Computer Science (CO-SCIENCE)
Vol. 5 No. 1 Januari 2025
Akreditasi Sinta 4 Nomor SK : 230/E/KPT/2022

Customer Churn Prediction Pada Sektor Perbankan Dengan Model Logistic Regression dan Random Forest

Elly Mufida*[1], Doni Andriansyah[2], Hylenarti Hertyana[3]

[1,2,3]Universitas Bina Sarana Informatika;
Jl. Kramat Raya 98. Jakarta Pusat, Indonesia
e-mail: *[1]elly.elm@bsi.ac.id, [2]doni.dad@bsi.ac.id, [3]hylenarti.hha@bsi.ac.id

(*) Corresponding Author

Artikel Info : Diterima : 03-12-2024 | Direvisi : 01-01-2025 | Disetujui : 31-01-2025

Abstrak - Customer churn adalah fenomena yang merugikan di sektor perbankan karena dapat mengurangi pendapatan dan meningkatkan biaya akuisisi pelanggan baru. Penelitian ini bertujuan untuk membandingkan performa dua model, Logistic Regression dan Random Forest, untuk memprediksi customer churn menggunakan dataset dari Kaggle. Proses penelitian melibatkan preprocessing data seperti normalisasi z-score dan pembagian dataset menjadi data pelatihan (70%) dan data pengujian (30%). Model dievaluasi menggunakan confusion matrix dengan nilai akurasi, precision, recall, dan F1-Score. Logistic Regression mencapai akurasi 76,85%, precision 79%, recall 94%, dan F1-Score 86%, menunjukkan performa cukup baik namun rentan terhadap false positives. Sebaliknya, Random Forest menunjukkan performa lebih unggul dengan akurasi 83,12%, precision 84%, recall 96%, dan F1-Score 90%. Random Forest cocok untuk masalah dengan kebutuhan recall tinggi karena lebih andal dalam mendeteksi customer churn potensial. Untuk meningkatkan performa model lebih lanjut, disarankan melakukan optimasi hyperparameter dan analisis feature importance. Model prediksi churn ini diharapkan dapat membantu bank mengurangi churn dan meningkatkan retensi pelanggan.

Kata Kunci : Customer Churn. Logistic Regression, Random Forest. Unsupervised Learning

Abstracts *– Customer churn is a detrimental phenomenon in the banking sector because it can reduce revenue and increase the cost of acquiring new customers. This research aims to compare the performance of two models, Logistic Regression and Random Forest, to predict customer churn using datasets from Kaggle. The research process involves data preprocessing such as z-score normalization and dividing the dataset into training data (70%) and testing data (30%). The model was evaluated using a confusion matrix with Accuracy, precision, recall and F1-Score values. Logistic Regression achieved 76.85% Accuracy, 79% precision, 94% recall, and 86% F1-Score, showing quite good performance but susceptible to false positives. In contrast, Random Forest shows superior performance with 83.12% Accuracy, 84% precision, 96% recall, and 90% F1-Score. Random Forest is suitable for problems with high recall requirements because it is more reliable in detecting potential customer churn. To further improve model performance, it is recommended to perform hyperparameter optimization and feature importance analysis. This churn prediction model is expected to help banks reduce churn and increase customer retention.*

Keywords : Customer Churn. Logistic Regression, Random Forest. Unsupervised Learning

PENDAHULUAN

Customer churn adalah istilah yang digunakan untuk menggambarkan hilangnya pelanggan dari suatu bisnis atau layanan. Customer churn disebut juga dengan customer turnover, customer attrition, atau customer deflection (Pondel et al., 2021) (Geiler et al., 2022). Pelanggan yang melakukan churn berarti berhenti menggunakan produk atau layanan yang ditawarkan oleh perusahaan, baik secara langsung dengan menutup akun atau dengan tidak memperpanjang langganan, atau secara tidak langsung dengan berhenti bertransaksi. Churn pelanggan

adalah fenomena yang merugikan bagi industri perbankan karena kehilangan pelanggan dapat mengurangi pendapatan. Mempertahankan pelanggan lebih ekonomis daripada mendapatkan yang baru (Hussain et al., 2023).

