Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 3, May 2024, Page 605-614
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i3.5215

Prediksi Perpindahan Pelanggan Pada Toko Online Menggunakan Metode Tree-Based Gradient Boosted Models

Selfia Hafidatus Sholeha*, Mochammad Faid, Moh. Ainol Yaqin

Fakultas Teknik, Informatika, Universitas Nurul Jadid, Probolinggo, Indonesia
Email: [1,*]selfiahafida02@gmail.com , [2]mfaid@unuja.ac.id, [3]ainolyaqin09@unuja.ac.id
Email Penulis Korespondensi: selfiahafida02@gmail.com
Submitted: 22/05/2024; Accepted: 30/05/2024; Published: 30/05/2024

Abstrak−Pelanggan adalah asset penting bagi kesukesan sebuah perusahaan dan memastikan kepuasan mereka adalah yang terpenting. Namun, perpindahan pelanggan yang terus menerus dapat menyebabkan berkurangnya nilai yang mengalir dari pelanggan, yang berpotensi membahayakan keunggulan kompetetif perusahaan. Perpindahan pelanggan, dimana konsumen memilih produk dari merek lain, di pengaruhi oleh berbagai faktor seperti promosi, harga, ketersediaan roduk, dan tingkat kepuasan pelanggan. Sementara penelitian tentang prediksi churn banyak yang terkonsentrasi di industri telekomunikasi, ritell, dan perbankan dan hanya sedikit yang melakukan penelitian prediksi churn terhadap toko online. Penelitian ini bertujan untuk memanfaatkan data mining dengan focus pada algoritma machine learning, khususnya metode tree-based gradient boosted models yang menerapkan model XGBoost, LightGBM, dan CatBoost, untuk memprediksi churn pelanggan di toko online. Metodelogi penelitian melibatkan pengumpulan data, pre-processing data, pemilihan dan pelatihan model, evaluasi model, analisis dan hasil. Penelitian ini menggunakan platform google collab dan beberapa library seperti library pandas, numpy, matplotlib, dan sebagainya. Hasil dari penelitian ini menunjukkan bahwa model XGBoost mencapai akurasi tertinggi dalam memprediksi perpindahan pelanggan, dengan kurva ROC sebesar 0,66 dan nilai akurasi sebesar 0.80032. Analisis feature importance menyoroti variable gender sebagai faktor penting dalam kinerja model. Penilitian ini berkontribusi dalam meningkatkan layanan pelanggan, meminimalisir terjadinya churn, dan pada akhirnya meningkatkan profitabilitas perusahaan di sektor toko online. Saran untuk penelitian di masa depan termasuk memperluas sumber data, menguji dengan lebih banyak metrik evaluasi, mengeksplorasi faktor churn tambahan dan membandingkan dengan metode prediksi lain untuk validasi.

Kata Kunci: Perpindahan Pelanggan; Toko Online; Data Mining; Tree-Based Gradient Boosted Models

Abstract−Customers are a critical asset to a company's success and ensuring their satisfaction is paramount. However, continuous churn can lead to reduced value flowing from customers, potentially jeopardizing a company's competitive advantage. Customer churn, where consumers choose products from other brands, is influenced by various factors such as promotion, price, product availability, and customer satisfaction levels. While much of the research on churn prediction is concentrated in the telecommunications, retail, and banking industries and only a few have conducted churn prediction research on online stores. This research aims to utilize data mining with a focus on machine learning algorithms, especially the tree-based gradient boosted models method that applies XGBoost, LightGBM, and CatBoost models, to predict customer churn in online stores. The research methodology involves data collection, data pre-processing, model selection and training, model evaluation, analysis and results. This research uses several libraries such as pandas library, numpy, matplotlib, and so on. The results of this study show that the XGBoost model achieved the highest accuracy in predicting customer churn, with an ROC curve of 0.66 and an accuracy value of 0.80032. The feature importance analysis highlights the gender variable as an important factor in model performance. This research contributes to improving customer service, minimizing churn, and ultimately increasing company profitability in the online store sector. Suggestions for future research include expanding data sources, testing with more evaluation metrics, exploring additional churn factors and comparing with other prediction methods for validation.

Keywords: Customer Churn; Online Stores; Data Mining; Tree-Based Gradient Boosted Models

1. PENDAHULUAN

Pelanggan adalah aset yang sangat penting bagi kesuksesan sebuah perusahaan. Oleh karena itu, segala cara dilakukan untuk memastikan kepuasan para pelanggan[1]. Pergantian pelanggan yang terus-menerus akan mengakibatkan penurunan nilai yang diperoleh perusahaan dari para pelanggannya. Dalam situasi yang tidak stabil, pelanggan yang terus berpindah dapat mengakibatkan perusahaan kehilangan keunggulan kompetetifnya di pasar. Apabila akusisi pelanggan baru tidak dapat mengimbangi kebutuhan pertumbuhan perusahaan, maka perusahaan akan menghadapi dilema dalam mempertahankan eksistensinya[2].

Perpindahan pelanggan atau yang biasa disebut *customer churn* terjadi ketika konsumen memilih produk dengan merk lain dari yang biasa mereka beli. Faktor-faktor yang mendorong perpindahan pelanggan ini beragam, seperti promosi, harga, penataan di toko, ketersediaan barang, inovasi produk, keinginan untuk mencoba hal baru, dan perubahan kualitas, atau tingkat kepuasan pelanggan[3]. Perpindahan pelanggan merujuk pada fenomena ketika pelanggan berhenti menggunakan produk atau layanan suatu perusahaan dan beralih ke pesaing atau menghentikan penggunaan sepenuhnya[4]. Perpindahan pelanggan memiliki dampak yang besar pada perusahaan, yang mengakibatkan kemungkinan terjadinya keuntungan atau kerugian dan bahkan penutupan bisnis (bangkrut)[5].

