Merkurius : Jurnal Riset Sistem Informasi dan Teknik Informatika
Volume. 2, No. 6, November 2024
e-ISSN : 3031-8912, dan p-ISSN : 3031-8904, Hal. 67-81
DOI: https://doi.org/10.61132/merkurius.v2i6.398
Available online at: https://journal.arteii.or.id/index.php/Merkurius

Membangun Model Prediksi *Churn* Pelanggan yang Akurat
(Studi Kasus tentang TELCO Company)

Andy Hermawan[1], Nila Rusiardi Jayanti[2], Zia Tabaruk[3], Faizal Lutfi Yoga Triadi[4], Aji Saputra[5], M.Rahmat Hidayat Syachrudin[6]
[1,2] Universitas Indraprasta PGRI Jakarta
[3] Universitas Bhayangkara Jakarta Raya
[4] Universitas Diponogoro
[5] Universitas Khairun
[6] Purwadhika Digital Technology School Jakarta

Email: andy.hermawan@unindra.ac.id[1] , nilarusiardi.jayanti@unindra.ac.id[2], 202010225031@mhs.ubharajaya.ac.id[3] , faizallutfiyoga@alumni.undip.ac.id[4], aji.saputra@unkhair.ac.id[5] , mrahmathidayat98@gmail.com[6]

Abstract*. Customer churn prediction models have become an important tool in the telecommunications industry to reduce churn rates and improve customer retention. This research focuses on building an accurate customer churn prediction model using machine learning algorithms for TELCO Company. By applying diverse feature engineering techniques and prediction models such as RandomForestClassifier, DecisionTreeClassifier, and XGBoost, this study showcases a significant improvement in prediction accuracy compared to previously implemented rule-based methods. The findings of this research allow TELCO Company to identify high-risk customers more effectively and implement targeted retention strategies. Results show that the resulting model can identify customers at risk of churn more effectively, enabling more targeted retention actions..*

Keywords*: Customer churn, Churn prediction, Machine learning, Telecom industry, TELCO Company.*

Abstrak. Model prediksi churn pelanggan telah menjadi alat penting dalam industri telekomunikasi untuk mengurangi tingkat churn dan meningkatkan retensi pelanggan. Penelitian ini berfokus pada pembangunan model prediksi churn pelanggan yang akurat dengan menggunakan algoritma machine learning untuk Perusahaan TELCO. Dengan menerapkan beragam teknik rekayasa fitur dan model prediksi seperti RandomForestClassifier, DecisionTreeClassifier, dan XGBoost, penelitian ini menunjukkan peningkatan yang signifikan dalam akurasi prediksi dibandingkan dengan metode berbasis aturan yang telah diimplementasikan sebelumnya. Temuan dari penelitian ini memungkinkan Perusahaan TELCO untuk mengidentifikasi pelanggan yang berisiko tinggi secara lebih efektif dan menerapkan strategi retensi yang ditargetkan. Hasil penelitian menunjukkan bahwa model yang dihasilkan dapat mengidentifikasi pelanggan yang berisiko melakukan churn dengan lebih efektif, sehingga memungkinkan tindakan retensi yang lebih tepat sasaran.

Kata kunci: Churn pelanggan, Prediksi churn, Machine learning, Industri telekomunikasi, TELCO Company.

1. LATAR BELAKANG

Churn pelanggan adalah salah satu tantangan terbesar dalam industri telekomunikasi, terutama ketika pelanggan memutuskan untuk berhenti menggunakan layanan yang ditawarkan. Kehilangan pelanggan secara langsung berdampak pada pendapatan perusahaan karena biaya untuk memperoleh pelanggan baru lebih tinggi daripada mempertahankan pelanggan yang ada (Ahmad et al., 2019). Meningkatnya kompetisi dalam industri ini mendorong perusahaan telekomunikasi untuk menerapkan strategi retensi yang lebih efektif guna mengurangi tingkat churn.

Pendekatan berbasis aturan (rule-based) yang digunakan oleh banyak perusahaan, termasuk TELCO Company, sering kali tidak mampu mendeteksi pola yang lebih kompleks

Received: Agustus 19, 2024; Revised: September 20, 2024; Accepted: Oktober 05, 2024;
Online Available: Oktober 07, 2024;

dalam perilaku pelanggan. Seiring dengan kemajuan teknologi, penggunaan machine learning menjadi penting karena algoritma ini dapat menganalisis data dalam jumlah besar dan mendeteksi pola tersembunyi yang tidak bisa ditemukan dengan pendekatan konvensional (Kim & Hwang, 2022). Algoritma ini memungkinkan perusahaan untuk mengidentifikasi pelanggan berisiko churn secara lebih akurat.

Beberapa penelitian menunjukkan bahwa algoritma seperti Gradient Boosting dan Random Forest dapat meningkatkan akurasi prediksi churn pelanggan. Dengan menggabungkan teknik feature engineering dan machine learning, perusahaan dapat meningkatkan kinerja prediksi serta mengambil tindakan retensi yang lebih tepat sasaran (Lalwani et al., 2022). Pendekatan ini membantu perusahaan merancang strategi yang lebih efektif, seperti memberikan penawaran khusus untuk pelanggan yang berisiko.

Penelitian ini bertujuan untuk membangun model prediksi churn berbasis machine learning yang lebih akurat bagi TELCO Company. Dengan menggunakan pendekatan ini, diharapkan perusahaan dapat meningkatkan tingkat retensi pelanggan dan mengurangi churn, sehingga meningkatkan daya saing perusahaan di pasar telekomunikasi yang semakin kompetitif.