Penelitian mengenai Customer churn digunakan pada berbagai kasus, terutama dalam industri yang berbasis langganan atau layanan berulang (recurring service), seperti: industri telekomunikasi (Abdulsalam et al., 2022) (Herdian & Girsang, 2023) (Sidiq & Anggraini, 2023), Perbankan dan Keuangan (Hussain et al., 2023) (Patricia et al., 2023), layanan streaming, e-Commerce (Sholeha et al., 2024), Software as a Service (SaaS), Perhotelan dan Pariwisata (Taherkhani et al., 2023), serta Retail berbasis Keanggotaan (Firmansyah & Yulianto, 2021). Pada penelitian sebelumnya, churn prediction dibangun tidak hanya menggunakan model prediksi tunggal, seperti churn prediction pada sektor perbankan dengan model decision tree dan naive bayes (Aksama & Wahyuniati, 2022), model Logistic Regression pada prediksi IBM Telco Customer churn (Sidiq & Anggraini, 2023), namun juga menggunakan model ensemble seperti yang ada pada tabel 1. Ensemble learning menggabungkan beberapa algoritma machine learning untuk menghasilkan model prediksi. yang lebih kuat dibandingkan model individu. Namun, mengkombinasikan model dengan akurasi yang bervariasi tidak selalu meningkatkan hasil prediksi; model dengan akurasi rendah dapat mengaburkan kontribusi model dengan akurasi tinggi (Naderalvojoud & Hernandez-Boussard, 2023).

Persaingan yang terus meningkat dalam industri membuat perusahaan untuk serius mengendalikan customer churn (Geiler et al., 2022). Customer churn dapat dianggap sebagai peluang keuntungan yang hilang, karnena untuk mendapatkan pelanggan baru biasanya lima hingga enam kali lebih tinggi daripada biaya untuk mempertahankan pelanggan yang sudah ada (Pondel et al., 2021). Banyak perusahaan yang lebih memilih mempertahankan pelanggan yang sudah ada dibanding dengan mengakuisisi pelanggan baru (Pondel et al., 2021) (Naderalvojoud & Hernandez-Boussard, 2023).

Dataset yang digunakan untuk model customer churn prediction termasuk ke dalam supervised learning dengan target binominal, sehingga banyak algoritma yang dapat digunakan. Tabel 1 merangkum beberapa algoritma yang digunakan pada customer churn prediction. Pengujian dan evaluasi yang digunakan bergantung pada algoritma yang digunakan. Pada penelitian sebelumnya, Convusion Matrix dan ROC adalah instrumen yang paling banyak digunakan untuk menguji dan mengevaluasi model customer churn prediction.

Tabel 1. Rangkuman penelitian sebelumnya mengenai Customer Churn Prediction

Peneliti	Kasus Churn	Algoritma yang digunakan	Pengujian dan Evaluasi yang digunakan	Hasil Penelitian
(Hussain et al., 2023)	Sektor Perbankan	Logistic Regression, Random Forest, Decision Tree, dan Extreme Gradient Boosting (XGBoost)	Accuracy dan F1 Score	XGBoost dipilih sebagai model terbaik berdasarkan evaluasi performa
(Abdulsalam et al., 2022)	Sektor Telekomunikasi	CART dan ANN	Accuracy, sensitivitas, spesifisitas, presisi, F-score, dan koefisien korelasi Matthews	Model ANN menunjukkan performa yang lebih baik dibandingkan dengan model CART
(Patricia et al., 2023)	Sektor Perbankan	Gradient Boosting dan Random Forest	Correlation, Accuracy, F1-Score dan AUC-ROC	Model Gradient Boosting menunjukkan performa terbaik dengan tingkat kesalahan klasifikasi (misclassification rate) lebih rendah dibandingkan model Random Forest
(Sidiq & Anggraini, 2023)	Sektor Telekomunikasi	Logistic Regression, Random Forest, Support Vector Machine (SVM), Gradient Boosting, AdaBoost, dan XGBoost	Accuracy, precision, recall, and F1-score	XGBoost memberikan hasil terbaik