Dalam era digital ini, toko online atau *e-commerce* telah menjadi salah satu pilar utama dalam ekonomi global. Dengan semakin banyaknya konsumen yang beralih ke platform online untuk memenuhi kebutuhan sehari-hari, persaingan di industri ini semakin ketat. Dalam konteks ini, kemampuan untuk mempertahankan

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 3, May 2024, Page 605-614
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i3.5215

pelanggan menjadi faktor kunci yang menentukan keberhasilan dan berkelanjutan bisnis. Konsumen kini memiliki lebih banyak pilihan dan kemudahan dalam berbelanja, salah satunya melalui *platform* toko online yang semakin popular[6]. Toko online adalah metode bagi konsumen untuk melakukan pembelian barang secara online. Ini merupakan proses penjualan langsung yang menggunakan internet, baik untuk konsumen (bisnis ke konsumen) maupun antar bisnis (bisnis ke bisnis).

Keberadaan toko online atau *e-commerce* membawa banyak keuntungan dalam pemasaran produk, ini di anggap sebagai elemen krusial dari internet saat ini, terutama dengan perkembangan teknologi dan internet yang cepat di Indonesia. Perubahan dalam bisnis termasuk cara beriklan, jual beli, dan interaksi manusia telah dipengaruhi oleh kemajuan ini. Penjualan online telah terbukti meningkatkan penerimaan masyarakat[7]. Adanya *e-commerce* telah mempermudah dan memudahkan proses pemenuhan kebutuhan. Aktivitas ini dapat dilakukan dari rumah tanpa berinteraksi langsung, sehingga menghemat waktu[8]. Dengan kondisi teknologi modern saat ini, pelanggan yang ingin mengakses belanja online tidak perlu berada di tempat secara fisik, karena ada banyak lokasi di Indonesia yang menawarkan akses internet menggunakan Wi-Fi melalui laptop, notebook, atau *personal digital assistant(PDA)*[9].

Perpindahan pelanggan merupakan masalah krusial yang dapat membawa dampak negatif besar bagi perusahaan *e-commerce*. Penelitian ini menunjukkan bahwa biaya untuk memperoleh pelanggan baru bisa lima kali lebih besar daripada mempertahankan pelanggan yang sudah ada. Selain itu, tingkat churn yang tinggi dapat menyebabkan penurunan pendapatan, meningkatnya biaya akusisi,dan bahkan potensi kebangkrutan.

Namun Sebagian besar dari peneitian prediksi perpindahan pelanggan hanya terkonsentrasi di industri telekomunikasi, perbankan ritel, dan industry lainnya, dan hanya ada sedikit penelitian tentang prediksi *churn* di toko online. Pada saat ini toko online sangat popular di kalangan masyarakat, seperti beberapa *marketplace* yang banyak digunakan di Indonesia seperti Tokopedia, bukalapak, blibli, Lazada, dan shopee menjadi pilihan utama bagi jutaan konsumen[10]. Dengan tingginya jumlah transaksi harian dan beragamnya produk yang ditawarkan, toko online menghadapi tantangan besar dalam memahami dan mengelola perilaku pelanggan. *Customer churn* yang tinggi tidak hanya brdampak pada pendapatan tetapi juga pada reputasi perusahaan dipasar yang sangat kompetitif ini.

Para peneliti terdahulu telah melakukan studi mendalam tentang prediksi perpindahan pelanggan di industri telekomunikasi, perbankan, dan lainnya dengan menggunakan berbagai metode peramalan. Seperti pada penelitian[11] menggunakan metode *Logistic Regression* dan *Decission Tree* untuk memprediksi perpindahan pelanggan atau *customer churn*, yang dimana Logistic Regression memiliki potensi lebih tinggi dalam memprediksi churn pelanggan. Sedangkan pada penelitian [12] menggunakan metode CRISP-DM dan algoritma ridge classifer untuk memprediksi perpindahan pelanggan pada industri telekomunikasi. Model dipilih untuk tahap deployment. Dan pada penelitian [4] menggunakan algoritma C4.5 yang di optimalkan oleh Particle Swarm Optimization (PSO) pada industri telekomunikasi. Penggunaan PSO dalam algoritma C4.5 meningkatkan akurasi prediksi churn. Setelah optimasi dengan PSO. Model C4.5+PSO juga menunjukkan peningkatan dalam presisi, recall, dan F1-score. Penelitian [13] prediksi *customer churn* pada PT.Hutchison 3 Indonesia, yang menggunakan penerapan algoritma Naïve Bayes yang menghasilkan nilai akurasi sebesar 91,3%. Presisi,recall, dan F1 score sebesar 95 % menunjukkan kemampuan model dalam mengklasifikasi data *churn* dengan keakuratan yang tinggi. Dan pada penelitian [14] prediksi churn nasabah bank menggunakan klasifikasi random forest dan decision tree dengan evaluasi confusion matrix. Hasil evaluasi yang di peroleh random forest lebih dibandingkan dengan decision tree yaitu sebesar 78% untuk random forest, 72% untuk decision tree, sehingga random forest adalah alat yang lebih efisien dan efektif dalam memprediksi churn nasabah dan memberikan kontribusi signifikan dalam analisis prediktif di sektor perbankan.

Meskipun banyak penelitian telah dilakukan di sektor telekomunikasi dan perbankan, penelitian tentang prediksi perpindahan pelanggan masih terbatas. Selain itu, metode Tree-based gradient boosted models, yang dikenal memiliki kemampuan prediktif yang kuat, belum banyak diimplementasikan dalam konteks toko online. Oleh karena itu, penelitian ini bertujuan untuk mengisi kesenjangan tersebut dengan mengembangkan model prediksi churn yang spesifik untuk toko online menggunakan metode tersebut. Model ini sering dipilih karena kemampuannya menghasilkan prediksi yang kuat dan umumnya tahan terhadap overfitting, dan model ini juga menawarkan implementasi yang efisien untuk algoritma gradient boosting dan digunakan secara luas didunia akademis dan industri untuk berbagai tugas pembelajaran mesin (*machine learning*).