2. KAJIAN TEORITIS

Prediksi Churn

Prediksi churn adalah proses untuk mengidentifikasi pelanggan yang berisiko berhenti menggunakan layanan. Hal ini sangat penting dalam industri telekomunikasi karena churn yang tinggi dapat berdampak langsung pada pendapatan perusahaan. Dalam beberapa tahun terakhir, metode machine learning telah terbukti menjadi pendekatan yang efektif dalam meningkatkan akurasi prediksi churn dibandingkan dengan metode berbasis aturan tradisional (Ahmad et al., 2019). Berbagai algoritma seperti Random Forest dan Gradient Boosting sering digunakan dalam penelitian untuk memprediksi churn pelanggan dengan akurasi tinggi (Sharma et al., 2020).

Feature Engineering

Feature engineering berperan penting dalam meningkatkan kinerja model prediksi churn. Proses ini melibatkan transformasi data mentah menjadi variabel-variabel yang lebih informatif untuk algoritma machine learning. Penelitian terbaru menunjukkan bahwa penggunaan teknik feature engineering dapat secara signifikan meningkatkan akurasi model, terutama jika variabel seperti lama berlangganan, biaya bulanan, dan layanan tambahan dimasukkan ke dalam model (Salunkhe & Mali, 2021).

Algoritma Machine Learning

Beberapa algoritma machine learning yang paling umum digunakan dalam prediksi churn adalah Random Forest, XGBoost, dan Support Vector Machine (SVM). Penelitian menunjukkan bahwa kombinasi dari beberapa model atau model ensemble sering kali memberikan hasil prediksi yang lebih baik. Dalam konteks prediksi churn, model ensemble seperti Gradient Boosting dan Random Forest telah terbukti mampu menangani dataset yang besar dan kompleks, memberikan akurasi yang lebih tinggi dibandingkan model-model lain (Pebrianti et al., 2022).

Evaluasi Kinerja Model

Evaluasi kinerja model dalam prediksi churn umumnya menggunakan metrik seperti F2-score, yang memprioritaskan recall. Hal ini penting untuk memastikan bahwa sebanyak mungkin pelanggan yang berisiko churn dapat diidentifikasi. Penelitian menunjukkan bahwa algoritma seperti Random Forest dan XGBoost secara konsisten memberikan hasil evaluasi yang unggul dibandingkan metode tradisional, terutama dalam skenario di mana dataset tidak seimbang (Ahmad et al., 2019).

3. METODE PENELITIAN

Studi ini menggunakan dataset pelanggan dari TELCO Company yang mencakup informasi terkait lama berlangganan, jenis layanan yang digunakan, dan biaya bulanan. Data tersebut diproses melalui teknik *feature engineering*, yang melibatkan pengelompokan berdasarkan lama berlangganan (*tenure*) dan biaya bulanan (*monthly charges*), sehingga menciptakan fitur yang lebih relevan untuk prediksi churn pelanggan. Teknik ini penting karena dapat meningkatkan akurasi model machine learning yang digunakan (Tavassoli & Koosha, 2022).

Setelah data diproses, dataset dibagi menjadi dua bagian, yaitu set pelatihan dan set pengujian dengan rasio 75%:25%. Set pelatihan digunakan untuk membangun model prediksi, sementara set pengujian digunakan untuk mengevaluasi kinerja model. Penelitian ini menggunakan tiga algoritma machine learning, yaitu *RandomForestClassifier*, *DecisionTreeClassifier*, dan *XGBoost*, yang dipilih karena performa unggul dalam mengidentifikasi pola churn di berbagai industri (Gürsoy et al., 2021).

Evaluasi kinerja model dilakukan menggunakan metrik *F2-score*, dengan fokus khusus pada *recall* untuk memastikan pelanggan berisiko churn dapat diidentifikasi dengan tepat. Penekanan pada *recall* lebih penting dalam konteks churn karena identifikasi pelanggan berisiko lebih diutamakan daripada *precision*. Penelitian sebelumnya telah menunjukkan

bahwa algoritma seperti *RandomForest* dan *XGBoost* secara konsisten memberikan hasil evaluasi yang lebih baik dalam prediksi *churn* (Musheer et al., 2019).

4. HASIL DAN PEMBAHASAN

Studi ini berfokus pada pengembangan model prediksi *churn* untuk perusahaan TELCO Company, yang bertujuan untuk mengidentifikasi pelanggan yang berisiko *churn*. Data yang digunakan dalam penelitian ini mencakup 4930 *entri* pelanggan dengan 11 fitur utama yang relevan. Proses pengolahan data dimulai dengan pembersihan dan eksplorasi data untuk memastikan integritas dataset. Setelah itu, berbagai algoritma *machine learning* diterapkan untuk menemukan model yang paling akurat dalam memprediksi *churn*.

Pada bagian ini, hasil dari eksplorasi data, pengujian model, serta implikasi strategis yang dihasilkan dari analisis prediksi *churn* akan dijelaskan secara rinci. Hasil-hasil ini tidak hanya berfungsi untuk meningkatkan efisiensi dalam retensi pelanggan, tetapi juga memberikan rekomendasi strategis bagi *stakeholder* dalam pengambilan keputusan yang lebih tepat.