(Sholeha et al., 2024)	Sektor e-Commerce	XGBoost, LightGBM, dan CatBoost	Accuracy, precision, recall, F1-score, dan kurva ROC	Model XGBoost mencapai performa terbaik, diikuti oleh LightGBM dan CatBoost
(Taherkhani et al., 2023)	Sektor Pariwisata	metode hybrid berbasis text mining dan algoritma Random Forest	Accuracy, precision, recall, F1-score	metode text mining yang dikombinasikan dengan algoritma optimasi dan klasifikasi dapat meningkatkan akurasi prediksi churn di industri perhotelan

Preprocessing data adalah langkah penting dalam analisis data dan machine learning untuk membersihkan, memformat, dan menyiapkan data agar dapat diproses lebih lanjut oleh model. Proses ini bertujuan untuk memastikan kualitas data dan meningkatkan akurasi hasil model. Langkah-langkah umum dalam preprocessing data adalah: Data Cleaning (Pembersihan Data), Data Transformation (Transformasi Data), Feature Engineering (Rekayasa Fitur), Handling Imbalanced Data (Penyeimbangan Data yang Tidak Seimbang). Data cleaning adalah proses untuk mendeteksi, memperbaiki, atau menghapus data yang tidak akurat, tidak lengkap, atau tidak relevan dari dataset untuk meningkatkan kualitas data agar lebih siap digunakan dalam analisis atau pemodelan (García et al., 2015). Relief-F digunakan sebagai metode seleksi fitur untuk memilih fitur yang relevan dari dataset sebelum diterapkan pada model klasifikasi (Abdulsalam et al., 2022). Permutation Feature Importance (PFI) digunakan untuk mengidentifikasi fitur-fitur paling berpengaruh terhadap churn (Sidiq & Anggraini, 2023). Synthetic Minority Over-sampling Technique (SMOTE) dapat digunakan untuk menangani ketidakseimbangan kelas, namum pada oversampling cenderung menurunkan akurasi, meskipun dapat meningkatkan nilai precision dari model (Sidiq & Anggraini, 2023) (Geiler et al., 2022). Principal Component Analysis (PCA) dan Feature Selection dapat digunakan untuk menyederhanakan dan meningkatkan efisiensi data (Lohiya et al., n.d.). Pada penelitian lain juga dilakukan tuning hyperparameter untuk mendapatkan akurasi yang optimal dengan menggunakan metode GridSearchCV untuk Prediksi Nasabah Churn pada Industri Perbankan (Amalia & Asmunin, 2024). Yang dilakukan pada metode GridSeacrhCV adalah mencoba satu per satu kombinasi parameter, kemudian memvalidasi setiap kombinasinya. GridSearchCV dapat diterapkan secara maksimum apabila batas atas dan batas bawah dari masing-masing parameter diketahui (Amalia & Asmunin, 2024).

METODE PENELITIAN

Penelitian ini dilakukan untuk memprediksi churn pelanggan pada sektor perbankan berbasis supervised learning dengan pendekatan machine learning berdasarkan data historis, serta menggunakan beberapa tahapan seperti pada gambar 1. Terdapat 8 tahapan yang dilakukan, yaitu: merumuskan masalah penelitian, pengumpulan data, data preprocessing, data spliting, pembangunan model, evaluasi model, analisis, dan kesimpulan.

(1). Rumusan masalah. Penulis merumuskan masalah dengan menentukan model terbaik untuk churn customer prediction, menggunakan model ensemble learning dan bukan ensemble learning. Model yang digunakan adalah Logistic Regression dan Random Forest. Logistic Regression adalah model yang bukan ansamble, sedangkan Random Forest adalah model ansemble dengan voting.

(2). Pengumpulan data. Dataset yang digunakan pada penelitian adalah data mengenai customer churn pada sektor perbankan yang diambil dari laman https://www.kaggle.com/code/rainertimothy/customer-churn-prediction-using-ml.