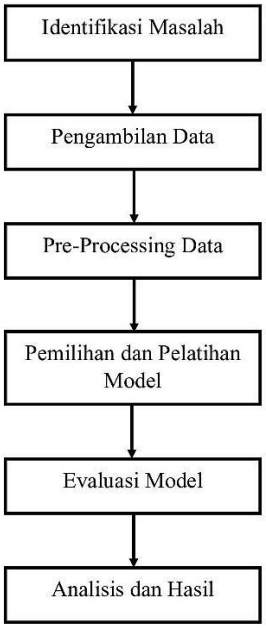
Jika masalah perpindahan pelanggan ini tidak segera di atasi, perusahaan *e-commerce* berisiko menghadapi penurunan pangsa pasar, kerugian finansial yang signifikan, dan penurunan loyalitas pelanggan dalam jangka Panjang. Dengan demikian, menemukan metode yang efektif untuk memprediksi dan mengurangi terjadinya perpindahan pelanggan adalah kebutuhan.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 3, May 2024, Page 605-614
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i3.5215

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Pada penelitian ini dilakukan dengan beberapa tahapan. Tahapan tersebut dimulai dari mengidentifikasi masalah, pengambilan data, pre-processing data, pemilihan dan pelatihan model, evaluasi model, hingga proses analisis dan hasil.



Gambar 1. Tahapan penelitian

Gambar 1 merupakan gambaran umum tentang tahapan-tahapan penelitian dalam memprediksi perpindahan pelanggan menggunakan metode Tree-Based Gradient Boosted Models. Proses awal dari penelitian tersebut adalah mengidentifikasi masalah dengan tujuan memahami akar penyebab masalah secara lebih mendalam dan menyeluruh, dalam penelitian ini ditemukan permasalahan terjadinya perpindahan pelanggan yang dapat mengakibatkan turunnya performa toko dan dapat mengakibatkan kerugian. Langkah selanjutnya adalah pengambilan data, dataset yang digunakan dalam penelitian ini diambil dari halaman website *Kaggle.com*.

Selanjutnya melakukan pre-processing data seperti penanganan nilai yang hilang (*missing values*), menghapus variabel yang tidak diperlukan, visualisasi, deteksi outlier, dan feature engineering. Feature engineering adalah Teknik yang diterapkan setelah data input dikumpulkan dan dibersihkan. Hal ini dapat dilakukan sebelum melakukan membangun model *machine learning*[15].

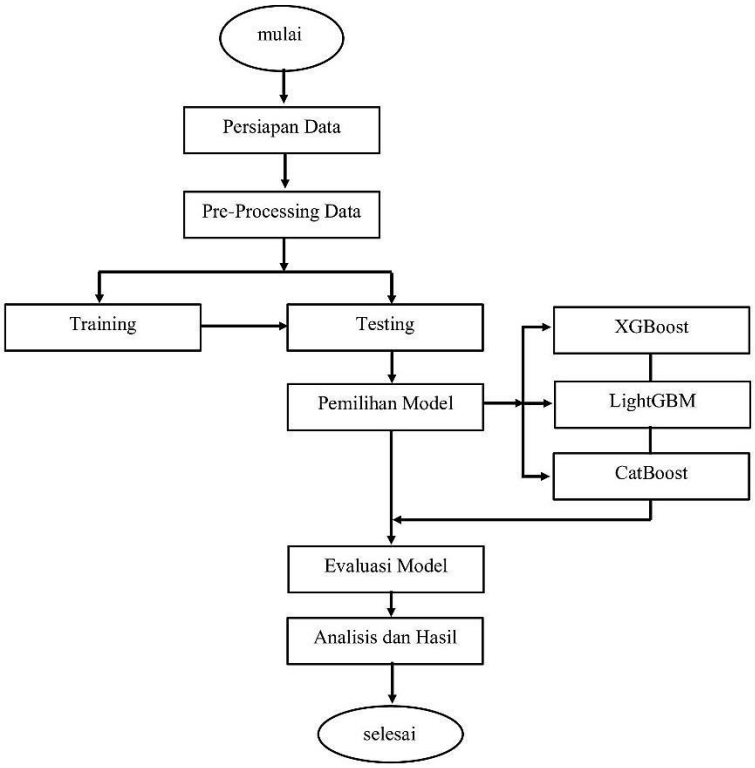
Tahap pemilihan dan pelatihan model, pada tahap ini metode tree-based gradient boosted models menawarkan beberapa model, dan penelitian ini akan menggunakan implementasi model yang populer dalam metode tersebut yaitu model XGBoost, LightGBM, dan CatBoost. Tahapan selanjutnya pelatihan model, yang dimana membagi data menjadi set data pelatihan dan set data pengujian menggunakan sklearn train_test_split.

Kemudian tahap evaluasi model adalah proses untuk mengukur kinerja atau kemampuan model yang telah dibuat dalam memprediksi atau mengkasifikasi data. Pada penelitian ini menggunakan beberapa metode evaluasi model yaitu, akurasi, presisi, recall, F1-score, dan ROC curve. Dan yang terakhis adalah tahap analisis dan hasil, Pada tahap ini menjelaskan model yang memperoleh kinerja model terbaik dan melakukan feature importance untuk mrngetahui variabel yang penting dalam kinerja model.

2.2 Tree-Based Gradient Boosted Models

Tree-Based gradient boosted models atau model gradient boosting berbasis keputusan adalah Teknik yang popular untuk mengklasifikasikan dan meramalkan masalah. Metode ini meningkatkan prosedur pembelajaran dengan menyederhanakan tujuan dan mengurangi jumlah iterasi yang di perlukan untuk mencapai solusi yang cukup optimal[16]. Ada beberapa model yang populer dan efektif yang menggunakan metode tersebut dalam *machine learning*, di antaranya adalah gradient boosting machine (GBM), extreme gradient boosting (XGBoost), LightGBM, CatBoost, dan Hist gradient boosting. Semua model ini menggunakan ide gradient boosting berbasis pohon keputusan sebagai pembelajar utamanya. Namun, setiap model memiliki keunikan dalam fitur dan optimisasi, sehingga mereka sesuai digunakan dalam situasi yang berbeda. Pada penelitian ini menggunakan tiga model sebagai pemecahan masalahnya yaitu model XGBoost, LightGBM, dan CatBoost. Berikut adalah diagram alir penerapan metode tree-based gradient boosted models yang ditunjukkan pada gambar 2.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 3, May 2024, Page 605-614
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i3.5215



Gambar 2. Diagram alir penerapan metode tree-based gradient boosted models yang menggunakan algoritma XGBoost, LightGBM, dan CatBoost

Diagram alir tersebut menunjukkan proses penerapan metode tree-based gradient boosted models yang di awali dengan persiapan data seperti membaca dataset, menampilkan ukuran data, dan melihat tipe data. Selanjutnya adalah pre-processing data, pada Langkah ini hal yang di lakukan adalah cek *missing values* atau nilai hilang serta menangani *missing values*, selanjutnya dilakukan penghapusan variabel yang tidak diperlukan untuk mempermudah melakukan analisis dan prediksi, visualisasi, mendeteksi data outlier, feauture engineering, dan melakukan pembagian data menjadi data pelatihan (*training*) dan data pengujian (*testing)* menggunakan sklearn *train_test_split*. Train-test split merupakan metode yang digunakan untuk memperkirakan kinerja algoritma *machine learning* saat diterapkan pada data yang tidak digunakan dalam proses pelatihan model. Metode ini membagi data menjadi data test dan data train sehingga memungkinkan evaluasi prediksi algoritma pada data yang belum pernah dilihat sebelumnya[17].