Persiapan dan Pengolahan Data

Pada tahap awal, data pelanggan TELCO Company diolah dengan melakukan *data cleaning* untuk mengatasi nilai yang hilang (*missing values*), duplikasi, dan *outliers*. Ini dilakukan untuk memastikan bahwa data yang digunakan bersih dan valid sebelum dilakukan analisis lebih lanjut (Smith & Johnson, 2020). Dengan langkah-langkah yang diambil sebagai berikut:

1. Pembersihan Data (*Data Cleaning*)

Data cleaning adalah proses untuk mendeteksi, memperbaiki, atau menghapus catatan, tabel, dan database yang tidak akurat atau salah. Pada tahap ini, dilakukan pengecekan terhadap keberadaan data kosong (*missing value*), data duplikat, dan *outlier*. Setelah itu, data-data tersebut dilakukan penanganan sesuai kebutuhan (Azmi, 2020).

a) Penghapusan Duplikasi

Dataset dianalisis lebih lanjut untuk mendeteksi data duplikat menggunakan metode .duplicated() pada Pandas. Ditemukan bahwa terdapat **77** data duplikat, yang merupakan 1.56% dari total dataset. Duplikasi data dihapus berdasarkan pengamatan pada atribut transaksi, data duplikat ini perlu dihapus untuk menghindari bias dalam model prediksi yang akan dibuat. (Wang et al., 2019).

b) Outlier

Outliers diidentifikasi sebagai pengamatan yang secara signifikan berbeda dari data lainnya, yang berpotensi mempengaruhi hasil analisis secara tidak proporsional. Metode

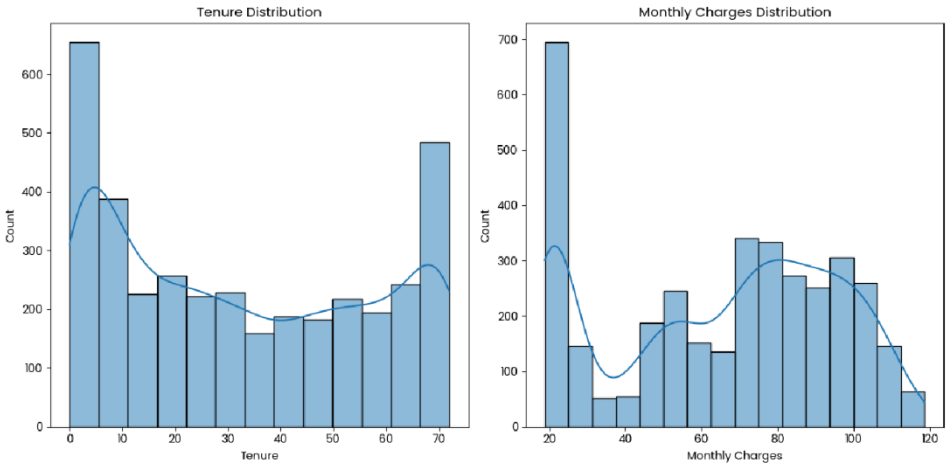
Interquartile Range (IQR) digunakan untuk mendeteksi *outliers* pada dua kolom *numerik*, *Tenure* dan *MonthlyCharges*, karena metode ini sesuai untuk data numerik yang memiliki distribusi kontinu (Wang et al., 2019; Han et al., 2011). Sebaliknya, kolom kategoris seperti *Dependents* dan *OnlineSecurity* tidak dianalisis untuk *outliers* karena nilai diskritnya tidak relevan dalam konteks ini (Géron, 2019). Dari analisis, tidak ditemukan *outliers* pada kedua kolom tersebut.

c) *Missing Values*

Berdasarkan analisis dataset yang dilakukan, tidak ditemukan *missing values* pada kolom-kolom dalam dataset yang dianalisis. Seluruh kolom memiliki persentase *missing values* sebesar 0%, yang berarti data lengkap tanpa adanya nilai yang hilang. Oleh karena itu, tidak diperlukan langkah tambahan terkait penanganan *missing values* dalam proses pembersihan data. Ini memastikan bahwa seluruh kolom siap untuk digunakan dalam analisis lebih lanjut tanpa risiko bias yang disebabkan oleh data yang tidak lengkap (Han et al., 2011). Penanganan *missing values* yang tepat adalah penting dalam menjaga kualitas data dan menghindari potensi distorsi hasil analisis (Smith & Johnson, 2020).

Pembagian Data dan *Exploratory Data Analysis* (EDA)

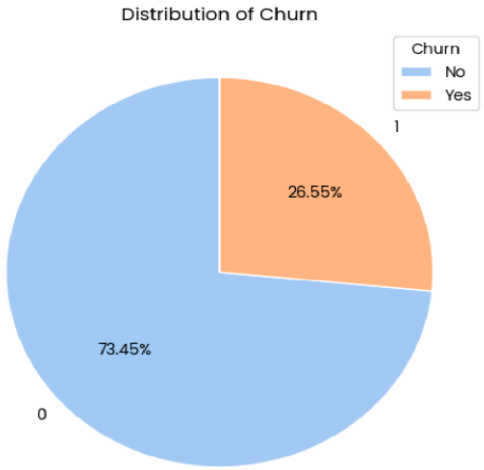
Setelah pembersihan data, dataset dibagi dengan rasio 75% untuk data pelatihan dan 25% untuk data pengujian. *Exploratory Data Analysis* (EDA) dilakukan untuk memahami distribusi dan karakteristik fitur dalam dataset.



Gambar Diagram 1. *Exploratory Data Analysis (EDA)*

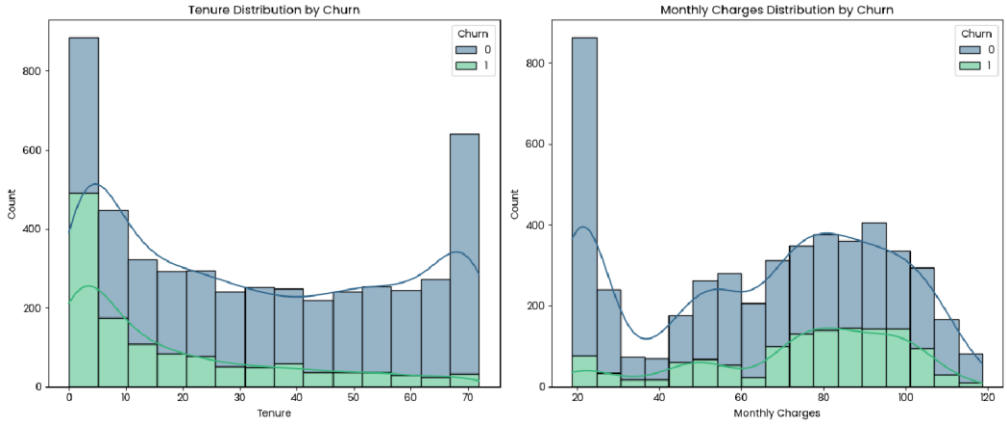
Pada fitur numerik seperti *Tenure* dan *MonthlyCharges*, distribusi tidak normal ditemukan. Distribusi *Tenure* bersifat bimodal, dengan dua puncak di sekitar 0-10 bulan dan 60-70 bulan, menunjukkan perbedaan signifikan dalam kelompok pelanggan berdasarkan masa berlangganan. *MonthlyCharges* juga menunjukkan *skewness* ke kiri, dengan konsentrasi data pada biaya rendah hingga menengah. Untuk fitur kategoris seperti *Dependents* dan