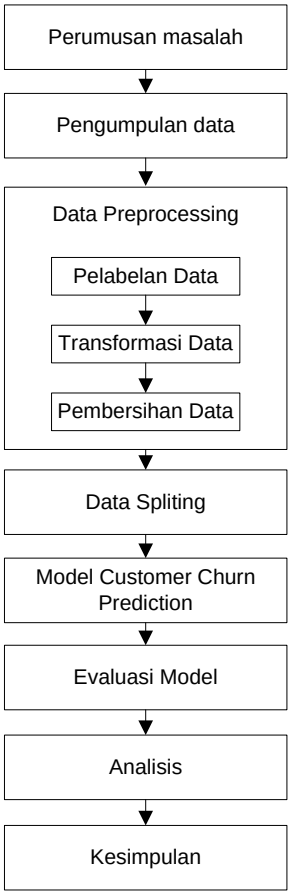
(3). Data Preprocessong. Sebelum data digunakan pada pebangunan model, dilakukan preprocessing terhadap dataset yang bertujuan untuk memastikan kualitas data dan meningkatkan akurasi hasil model. Langkah-langkah yang dilakuan dalam preprocessing data meliputi: pelabelan data dengan menentukan target variable (Churn: Yes/No), transformasi data, dan pembersihan data.

(4). Data Splitting. Tahap selanjutnya adalah data splitting, yaitu membagi data set menjadi data training (train set) dan data uji (test set), dengan proporsi 80%:20%.

(5). Model Customerc Churn Prediction. Penulis menggukana dua model machine learning untuk dapat memprediksi customer churn prediction, yaitu algoritma Logistic Regression untuk pemodelan linier dan Random Forest untuk peningkatan akurasi. Pada tahapan pembentukan model, penulis juga menentukan hyperparameter yang palling baik pada setiap model agar mendapatkan akurasi yang tinggi.

(6). Evaluasi Model.Penulis menggunakan Accuracy, Precision dan Recall, dan F1-Score untuk mengevaluasi model, kemudian menganalisis hasil penelitian berdasarkan nilai evaluasi tersebut.

(7). Kesimpulan. Tahap terakhir adalah menyimpulkan hasil penelitian untuk menjawab permasalahan penelitian. Pada tahapan ini, penulis juga memberikan rekomendasi kepada sector perbankan untuk dapat membuat tindak lanjut bagi pelanggan yang diprediksi akan melakukan churn.



(Sumber: Hasil penelitian, 2024)

Gambar 1. Metode Penelitian

HASIL DAN PEMBAHASAN

Perumusan Masalah

Berdasarkan Undang-undang No 10 tahun 1998 tentang perubahan atas Undang-Undang Nomor 7 Tahun 1992 tentang Perbankan, produk Atau layanan pada sektor perbankan antara lain adalah: Kliring, Transfer, Inkaso, Safe deposit box, Bank Garansi, Payment Point, Credit Card, Travellers Cheque, Surat Berharga, dan Automated Teller Machine (ATM). Setiap customer pada perbankan dapat menggunakan lebih dari satu layanan. Churn yang dilakukan customer pada sektor perbankan dapat dilakukan secara langung maupun tidak langsung. Permasalahan yang ada dalam lingkungan bisnis yang sangat kompetitif, termasuk sektor perbankan menjadikan retensi pelanggan sebagai prioritas karena biaya untuk mendapatkan pelanggan baru lebih tinggi daripada mempertahankan yang sudah ada. Diperlukan model dengan akurasi tinggi yang dapat memprediksi customer untuk melakukan churn. Hasil prediksi tersebut dapat digunakan oleh pengelola perbankan untuk melakukan tindakan lebih lanjut agar dapat mengurangi jumlah churn pada customer yang diprediksi melakukan churn.

Pengumpulan Data

Dataset yang digunakan pada penelitian ini diambil dari https://www.kaggle.com/code/rainertimothy/customer-churn-prediction-using-ml, memiliki 13 atribut seperti ditunjukan pada table 2.