Setelah melakukan pembagian data menjadi data *train* dan data *test*, Langkah selanjunya adalah memilih model algoritma yang akan digunakan untuk melakukan evaluasi model. Pada penelitian ini memilih algoritma XGBoost, LightGBM, dan CatBoost untuk penyelesaian masalahnya. Pada saat melakukan uji pada model, penelitian ini menggunakan *cross_val_score* dalam scikit-learn yang digunakan untuk mengevaluasi kinerja model machine learning dengan Teknik *cross validation*. *Cross validation*, juga dikenal sebagai estimasi rotasi, adalah Teknik validasi model untuk menilai generalisasi hasil analisis statistic pada data independent. Teknik ini dignakan untuk memprediksi model dan meperkirakan akurasinya dalam praktik[18].

$$\text{akurasi} = \left(\frac{\sum klasifikasi\ benar}{\sum data\ uji}\right) \times 100\% \quad (1)$$

Rumus tersebut merupakan rumus K-Fold cross validation, dimana variabel akurasi adalah hasil keakuratan, klasifikasi benar adalah jumlah prediksi yang benar, dan data uji adalah jumlah data yang di uji[19]. Kemudian melakukan evaluasi model dan yang terakhir melakukan analisis dan hasil.

3. HASIL DAN PEMBAHASAN

3.1 Pengambilan Data

Pada penelitian ini pengambilan data, data yang dipilih adalah data sekunder yang telah disediakan oleh website *kaggle.com.* Kaggle adalah platform terkenal didunia data science dan machine learning yang menawarkan lebih dari 60000 dataset dan memiliki komunitas data terbesar saat ini[20]. Data yang di ambil merupakan dataset baru yang telah update 7 bulan yang lalu.

Kumpulan data ini terdiri dari berbagai atribut yang berkaitan dengan pelanggan toko online, yang memungkinkan kami untuk menyimpulkan hubungan yang konsisten antara Tindakan pelanggan dan *churn*.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 3, May 2024, Page 605-614
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i3.5215

Pada dataset ini memiliki ukuran data 25000 data observasi dan 13 variabel, yang terdiri dari *Customer_ID*, *Purchase_Date*, *Product_Category*, *Product_ Price*, *Quantity*, *Total_Purchase_Amount*, *Payment-Method*, *Customer_Age*, *Returns*, *Customer_Name*, *Age*, *Gender*, *dan Churn.*

3.2 Persiapan Data

Setelah melakukan pengambilan dataset, Langkah selanjutnya adalah melakukan persiapan data seperti mengimport dataset ke google colab.

```
#Memuat 5 data pertama dari dataset
data.head()
```

	Customer_ID	Purchase_Date	Product_Category	Product_Price	Quantity	Total_Purchase_Amount	Payment_Method	Customer_Age	Returns	Customer_Name	Age	Gender	Churn
0	46251	9/8/2020 9:38	Electronics	12	3	740	Credit Card	37	0.0	Christine Hernandez	37	Male	0.0
1	46251	3/5/2022 12:56	Home	468	4	2739	PayPal	37	0.0	Christine Hernandez	37	Male	0.0
2	46251	5/23/2022 18:18	Home	288	2	3196	PayPal	37	0.0	Christine Hernandez	37	Male	0.0
3	46251	11/12/2020 13:13	Clothing	196	1	3509	PayPal	37	0.0	Christine Hernandez	37	Male	0.0
4	13593	11/27/2020 17:55	Home	449	1	3452	Credit Card	49	0.0	James Grant	49	Female	1.0

Gambar 3. Import dataset kedalam google colab

Gambar 3 menunjukkan 5 baris pertama dari dataset yang telah di import kedalam google colab dengan libarary pandas yang berfungsi untuk membaca dataset dari file CSV, mengolah data dalam bentuk dataframe, dan sebagainya.

3.3 Pre-Processing Data

Tahap pre-processing ini operasi yang dilakukan adalah penanganan nilai yang hilang (*missing values*), menghapus variable yang tidak diperlukan, visualisasi, deteksi outlier, dan feature engineering. Langkah selanjutnya adalah pre-processing data, pada tahap ini melakukan *checking missing values* dan melakukan penanganan *missing values* untuk memastikan kualitas data yang digunakan.

```
# check missing values (nilai yang hilang)
data.isnull().sum()
```

```
Customer_ID                0
Purchase_Date              0
Product_Category           0
Product_Price              0
Quantity                   0
Total_Purchase_Amount      0
Payment_Method             0
Customer_Age               0
Returns                47596
Customer_Name              0
Age                        0
Gender                     0
Churn                      0
dtype: int64
```

Gambar 4. Cek missing value (nilai yang hilang)