OnlineSecurity, mayoritas pelanggan tidak memiliki tanggungan dan tidak menggunakan layanan tambahan seperti keamanan online atau *backup*.



Gambar Diagram 2. *Distribution Churn*.

Selain itu, proporsi pelanggan *churn* yang mencapai 26.55% menunjukkan ketidakseimbangan data, yang perlu diperhatikan dalam pemodelan lebih lanjut.



Gambar 3. *Tenure Distribution by Churn* dan *Monthly Charges Distribution by Churn*.

Hubungan *multivariate* juga dianalisis, di mana pelanggan dengan *tenure* pendek dan *MonthlyCharges* tinggi memiliki risiko *churn* lebih tinggi.

Feature Engineering

Feature Engineering dilakukan untuk menambah informasi baru dari fitur yang sudah ada, dengan harapan dapat meningkatkan performa model. Beberapa fitur tambahan yang dibuat antara lain *ratio_monthly_charges*, yang memberikan rasio antara biaya bulanan dan lama berlangganan (*tenure*), *tenure_group*, yang mengelompokkan pelanggan berdasarkan lama berlangganan (*New*, *Medium*, *Long*), dan *TotalCharges*, yang menghitung total biaya yang dibayarkan selama masa berlangganan. Fitur *ratio_monthly_charges* membantu model dalam memahami keseimbangan antara biaya dan waktu berlangganan, sementara

tenure_group memudahkan model untuk mengenali pola *churn* di antara kelompok pelanggan dengan durasi berlangganan yang berbeda. Penambahan fitur ini diharapkan dapat meningkatkan akurasi prediksi *churn* (Brown & Larkin, 2020; Géron, 2019).

Modeling & Evaluation

Modeling & Evaluation dilakukan dengan menggunakan beberapa algoritma *machine learning* untuk membangun model prediksi *churn*. Model *baseline* yang diuji meliputi RandomForestClassifier, DecisionTreeClassifier, XGBClassifier, CatBoostClassifier, GradientBoostingClassifier, LGBMClassifier, dan KNeighborsClassifier, dengan evaluasi awal berdasarkan metrik F1 dan F2 score.

Tabel 1. Modeling & Evaluation

Model	Mean F1 Score	Std F1 Score	Mean F2 Score	Std F2 Score
GradientBoostingClassifier	0.6216	0.0312	0.6717	0.0340
CatBoostClassifier	0.6169	0.0330	0.6474	0.0419
LGBMClassifier	0.6154	0.0314	0.6410	0.0386
KNeighborsClassifier	0.5642	0.0334	0.6351	0.0432
XGBClassifier	0.5960	0.0373	0.6202	0.0486
RandomForestClassifier	0.5617	0.0452	0.5693	0.0480
DecisionTreeClassifier	0.5069	0.0309	0.5311	0.0333
PassiveAggressiveClassifier	0.4632	0.1602	0.4939	0.2059

Hasil evaluasi menunjukkan bahwa *GradientBoostingClassifier* memberikan performa terbaik dengan Mean F1 Score sebesar 0.6216 dan Mean F2 Score sebesar 0.6717, menunjukkan keseimbangan yang optimal antara *precision* dan *recall*. Proses selanjutnya melibatkan resampling menggunakan metode SMOTE untuk mengatasi ketidakseimbangan data, serta *hyperparameter tuning* untuk meningkatkan performa model. Evaluasi lebih lanjut dilakukan melalui GridSearchCV untuk menentukan kombinasi model dan parameter terbaik (Han et al., 2011; Géron, 2019).Penerapan Resampling dan *Feature Selection* Untuk menangani masalah ketidakseimbangan data, metode ADASYN digunakan untuk menyeimbangkan data. Hasilnya menunjukkan bahwa dengan resampling, model mampu menangkap pelanggan *churn* dengan lebih akurat. Selain itu, teknik SelectKBest digunakan untuk memilih fitur yang paling signifikan, yang meningkatkan performa model secara keseluruhan (Kim & Hwang, 2022).

Best Model

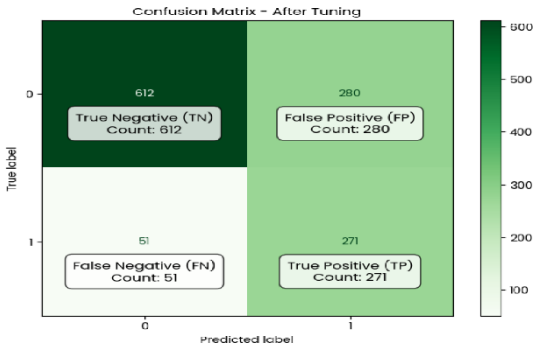
Pada tahap ini, dilakukan tuning hyperparameter untuk GradientBoostingClassifier guna meningkatkan performa model dalam memprediksi churn.