Tabel 2. Variabel pada dataset

No	Nama atribut	Keterangan
1	Customer ID	Identitas yang bersifat unik bagi setiap custemer
2	Surname	Nama depan dan nama belakan dari custumer
3	Credit Score	Sebua nilai numerik yang merepresentasikan skor custemer's credit
4	Geography	Negara tempat custemer menetap (resides)
5	Gender	jenis kelamin customer (Male or Female)
6	Age	Usia custumer
7	Tenure	Jumlah tahun customer bergabung bersama bank
8	Balance	Saldo rekening customer
9	NumOfProducts	Jumlah layanan atau produk perbankan yang dimiliki atau digunakan oleh customer, seperti rekening tabungan, kartu kredit, pinjaman, atau deposito
10	HasCrCard	Berisi data biner yang menunjukkan kepemilikikan kartu kredit, 1 = yes berarti pelanggan memiliki kartu kredit, dan 0 = no berarti pelanggan tidak memiliki kartu kredit
11	IsActiveMember	Berisi data biner yang menunjukkan apakah customer merupakan anggota aktif di perbankan (1 = yes, 0 = no)
12	EstimatedSalary	Estimasi gaji customer
13	Exited	Atribut target yang berisi informasi apakah customer melakukan churn (1 = yes, 0 = no).

Gambar 2 menampilan 10 baris pertama dari 10002 baris dalam dataset yang digunakan pada penelitian. Dari gambar tersebut terdapat 1 atribut rowNumber dan 12 atribut bukan target dan 1 atribut churn.

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42.0	2	0.00	1	1.0	1.0	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41.0	1	83807.86	1	0.0	1.0	112542.58	0
2	3	15619304	Onio	502	France	Female	42.0	8	159660.80	3	1.0	0.0	113931.57	1
3	4	15701354	Boni	699	France	Female	39.0	1	0.00	2	0.0	0.0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43.0	2	125510.82	1	NaN	1.0	79084.10	0
5	6	15574012	Chu	645	Spain	Male	44.0	8	113755.78	2	1.0	0.0	149756.71	1
6	7	15592531	Bartlett	822	NaN	Male	50.0	7	0.00	2	1.0	1.0	10062.80	0
7	8	15656148	Obinna	376	Germany	Female	29.0	4	115046.74	4	1.0	0.0	119346.88	1
8	9	15792365	He	501	France	Male	44.0	4	142051.07	2	0.0	NaN	74940.50	0
9	10	15592389	H?	684	France	Male	NaN	2	134603.88	1	1.0	1.0	71725.73	0

(Sumber: Hasil penelitian, 2024)
Gambar 2. Tampilan 10 baris pertama dari dataset

Data Preprocessing

Sebelum dataset siap digunakan pada Model, dilakukan preprosesing data yaitu: menghilangkan tiga kolom yang tidak diperlukan (RowNumber, CustomerId, dan Surname) sehingga hanya terseisa 9 kolom, merubah nama kolom Exited menjadi Churn, dan memutasi missing value. Setelah dilakukan data preprocessing, terdapat 6380 data dengan 9 kolom. Sebelum data dapat digunakan pada model, dilakukan normalisasi dengan metode z-score normalization.

Split Dataset

Dataset yang digunakan pada penelitian ini termasuk ke dalam supervised learning, dan merupakan kasus klasifikasi. Untuk membuat model machine learning, penulis melakukan split dataset/pemisahan dataset dengan komposisi 70% data uji dan 30% data training. Gambar 3 berikut adalah code Python yang digunakan untuk splitting data.

```
X = df.drop('Churn', axis=1)
y = df['Churn']
Scaler = StandardScaler()
```

```
X = Scaler.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

X_train.shape, X_test.shape
```

(Sumber: Hasil penelitian, 2024)

Gambar 3. Code Python untuk model splitting datasest

Model Customer Churn Prediction.

Dalam membentuk model churn prediction, penulis menggunakan model Logistic Regression dan Random Forest, yang masing-masing model tersebut akan dievaluasi dengan nilai Accuracy, Precision, Recall, dan F1-Score.

A. Logistic Regression

Model pertama yang direkomendasikan adalah Logistic Regression. Logistic Regression adalah model yang umum digunakan pada kasus binominal, dimana tujuan utama dari model ini adalah mengkkasifikasikan data menjadi dua bagian. Meskipun bernama Regresion, namun model ini tidak digunakan pada masalah regresi, namun pada masalah klasifikasi. Cara kerja model Logistic Regression adalah sebagai berikut:

(1). Transformasi fungsi linear, dengan menggunakan persamaasn fungsi linear pada rumus (1).