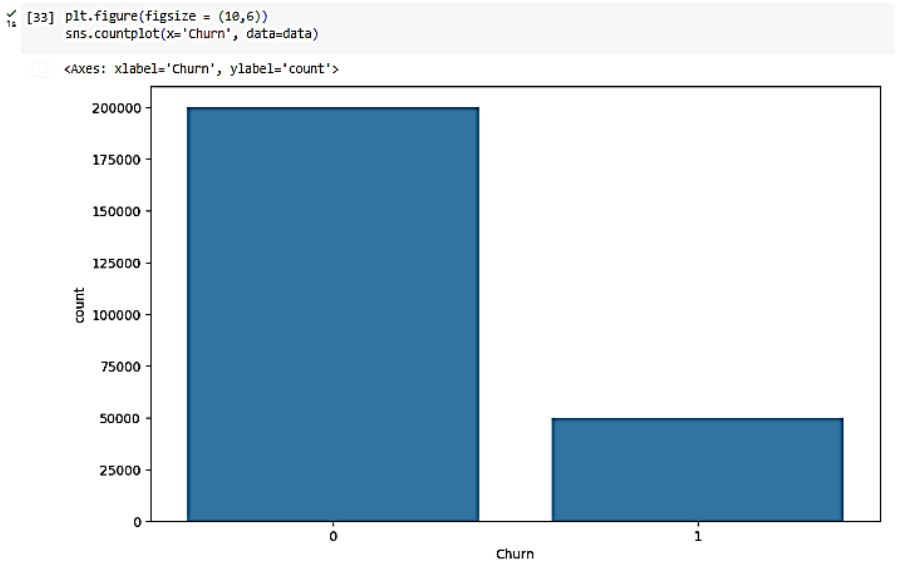
```
Customer_ID                0
Purchase_Date              0
Product_Category           0
Product_Price              0
Quantity                   0
Total_Purchase_Amount      0
Payment_Method             0
Customer_Age               0
Returns                    0
Customer_Name              0
Age                        0
Gender                     0
Churn                      0
dtype: int64
```

Gambar 5. Setelah dilakukan penanganan missing value

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 3, May 2024, Page 605-614
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i3.5215

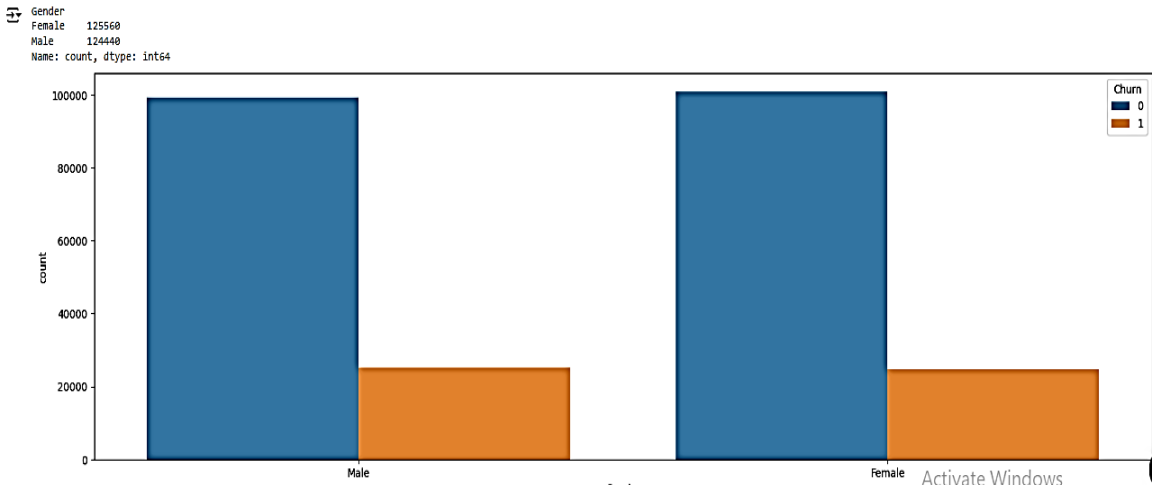
Dari gambar 4 di tunjukkan bahwa adanya *missing values* atau nilai yang hilang pada variabel returns dengan jumlah 47596 nilai yang hilang, setelah dilakukan penangan *missing values* hasilnya dapat dilihat pada gambar 5. Langkah selanjutnya dalam pre-processing data adalah menghapus variabel yang tidak diperlukan agar mempermudah penelitian. Dalam penelitian ini variabel yang di hapus adalah "*Customer_id*, *Customer_Name*, *Purchase_Date*, *Customer_age*, dan *Payment_Method"*.

Setelah melakukan cek *missing values*, penanganan *missing values*, dan menghapus variabel-variabel yang tidak diperlukan, langkah selanjutnya adalah visualisasi dasar untuk memahami bagaimana data di distribusikan. Pada langkah visualisasi yang dilakukan adalah menganilisis variabel target yang dipilih, dan pada penelitian ini memilih variabel "churn" sebagai variabel targetnya.



Gambar 6. Visualisasi variabel *churn*

Hasil dari analasis varibel target(churn) adalah jumlah pelanggan yang melakukan churn atau berpindah lebih sedikit dari pelanggan tetap. Selanjutnya kami memvisualisasikan hubungan variabel target dan variabel kategori dan numerik. Pada tahap ini menggunakan variabel *Gender*, *Age*, dan *product_category* sebagai variabel yang berhubungan dengan variabel target.