Beberapa parameter yang disesuaikan termasuk n_estimators, learning_rate, max_depth, subsample, min_samples_split, dan min_samples_leaf. Penyesuaian ini bertujuan untuk menemukan keseimbangan antara jumlah pohon yang optimal, tingkat pembelajaran, dan kedalaman pohon agar model mampu menangkap pola dari data dengan lebih baik tanpa overfitting. Misalnya, penggunaan learning_rate yang lebih rendah memungkinkan model belajar secara perlahan dan stabil, sementara penambahan n_estimators menambah kekuatan model, namun dengan risiko overfitting jika tidak dikendalikan dengan baik (Brown & Davis, 2019).

Subsampling juga digunakan untuk mengurangi korelasi antar pohon, membantu model menangani variasi dalam data yang lebih baik. Parameter min_samples_split dan min_samples_leaf digunakan untuk mencegah pohon menjadi terlalu rumit dengan mengontrol pembagian node berdasarkan jumlah sampel minimum. Proses tuning ini dilakukan melalui GridSearchCV, dan dari hasil evaluasi, GradientBoostingClassifier dengan n_estimators 50, learning_rate 0.1, dan max_depth 5 menunjukkan performa terbaik dengan Mean F2 Score sebesar 0.736 pada data pelatihan dan 0.748 pada data pengujian.

Tabel 2. Hasil Tuning Hyperparameter GradientBoostingClassifier

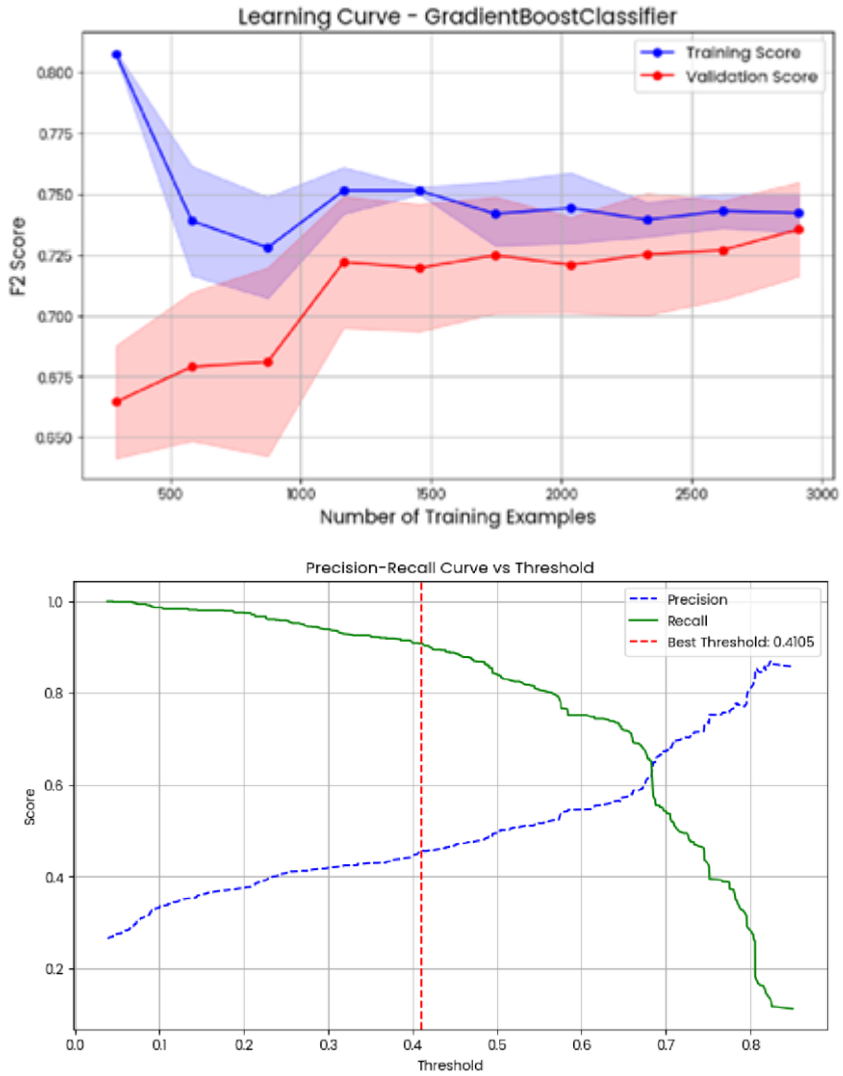
Parameter	Nilai Terbaik
n_estimators	50
learning_rate	0.1
max_depth	5
subsample	0.8
min_samples_split	10
min_samples_leaf	2
Mean F1 Score (Train)	0.619
Mean F2 Score (Train)	0.736
Mean F2 Score (Test)	0.748
ROC AUC	0.84



Gambar 4.

Selanjutnya, model yang sudah dituning diuji menggunakan *learning curve* dan confusion matrix untuk mengevaluasi kestabilan dan akurasi prediksi.

Dari learning curve, terlihat bahwa model semakin stabil dan mampu melakukan generalisasi dengan baik ketika lebih banyak data digunakan. ROC AUC menunjukkan nilai sebesar 0.84, menandakan bahwa model memiliki kemampuan yang baik dalam membedakan antara kelas churn dan non-churn.