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

Dimana z adalah nilai prediksi linear (logit), x_1, x_2, …, x_n adalah fitur input, dan β_0, β_1, …, β_n adalah koefisien model

(2). Transformasi Sigmoid atau logistic, digunakan untuk mengubah fungsi linear menjadi fungsi probabilitas. Rumus yang digunakan pada trasformasi sigmoid adalah: $p = \frac{1}{1+e^{-z}}$, dimana p adalah nilai probabilitas bahwa kejadian tersebut termasuk dalam kelas target.

Gambar 4 adalah code Python yang digunakan penulis untuk membentuk model Logistic Regression.

```
Log_reg = LogisticRegresion ()
Log_reg.fit(X_train, y_train)
y_test_pred = Log_reg.predict(X_test)
y_train_pred = Log_reg.predict(X_train)
```

(Sumber: Hasil penelitian, 2024)

Gambar 4. Code Python untuk membuat model Logistic Regression

B. Random Forest

Random Forest adalah model machine learning berbasis ensemble yang digunakan penulis pada kasus klasifikasi pada churn predection. Model ini bekerja dengan membuat banyak decision trees selama pelatihan dan menggabungkan hasilnya untuk membuat prediksi yang lebih stabil dan akurat. Tahapan yang dilakukan pada pembangunan model Random Forest adalah sebagai berikut:

(1). Bootstrap Sampling, yaitu membagi Dataset secara acak menjadi beberapa subset (dengan pengembalian), disebut sebagai bootstrap samples. Setiap subset digunakan untuk melatih satu decision tree.

(2). Pembuatan Decision Tree. Untuk setiap tree, hanya sebagian fitur yang dipilih secara acak untuk digunakan pada setiap node, dengan tujuan mencegah semua pohon menghasilkan keputusan yang sama dan meningkatkan keragaman dalam model.

(3). Agregasi hasil menggunakan metode voting mayoritas, yaitu dengan menggabungkan semua pohon untuk menentukan kelas akhir.

Dalam membentuk model Random Forest pada kasus churn prediction, penulis menentukan dua nilai hyperparameter yaitu: n_estimators=100, dan max_depth=10. Hyperparameter n_estimator menentukan jumlah decision tree dalam hutan. Semakin banyak pohon, semakin stabil hasilnya namun membutuhkan semakin banyak waktu untuk memprores. Hyperparameter max_depth menentukan kedalaman maksimum dari setiap pohon dengan tujuan untuk mencegah overfitting. Gambar 5adalah code Python untuk menetapkan hyperparameter pada model Random Forest.

```
rand_forest = RandomForestClassifier(n_estimators=100, max_depth=10)
rand_forest.fit(X_train, y_train)
y_test_pred = rand_forest.predict(X_test)
y_train_pred = rand_forest.predict(X_train)
```

(Sumber: Hasil penelitian, 2024)

Gambar 5. Code Python untuk membuat model Random Forest

Evaluasi Model

Penulis menggunajan metriks evaluasi untuk mengukur performa model dalam memprediksi data baru secara akurat dan andal. Evaluasi ini penting untuk memastikan bahwa model tidak hanya bekerja dengan baik pada data pelatihan tetapi juga dapat digeneralisasi untuk data yang belum pernah dilihat sebelumnya. Nilai-nilai yang digunakan pada metriks evaluasi meliputi: Accuracy, Precision dan Recall, dan F1-Score, yang dihitung beradasarkan hasil pengujian data testing terhadap model yang telah dibentuk oleh data traning. Terdapat 1914 data dari 6380 data pada dataset digunakan untuk mengevaluasi model. Accuracy menunjukkan persentase prediksi yang benar (baik positif maupun negatif) dibandingkan dengan total data yang diprediksi. Precision adalah kemampuan model untuk memprediksi positif dengan benar dibandingkan semua prediksi positif.

Gambar 5 menyajikan hasil confusion matrix dari dua model yang digunakan pada penelitian ini, yaitu Logistic Regression dan Random Forest.

