2022, doi: 10.1016/j.asoc.2022.108491.
[8] S. M. Shrestha and A. Shakya, "A Customer Churn Prediction Model using XGBoost for the Telecommunication Industry in Nepal," Procedia Comput. Sci., vol. 215, pp. 652–661, 2022, doi: 10.1016/j.procs.2022.12.067.
[9] S. Baghla and G. Gupta, "Performance Evaluation of Various Classification Techniques for Customer Churn Prediction in E-commerce," Microprocess. Microsyst., vol. 94, no. September, p. 104680, 2022, doi: 10.1016/j.micpro.2022.104680.
[10] J. Mantik and L. Qadrini, "Handling Unbalanced Data With Smote Adaboost," J. Mantik, vol. 6, no. 2, pp. 2332–2336, 2022.
[11] J. Pamina et al., "An effective classifier for predicting churn in telecommunication," J. Adv. Res. Dyn. Control Syst., vol. 11, no. 1 Special Issue, pp. 221–229, 2019.
[12] I. Hanif, "Implementing Extreme Gradient Boosting (XGBoost) Classifier to Improve Customer Churn Prediction," 2020, doi: 10.4108/eai.2-8-2019.2290338.
[13] I. M. Syahrani, "Comparation Analysis of Ensemble Technique With Boosting(Xgboost) and Bagging (Randomforest) For Classify Splice Junction DNA Sequence Category," J. Penelit. Pos dan Inform., vol. 9, no. 1, pp. 27–36, 2019, doi: 10.17933/jppi.v9i1.249.
[14] S. F. Sabbeh, "Machine-learning techniques for customer retention: A comparative study," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 2, pp. 273–281, 2018, doi: 10.14569/IJACSA.2018.090238.
[15] Kaggle, "Kaggle." [Online]. Available: https://www.kaggle.com/datasets/danielbeltsazar/shopee-new-user-behavior?select=sample_data_DStest.csv
[16] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res., vol. 30, no. 2, pp. 321–357, 2002.
[17] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," J. Teknol. dan Sist. Komput., vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.
[18] Y. A. Sir and A. H. H. Soepranoto, "Pendekatan Resampling Data Untuk Menangani Masalah Ketidakseimbangan Kelas," J. Komput. dan Inform., vol. 10, no. 1, pp. 31–38, 2022, doi: 10.35508/jicon.v10i1.6554.
[19] M. Libnao, M. Misula, C. Andres, J. Mariñas, and J. Mariñas, "ScienceDirect Procedia ScienceDirect Traffic incident prediction and classification system using naïve Traffic incident prediction and algorithm classification system using naïve bayes bayes algorithm," Procedia Comput. Sci., vol. 227, pp. 316–325, 2023, doi: 10.1016/j.procs.2023.10.530.
[20] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, "RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.
[21] J. G. C. Krüger, A. de S. Britto, and J. P. Barddal, "An explainable machine learning approach for student dropout prediction," Expert Syst. Appl., vol. 233, no. June, p. 120933, 2023, doi: 10.1016/j.eswa.2023.120933.
[22] D. Vora and K. Iyer, "Evaluating the Effectiveness of Machine Learning Algorithms in Predictive Modelling," Int. J. Eng. Technol., vol. 7, no. 3.4, p. 197, 2018, doi: 10.14419/ijet.v7i3.4.16773.
[23] S. K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, and S. H. Gawande, "Customer Churn Prediction in Telecom Sector using Machine Learning Techniques," Results Control Optim., vol. 14, no. October 2023, p. 100342, 2023, doi: 10.1016/j.rico.2023.100342.


Submitted: **05/12/2023**; Accepted: **30/01/2024**; Published: **31/01/2024**