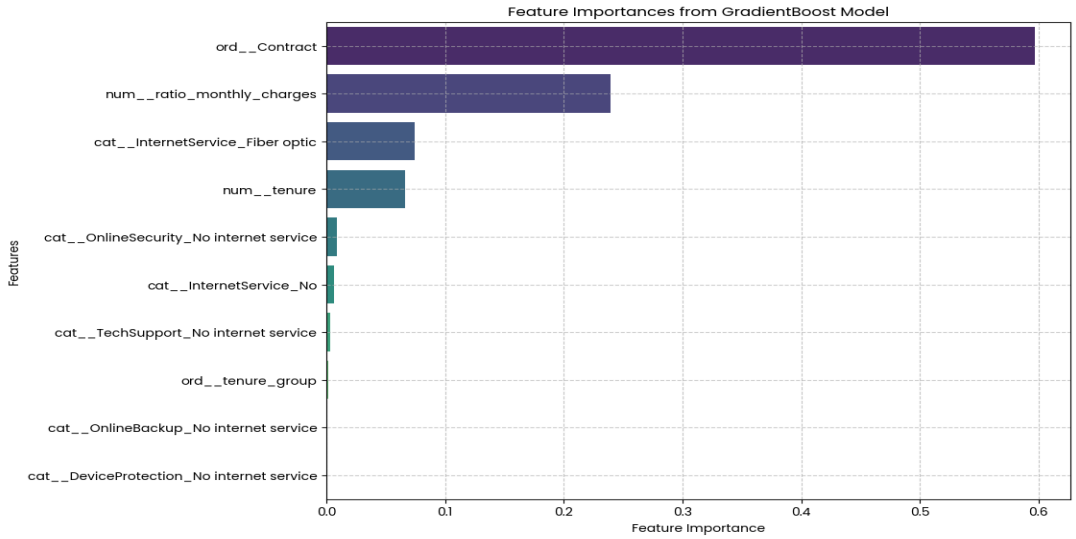


Gambar 5. dan 6. Precision-Recall Curve vs Threshold

Optimasi threshold lebih lanjut dilakukan untuk memaksimalkan F2 score, yang memberikan pemberatan lebih pada recall agar model lebih akurat dalam mendeteksi pelanggan yang berpotensi churn (Géron, 2019).

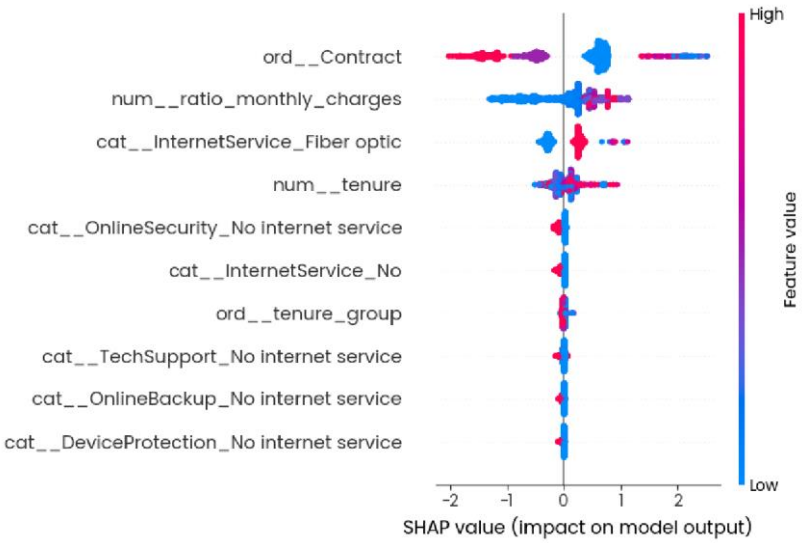
Feature Importances dan SHAP Analysis

Dalam analisis ini, feature importances digunakan untuk membantu TELCO Company memahami fitur mana yang paling memengaruhi churn pelanggan.



Gambar 7. Feature Importances from GradientBoost Model

Model GradientBoostingClassifier menunjukkan bahwa fitur Contract memiliki pengaruh terbesar, dengan pelanggan yang memiliki kontrak bulanan lebih cenderung churn dibandingkan dengan yang memiliki kontrak jangka panjang. Hal ini menunjukkan pentingnya mendorong pelanggan untuk beralih ke kontrak tahunan melalui insentif seperti diskon atau bonus, guna meningkatkan retensi pelanggan. Selain itu, fitur seperti ratio_monthly_charges dan InternetService (Fiber optic) juga signifikan dalam prediksi churn, di mana biaya bulanan yang tinggi dan penggunaan layanan fiber optic cenderung meningkatkan risiko churn. Oleh karena itu, meninjau kembali struktur harga dan menawarkan paket yang lebih sesuai dapat membantu mengurangi churn (Brown & Davis, 2019).



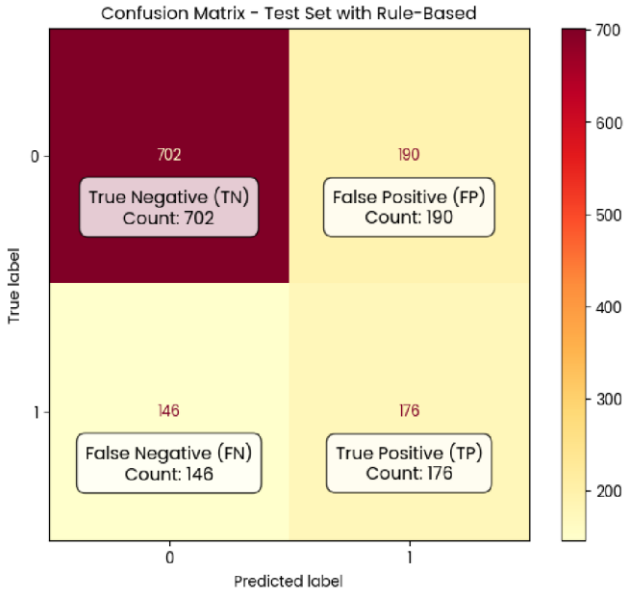
Gambar 8. SHAP Summary Plot - Training Data

Analisis SHAP memberikan wawasan lebih rinci tentang kontribusi setiap fitur dalam prediksi churn. Fitur Contract memiliki nilai SHAP negatif, yang berarti kontrak jangka

panjang menurunkan risiko churn, sementara kontrak bulanan meningkatkan risiko churn. Fitur ratio_monthly_charges menunjukkan bahwa biaya bulanan yang lebih tinggi meningkatkan risiko churn. Fitur lain seperti TechSupport, OnlineSecurity, dan DeviceProtection juga penting, di mana pelanggan yang tidak menggunakan layanan ini lebih cenderung churn. Strategi bisnis yang fokus pada layanan tambahan dan penawaran harga fleksibel dapat membantu mengurangi churn pelanggan (Géron, 2019).