		Actual Values	
		Positive	Negative
Predicted Values	Positive	1355	85
	Negative	360	116

(a) Model Logistic Regression

		Actual Values	
		Positive	Negative
Predicted Values	Positive	1377	61
	Negative	262	214

(b) Model Random Forest

(Sumber: Hasil penelitian, 2024)

Gambar 6. Hasil perhitungan Confusion Matrix dengan python

Dari nilai yang didapat dari hasil pengukuran Confusion Matrix, berikut adalah rumus yang digunakan untuk menghitung nilai evaluasi.

$$accuracy = \frac{True\_Posotive+True\_negatif}{Total\_Prediction} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1 - Score = 2.\frac{Precision.Recall}{Precision+Recall} \quad (5)$$

Tabel 1 menyajikan hasil perhitungan nilai evaluasi untuk masing-masing model yang digunakan, yaitu Logistic Regression dan Random Forest. Nilai evaluasi yang digunakan adalah Accuracy, Precision, Recall dan F1-Score seperti pada rumus (2) sampai rumus (4)

Tabel 1. Hasil Penrhitungan Nilai Evaluasi Model

Evaluasi	Logistic Regression	Random Forest
Accuracy	76,85%	12%
Precision	79%	84%
Recall	94%	96%
F1-Score	86%	90%

(Sumber: Hasil penelitian, 2024)

3.6. Analisis

Evaluasi yang digunakan oleh penulis adalah Accuracy, Precision dan Recall, dan F1-Score. Pada model Logistic Regression, didapat nilai Accuracy sebesar 76,85%, menunjukkan model berhasil memprediksi 76,85% dari seluruh data dengan benar, termasuk prediksi churn dan tidak churn. Nilai Precision sebear 79% menunjukkan dari semua prediksi churn, 79% benar-benar churn. Recall sebesar 94% menunjukkan model mampu menangkap 94% dari semua kejadian churn yang sebenarnya. F1-Score 86% menunjukkan keseimbangan yang cukup baik antara kemampuan model untuk mendeteksi churn dengan benar (recall) dan menghindari prediksi churn yang salah (precision).

Pada model Random Forest, nilai evaluasi yang didapat lebih besar dari model Logistic Regression, yaitu nilai Accuracy sebesar 83,12%, menunjukkan model berhasil memprediksi 83,12% dari seluruh data dengan benar, termasuk prediksi churn dan tidak churn. Nilai Precision sebear 84% menunjukkan dari semua prediksi churn, 79% benar-benar churn. Recall sebesar 96% menunjukkan model mampu menangkap 94% dari semua kejadian churn yang sebenarnya. F1-Score 90% menunjukkan keseimbangan yang cukup baik antara kemampuan model untuk mendeteksi churn dengan benar (recall) dan menghindari prediksi churn yang salah (precision).

KESIMPULAN

Model Logistic Regression ini menunjukkan performa yang cukup baik dengan recall yang sangat tinggi (94%) dan F1-Score yang solid (86%). Namun, precision (79%) dan akurasi (76,85%) masih perlu ditingkatkan, terutama jika false positives menjadi perhatian utama. Akurasi pada model Logistic Regression cukup baik, namun tidak mencerminkan kinerja yang sepenuhnya optimal pada dataset tidak seimbang. Model Random Forest menunjukkan performa yang lebih baik secara keseluruhan dibanding Logistic Regression, terutama dalam hal akurasi, precision, recall, dan F1-Score. Model Random Forest sangat cocok untuk masalah di mana recall tinggi diperlukan, seperti mendeteksi kejadian kritis yang membutuhkan sensitivitas tinggi. Rekomendasi penulis untuk perusahaan perbankan dalam rangka menekan angka churn pelanggan adalah melakukan analisa lebih lanjut terhadap faktor-faktor yang paling kuat mempengaruhi churn pelanggan. Selanjutnya perusahaan perbankan dapat menggunakan analisa SWOT untuk menentukan upaya tindak lanjut yang dapat mencegah churn pada pelanggan yang diprediksi melakukan churn.

Untuk penelitian selanjutnya mengenai churn prediction yang menggunakan model klasifikasi Logistic Regression perlu memastikan keseimbangan data dengan menghindari distribusi kelas yang tidak seimbang diantaranya dapat menggunakan metode Synthetic Minority Oversampling Technique dengan cara menciptakan data sintetis untuk kelas minoritas, sehingga distribusi kelas menjadi lebih seimbang, sehingga dapat meningkatkan performa model klasifikasi. Churn prediction dengan model Random Forest perlu dilakukan optimasi hyperparameter dan analisis feature importance untuk menentukan fitur mana yang paling berkontribusi dalam prediksi, sehingga performa Random Forest dapat ditingkatkan lebih lanjut.