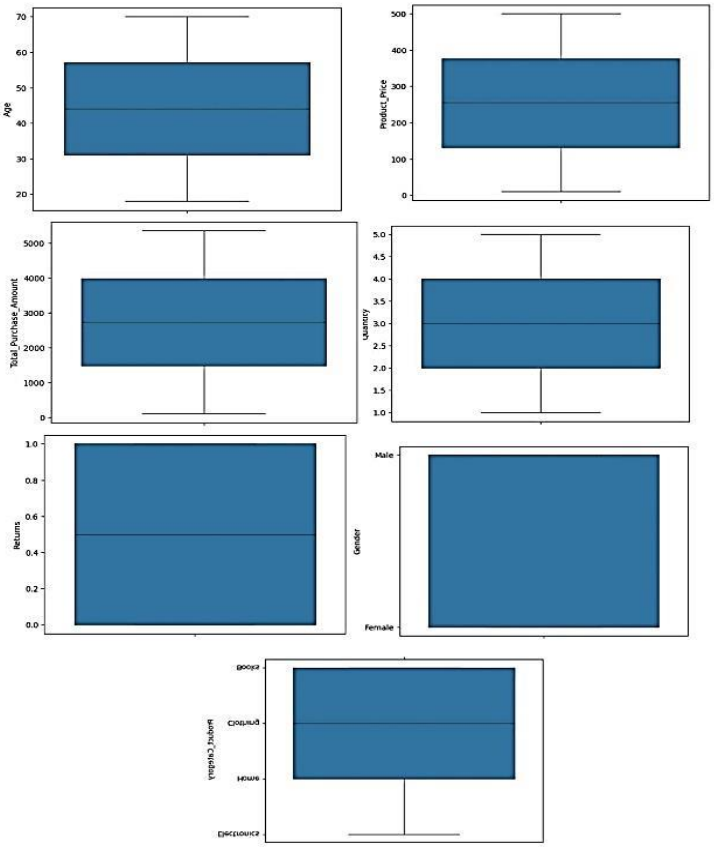


Gambar 7. Visualisasi variabel target (*churn*) dengan variabel *gender*

Hasil dari visualisasi hubungan variabel target dengan variabel *Gender* menganalisis bahwa pelanggan wanita lebih sering melakukan perpindahan dibandingkan dengan pelanggan laki-laki. Variabel *Age* atau umur menunjukkan usia kisaran 58 tahun seringkali melakukan tindakan churn, dan variabel *Product_Category* menunjukkan bahwa produk *clothing* (pakaian) sering membuat pelanggan melakukan churn.

Selanjutnya adalah melakukan deteksi outlier dalam kumpulan data, pada penelitian ini melakukan visualisasi dasar menggunakan *boxplot of the seaborn library* untuk mendeteksi outlier. Hasil deteksi outlier menggunakan boxplot ditunjukkan pada Gambar 8.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 3, May 2024, Page 605-614
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i3.5215



Gambar 8. Boxplot deteksi outliers

Dari gambar 8 tersebut menunjukkan tidak adanya outlier dalam data. Karena tidak adanya outlier dalam data, maka selanjutnya melakukan *feature engineering* untuk mengubah data kategorikal menjadi data numerik guna menyiapkan data untuk pemodelan dan karenanya menciptakan lebih banyak fitur dalam kumpulan data. Karena variable Product_Category adalah data kategori, maka penelitian memberikan one-hot encoding dengan menggunakan library pandas (pd.get_dummies) untuk membuat lebih banyak fitur dari variable Product_Category. Selain itu, penelitian ini juga membuat fungsi untuk mengubah data kategorikal dalam variable "Gender" menjadi data numerik, missal laki-laki = 0 sedangkan perempuan = 1.

3.4 Pemilihan dan Pelatihan Model

Metode tree-based gradient boosted models memiliki banyak model yang ditawarkan untuk membangun dan mengembangkan berbagaimodel berdasarkan beberapa teknik pemodelan yang berbeda. Pada tahap ini, peneliti memilih 3 model algoritma sebagai pemecahan masalahnya yaitu model XGBoost, LightGBM, dan CatBoost. Setelah melakukan pemilihan model, Langkah selanjutnya adalah melakukan pelatihan model yang dimanan membagi data menjadi set data training dan testing mengggunakan pusrtaka train test split.

3.5 Evaluasi Model

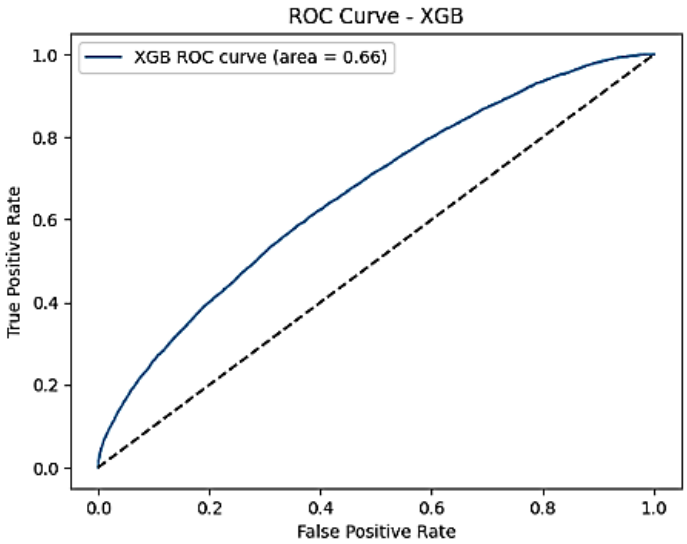
Setelah melakukan beberapa operasi seperti penyetelan hyperparameter, validasi silang, output tertinggi diambi oleh LightGBM Classifer kemudian di susul XGB Classifer, dan yang terahir CatBoost Classifer. Kinerja setiap pengklasifikasi dapat dilihat pada tabel 1.

Tabel 1. Hasil nilai akurasi

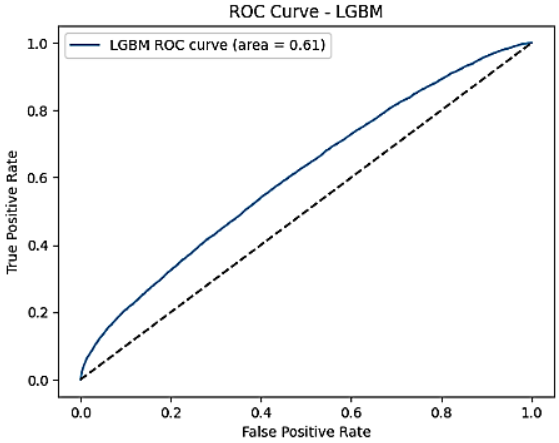
Algoritma	Akurasi
XGBoost Classifer	0.80032
LightGBM Classifer	0.80032
Catboost Classifer	0.80026

Skor akurasi mungkin bukan metrik terbaik untuk mengevaluasi performa model, oleh karena itu pada penelitian ini juga menggunakan f1 score, recall, precission, dan ROC *(Receiver Operaing Characteristic).* Hasil evaluasi menggunakan plot ROC kurva ditunjukkan pada gambar 9 yang merupakan hasil dari model XGBoost, gambar 10 hasil dari model LightGBM, dan gambar 11 hasil dari model catboost.