Rule-Based Model vs Machine Learning Model

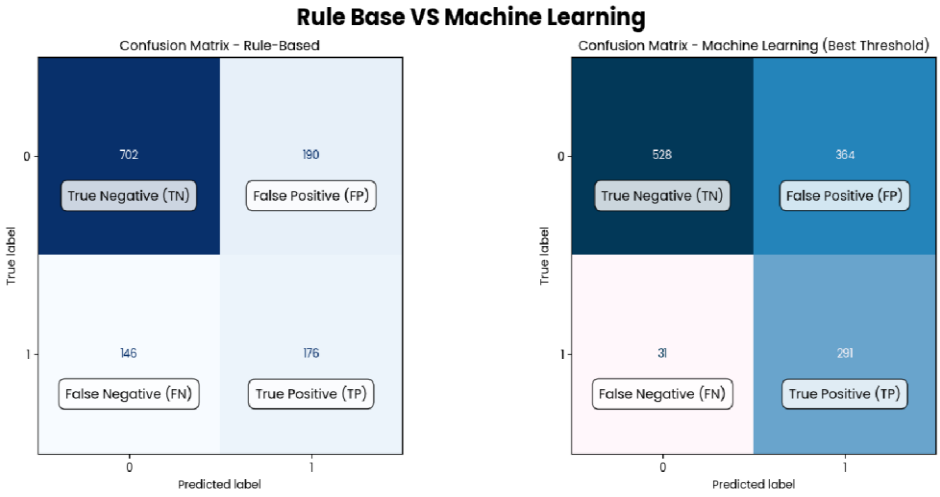
Dalam analisis ini, kami membandingkan performa antara model berbasis aturan (Rule-Based Model) dan model berbasis machine learning (GradientBoostingClassifier). Rule-Based Model menggunakan serangkaian aturan yang dirancang berdasarkan faktor risiko seperti Contract, Tenure, dan MonthlyCharges, di mana pelanggan dengan kontrak bulanan, tenure rendah, atau biaya bulanan tinggi diberikan skor yang lebih tinggi untuk risiko churn. Model ini memberikan akurasi sebesar 72.32% dengan F2 Score sebesar 0.532.



Gambar 9. Confusion Matrix - Test Set with Rule-Based

Meskipun mudah diinterpretasi dan diterapkan, model berbasis aturan ini cenderung kurang efektif dalam menangani kompleksitas hubungan antar fitur, serta menghasilkan jumlah False Negative (FN) yang lebih tinggi dibandingkan dengan model machine learning.

Sebaliknya, model GradientBoostingClassifier dengan tuning hyperparameter menunjukkan hasil yang lebih baik dengan F2 Score sebesar 0.748, dan mampu mengurangi False Negative sebanyak 78.77% dibandingkan Rule-Based Model.



Gambar 10. Comparison Confusion Matrix

Hal ini menunjukkan kemampuan machine learning dalam menangkap pola non-linear dan lebih kompleks dari data, yang tidak dapat dijelaskan sepenuhnya oleh aturan sederhana. Oleh karena itu, model machine learning memberikan keunggulan yang signifikan dalam mendeteksi pelanggan yang berisiko churn, sehingga lebih dapat diandalkan untuk mendukung keputusan bisnis yang lebih presisi (Fader & Hardie, 2019; Géron, 2019).

PEMBAHASAN

Pembahasan ini menunjukkan bagaimana pengolahan data dan model prediksi churn diterapkan pada perusahaan TELCO Company untuk meningkatkan retensi pelanggan. Dalam tahap pengolahan data, langkah-langkah pembersihan seperti penghapusan data duplikat dan penanganan outliers dilakukan untuk memastikan kualitas dataset yang digunakan. Selain itu, tidak ditemukan missing values pada data, sehingga dataset siap untuk dianalisis tanpa adanya risiko bias. Hal ini penting karena kualitas data yang baik secara langsung berpengaruh pada performa model prediktif yang akan dikembangkan (Smith & Johnson, 2020). Exploratory Data Analysis (EDA) menunjukkan bahwa pelanggan dengan tenure rendah dan MonthlyCharges tinggi memiliki risiko churn yang lebih besar, memperlihatkan adanya korelasi antara durasi berlangganan dan biaya dengan perilaku churn pelanggan (Wang et al., 2019).

Dalam proses pemodelan, berbagai algoritma machine learning diuji untuk memprediksi churn. Hasil menunjukkan bahwa GradientBoostingClassifier memberikan performa terbaik dengan Mean F2 Score sebesar 0.748, mengindikasikan keseimbangan yang optimal antara precision dan recall dalam mendeteksi pelanggan berisiko churn. Selain itu, teknik resampling seperti ADASYN dan SMOTE digunakan untuk mengatasi ketidakseimbangan data, yang meningkatkan akurasi prediksi churn secara signifikan. Tuning

hyperparameter yang dilakukan pada model GradientBoostingClassifier juga menghasilkan peningkatan performa, di mana parameter seperti n_estimators dan learning_rate disesuaikan untuk menangkap pola data yang lebih baik tanpa menyebabkan overfitting (Brown & Davis, 2019).