REFERENSI

Abdulsalam, S. O., Arowolo, M. O., Saheed, Y. K., & Afolayan, J. O. (2022). Customer Churn Prediction in Telecommunication Industry Using Classification and Regression Trees and Artificial Neural Network Algorithms. *Indonesian Journal of Electrical Engineering and Informatics*, *10*(2), 431–440. https://doi.org/10.52549/ijeei.v10i2.2985

Aksama, M. C., & Wahyuniati, A. (2022). Prediksi Churn Nasabah Bank Menggunakan Klasifikasi NaÃ¯ve

Bayes dan ID3. *Jurnal Processor*, *17*(1), 9–18. https://doi.org/10.33998/processor.2022.17.1.1170

Amalia, N., & Asmunin. (2024). Optimasi Algoritma Random Forest dengan Hyperparameter Tuning Menggunakan GridSearchCV untuk Prediksi Nasabah Churn pada Industri Perbankan. *Manajemen Informasi*, *16*(1), 1–9.

Firmansyah, & Yulianto, A. (2021). Prediksi Customer Churn Pada Bisnis Retail Menggunakan Algoritma Naïve Bayes. *Remik*, *6*(1), 41–47. https://doi.org/10.33395/remik.v6i1.11196

García, S., Luengo, J., & Herrera, F. (2015). Instance selection. In *Intelligent Systems Reference Library* (Vol. 72). https://doi.org/10.1007/978-3-319-10247-4_8

Geiler, L., Affeldt, S., & Nadif, M. (2022). A survey on machine learning methods for churn prediction. *International Journal of Data Science and Analytics*, *14*(3), 217–242. https://doi.org/10.1007/s41060-022-00312-5

Herdian, R., & Girsang, A. S. (2023). Implementation of hybrid methods in data mining for Predicting customer churn in the telecommunications sector. *Jurnal Mantik*, *7*(1), 2685–4236.

Hussain, S. A., Roushdy, M., Galal, A., & Haggag, R. M. Y. (2023). Prediction of Prediction Customer Churn In The Banking Sector: Review. *Journal of Southwest Jioautong University*, *58 No 4*, 719–735. https://doi.org/10.35741

Lohiya, S., Salunkhe, O., Parshionikar, S., & Thatte, S. (n.d.). *Telecom Customer Churn Prediction: A Review*. *8*(7), 158–170.

Naderalvojoud, B., & Hernandez-Boussard, T. (2023). Improving machine learning with ensemble learning on observational healthcare data. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2023*, 521–529.

Patricia, M., Oetama, R. S., & Prasetiawan, I. (2023). Unveiling Churn Prediction At Bank Ivory. *Jurnal Informatika Dan Teknik Elektro Terapan*, *11*(3s1). https://doi.org/10.23960/jitet.v11i3s1.3394

Pondel, M., Wuczyński, M., Gryncewicz, W., Łysik, Ł., Hernes, M., Rot, A., & Kozina, A. (2021). Deep learning for customer churn prediction in e-commerce decision support. *Business Information Systems*, *1*(November), 3–12. https://doi.org/10.52825/bis.v1i.42

Sholeha, S. H., Faid, M., & Yaqin, M. A. (2024). Prediksi Perpindahan Pelanggan Pada Toko Online Menggunakan Metode Tree-Based Gradient Boosted Models. *Journal of Computer System and Informatics*, *5*(3), 605–614. https://doi.org/10.47065/josyc.v5i3.5215

Sidiq, M. M., & Anggraini, D. (2023). Analysis and Classification of Customer Churn Using Machine Learning Models. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *7*(6), 1253–1259. https://doi.org/10.29207/resti.v7i6.4933

Taherkhani, L., Daneshvar, A., Amoozad Khalili, H., & Sanaei, M. R. (2023). Analysis of the Customer Churn Prediction Project in the Hotel Industry Based on Text Mining and the Random Forest Algorithm. *Advances in Civil Engineering*, *2023*. https://doi.org/10.1155/2023/6029121