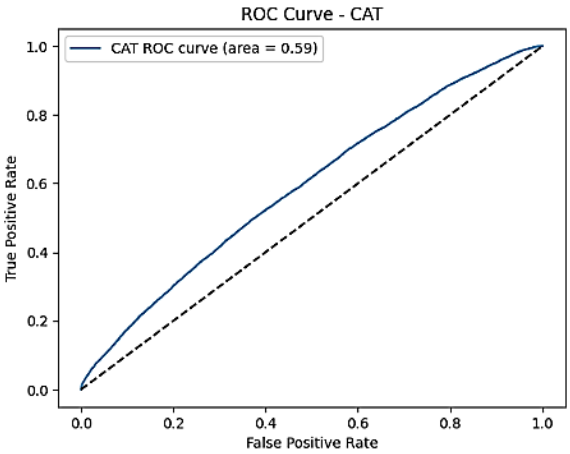
Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 3, May 2024, Page 605-614
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i3.5215



Gambar 9. Plot ROC curve XGBoost



Gambar 10. Plot ROC curve LightGBM



Gambar 11. Plot ROC curve CatBoost

Dari gambar kurva ROC diatas menunjukkan model XGB memperoleh skor keberhasilan terbaik dengan nilai akurasi 0.80032 dan kurva ROC 0,66.

3.6 Analisis dan Hasil

Tahap analisis dan hasil merupakan tahap akhir dari penelitian ini, setelah melakukan beberapa tahapan–tahapan, penelitian ini menghasilkan model XGBoost sebagai model terbaik dalam prediksi perpindahan pelanggan dengan kurva ROC 0,66 dan nilai akurasi, presisi, f1 score, dan recall dapat dilihat pada gambar 12.

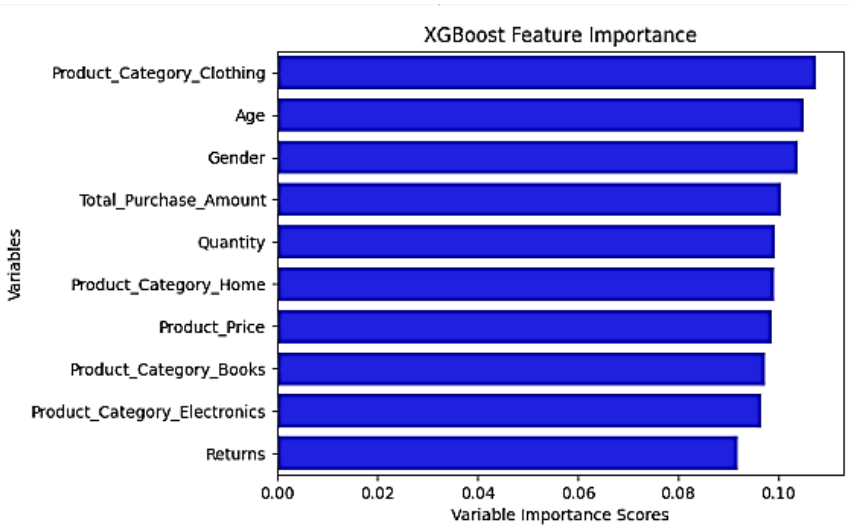
Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 3, May 2024, Page 605-614
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i3.5215

	precision	recall	f1-score	support
0	0.80	1.00	0.89	40016
1	0.00	0.00	0.00	9984
accuracy			0.80	50000
macro avg	0.40	0.50	0.44	50000
weighted avg	0.64	0.80	0.71	50000

Accuracy score of XGB model: 0.80032

Gambar 12. Hasil evaluasi model XGBoost

Karena model XGBoost memperoleh skor terbaik, peneliti melakukan *feature importance* yaitu fitur prediktitif (variabel) yang paling penting dalam kinerja model. Hasil dari barplot *feature importance* di tunjukkan pada gambar 13 berikut.



Gambar 13. Barplot feature importance XGB

Dari gambar 13 tersebut menunjukkan bar plot feature importance dari model xgboost untuk variabel-variabel yang digunakan dalam model tersebut. Hasil dari gambar bar plot feature importance tersebut menunjukkan variabel *Product_Category_Clothing* merupakan variabel penting dalam kinerja model.

Dengan hal tersebut model XGBoost dapat dilatih dengan baik untuk mendeteksi pola yang lebih kompleks dalam data dan mencapai nilai akurasi yang tinggi. Nilai akurasi yang tinggi diperlukan untuk dapat mengidentifikasi kasus-kasus pelanggan yang melakukan tindakan *churn*.

4. KESIMPULAN

Metode tree-based gradient boosted models, khususnya model XGBoost, berhasil digunakan untuk memprediksi perpindahan pelanggan pada toko online dengan nilai akurasi yang cukup tinggi (0,80032) dan kurva ROC sebesar 0,66. Evaluasi model dilakukan menggunakan beberapa metrik seperti f1 score, recall, presisi, dan plot ROC curve, yang membantu mengukur kinerja dan kehandalan model prediksi. Analisis feature importance menunjukkan bahwa variable *product_category_clothing* menjadi variabel penting dalam kinerja model XGBoost, yang dapat memberikan wawasan lebih tentang faktor-faktor yang mempengaruhi perpindahan pelanggan. Hal ini juga dapat membantu perusahaan meningkatkan layanan kepada pelanggan dan meminimalisir jumlah perpindahan pelanggan yang dapat berdampak pada keuntungan perusahaan. Penulis menyadari bahwa masih banyak kekurangan dalam penelitian ini, oleh karena itu penulis akan memberikan beberapa saran. Untuk meningkatkan nilai akurasi, penulis menyarankan untuk memperluas cakupan data dengan mengambil dataset dari sumber yang lebih beragam untuk meningkatkan generalisasi model prediksi. Melakukan pengujian model dengan lebih banyak metrik evaluasi dan teknik cross-validation untuk memastikan kehandalan dan stabilitas model. Menggali lebih dalam lagi faktor-faktor lain yang dapat memengaruhi perpindahan pelanggan, selain dari variable yang telah dipertimbangkan dalam penelitian ini. Melakukan perbandingan dengan metode-metode prediksi lainnya untuk memperkuat validitas hasil penelitian. Dengan demikian, penelitian ini dapat terus ditingkatkan untuk memberikan kontribusi yang lebih besar dalam memahami dan mengelola perpindahan pelanggan pada toko online secara efektif.

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 3, May 2024, Page 605-614
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i3.5215

REFERENCES

[1] V. R. R. Raj and R. A. Azad .V, "Customer Churn Prediction in Telecommunication Industry Having Data Certainty," *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 4099, pp. 113–122, 2020, doi: 10.32628/ijsrset207427.

[2] C. A. License, Q. Zeng, M. Chang, Q. Tong, and J. Su, "Retracted: A Prediction Model of Customer Churn considering Customer Value: An Empirical Research of Telecom Industry in China," *Discret. Dyn. Nat. Soc.*, vol. 2023, pp. 1–1, 2023, doi: 10.1155/2023/9876034.

[3] D. V. Hanifah and Y. P. Astuti, "Analisis Perpindahan Pelanggan Dan Strategi Persaingan Restoran Dengan Metode Markov Chain Dan Game Theory," *MATHunesa J. Ilm. Mat.*, vol. 11, no. 3, pp. 310–317, 2023, doi: 10.26740/mathunesa.v11n3.p310-317.

[4] M. Rizki Kurniawan, P. Nurul Sabrina, and R. Ilyas, "Prediksi Customer Churn Pada Perusahaan Telekomunikasi Menggunakan Algoritma C4.5 Berbasis Particle Swarm Optimization," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 5, pp. 3369–3375, 2024, doi: 10.36040/jati.v7i5.7476.

[5] A. Khattak, Z. Mehak, H. Ahmad, M. U. Asghar, M. Z. Asghar, and A. Khan, "Customer churn prediction using composite deep learning technique," *Sci. Rep.*, vol. 13, no. 1, pp. 1–17, 2023, doi: 10.1038/s41598-023-44396-w.

[6] D. Ika Sugiarti and R. Iskandar, "Pengaruh Consumer Review Terhadap Keputusan Pembeli Terhadap Toko Online Shopee," *J. Sos. Teknol.*, vol. 1, no. 9, pp. 954–962, 2021, doi: 10.59188/jurnalsostech.v1i9.195.

[7] E. T. Oktaria, Y. Yuniarthe, H. Hairudin, and ..., "Sarana Publikasi Dan Media Promosi Produk Kreatifitas Siswa Menggunakan E-Commerce Pada Smk Gading Rejo Kabupaten …," *J. Pengabdi. …*, vol. 2, pp. 78–83, 2023, [Online]. Available: https://www.jpu.ubl.ac.id/index.php/jpu/article/view/34%0Ahttps://www.jpu.ubl.ac.id/index.php/jpu/article/download/34/32

[8] L. Dwi, "Perbandingan Performa Model Prediksi Customer Churn Berbasis Machine Learning Pada Fashion E-Commerce," 2023.

[9] Z. Kedah, "Use of E-Commerce in The World of Business," *Startupreneur Bus. Digit. (SABDA Journal)*, vol. 2, no. 1, pp. 51–60, 2023, doi: 10.33050/sabda.v2i1.273.

[10] X. Xiahou and Y. Harada, "B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM," *J. Theor. Appl. Electron. Commer. Res.*, vol. 17, no. 2, pp. 458–475, 2022, doi: 10.3390/jtaer17020024.

[11] A. Mauludin, N. Aziz, A. Mauliddin, V. A. Sintalana, D. Hafiz, and A. A. Rismayadi, "Prediksi Customer Churn Menggunakan Logistic Regression dan Decission Tree," vol. 4, no. 1, pp. 11–19, 2023.

[12] Y. Yudiana, A. Yulia Agustina, and dan Nur Khofifah, "Prediksi Customer Churn Menggunakan Metode CRISP-DM Pada Industri Telekomunikasi Sebagai Implementasi Mempertahankan Pelanggan," *Indones. J. Islam. Econ. Bus.*, vol. 8, no. 1, pp. 01–20, 2023, [Online]. Available: http://e-journal.lp2m.uinjambi.ac.id/ojp/index.php/ijoieb

[13] R. Alfarez and V. Purwayoga, "PENERAPAN NAÏVE BAYES UNTUK PREDIKSI CUSTOMER CHURN (STUDI KASUS : PT HUTCHISON 3 INDONESIA)," vol. 05, no. 02, pp. 301–307, 2024.

[14] A. F. Azmi and A. Voutama, "KOMPUTA : Jurnal Ilmiah Komputer dan Informatika PREDIKSI CHURN NASABAH BANK MENGGUNAKAN KLASIFIKASI RANDOM FOREST DAN DECISION TREE DENGAN EVALUASI CONFUSION MATRIX KOMPUTA : Jurnal Ilmiah Komputer dan Informatika," vol. 13, no. 1, 2024.

[15] T. Verdonck, B. Baesens, M. Óskarsdóttir, and S. vanden Broucke, "Special issue on feature engineering editorial," *Mach. Learn.*, no. 0123456789, 2021, doi: 10.1007/s10994-021-06042-2.

[16] N. Subramani, S. V. Easwaramoorthy, P. Mohan, M. Subramanian, and V. Sambath, "A Gradient Boosted Decision Tree-Based Influencer Prediction in Social Network Analysis," *Big Data Cogn. Comput.*, vol. 7, no. 1, 2023, doi: 10.3390/bdcc7010006.

[17] J. Brownlee, "Train-Test Split for Evaluating Machine Learning Algorithms," machine learning mastery. Tanggal akses 20 Februari 2024 [Online]. Available: https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/

[18] H. Azis, P. Purnawansyah, F. Fattah, and I. P. Putri, "Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 81–86, 2020, doi: 10.33096/ilkom.v12i2.507.81-86.

[19] A. P. Windarto, S. Defit, and A. Wanto, "Optimalisasi Parameter dengan Cross Validation dan Neural Back-propagation Pada Model Prediksi Pertumbuhan Industri Mikro dan Kecil," *J. Sist. Inf. Bisnis*, vol. 11, no. 1, pp. 34–42, 2021, doi: 10.21456/vol11iss1pp34-42.

[20] V. V. Putri, A. Tholib, and C. Novia, "Deteksi Kaggle Bot Account Menggunakan Deep Neural Networks," *NJCA (Nusantara J. Comput. Its Appl.*, vol. 8, no. 1, p. 13, 2023, doi: 10.36564/njca.v8i1.304.