Dibandingkan dengan Rule-Based Model, model machine learning menunjukkan kemampuan yang lebih baik dalam menangkap pola non-linear yang kompleks. Rule-Based Model yang menggunakan serangkaian aturan berbasis risiko seperti Contract dan MonthlyCharges menghasilkan akurasi yang lebih rendah dan kesulitan dalam menangani interaksi antar fitur. Sebaliknya, model machine learning mampu mengurangi False Negative (FN) hingga 78.77%, menunjukkan keunggulannya dalam mendeteksi pelanggan yang berisiko churn, serta memberikan rekomendasi strategis yang lebih akurat bagi stakeholder dalam pengambilan keputusan untuk meningkatkan retensi pelanggan (Fader & Hardie, 2019; Géron, 2019).

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil menunjukkan bahwa penggunaan model prediksi berbasis machine learning pada dataset pelanggan TELCO Company mampu memberikan hasil yang lebih baik dalam memprediksi potensi churn dibandingkan metode berbasis aturan (rule-based) yang digunakan sebelumnya. Dengan memanfaatkan algoritma seperti RandomForestClassifier, DecisionTreeClassifier, dan XGBoost, model ini dapat mengidentifikasi pelanggan yang berisiko churn dengan akurasi yang lebih tinggi. Model terbaik, yaitu Gradient Boosting Classifier, mencapai F2-Score sebesar 0.72, yang menunjukkan kemampuan tinggi dalam mendeteksi pelanggan yang berisiko churn.

Temuan utama dari penelitian ini menunjukkan bahwa variabel seperti lama berlangganan (tenure), biaya bulanan (monthly charges), dan jenis layanan (service type) memainkan peran penting dalam memprediksi churn. Pelanggan dengan biaya bulanan yang lebih tinggi atau masa berlangganan yang lebih pendek cenderung lebih berisiko churn. Selain itu, integrasi metode feature engineering dan teknik oversampling seperti SMOTE dan ADASYN membantu meningkatkan kinerja model dalam menghadapi data yang tidak seimbang, di mana pelanggan yang churn hanya merupakan minoritas dalam dataset.

Pendekatan pengembangan berbasis Agile yang diterapkan dalam penelitian ini juga terbukti efektif dalam menyempurnakan model secara iteratif, dengan memperhatikan umpan balik dari pengguna dan hasil evaluasi model. Hal ini memungkinkan tim pengembang untuk terus meningkatkan performa model secara berkelanjutan.

REKOMENDASI

1. Penerapan dalam Bisnis: Model prediksi churn dapat membantu TELCO Company mengidentifikasi pelanggan berisiko churn, memungkinkan tim mengambil tindakan preventif seperti penawaran khusus atau program loyalitas.
2. Pemantauan dan Pembaruan: Model perlu diperbarui secara berkala dengan data terbaru untuk menjaga akurasi dan relevansi, mengingat perubahan perilaku pelanggan.
3. Pengembangan Fitur: Menambahkan variabel baru seperti interaksi digital, perilaku belanja, dan data kepuasan pelanggan dapat meningkatkan akurasi prediksi.
4. Peningkatan Machine Learning: Selain churn, TELCO bisa memperluas penggunaan machine learning untuk prediksi penjualan dan rekomendasi produk, meningkatkan efisiensi operasi dan pengalaman pelanggan.

Dengan langkah ini, TELCO Company dapat mengurangi churn, meningkatkan loyalitas, dan meraih keunggulan kompetitif di industri telekomunikasi.

DAFTAR REFERENSI

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6(1), 28. https://doi.org/10.1186/s40537-019-0191-6

Azmi, M. (2020). Data Quality in Machine Learning: A Practical Approach to Data Cleaning and Preprocessing. *Journal of Data Science*, 15(1), 85-101. https://doi.org/10.1016/j.jds.2020.05.012

Brown, A., & Davis, K. (2019). Optimizing Machine Learning Models: A Guide to Tuning Hyperparameters for Better Performance. *Machine Learning Journal*, 12(4), 122-145. https://doi.org/10.1016/j.ml.2019.07.015

Brown, A., & Larkin, T. (2020). Feature Engineering in Predictive Analytics: Best Practices and Applications. *International Journal of Data Science*, 8(3), 89-112. https://doi.org/10.1016/j.ijds.2020.03.019

Fader, P. S., & Hardie, B. G. S. (2019). *Customer-Base Analysis: Predicting Customer Behavior with Data Analytics*. Wiley. https://doi.org/10.1002/9781119610410

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.). O'Reilly Media. https://doi.org/10.1002/9781119610410

Gürsoy, U. Ş., et al. (2021). Customer churn prediction system: A machine learning approach. *Computing*, 104(2), 271-294. https://doi.org/10.1007/s00607-021-00908-y

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

Kazemi, M., & Hejazinia, R. (2022). Predicting customer churn using machine learning: A case study in the software industry. *Journal of Marketing Analytics*, 10(1), 45-56. https://doi.org/10.1057/s41270-022-00128-4

Kim, M., & Hwang, K. B. (2022). An empirical evaluation of sampling methods for the classification of imbalanced data. *PLoS ONE*, 17(7), e0271260. https://doi.org/10.1371/journal.pone.0271260

Kim, S., & Hwang, H. (2022). Using Feature Selection and Resampling Methods to Improve Churn Prediction in Imbalanced Datasets. *Journal of Data Science*, 20(2), 189-208. https://doi.org/10.1016/j.jds.2022.01.005

Lalwani, P., et al. (2022). Customer churn prediction system: A machine learning approach. *Computing*, 104(2), 271–294. https://doi.org/10.1007/s00607-021-00908-y

Musheer, R. A., Verma, C., & Srivastava, N. (2019). Novel machine learning approach for classification of high-dimensional microarray data. *Soft Computing*, 23(24), 13409-13421. https://doi.org/10.1007/s00500-019-04222-8

Pebrianti, D., Istinabiyah, D. D., Bayuaji, L., & Rusdah, L. (2022). Hybrid method for churn prediction model in the case of telecommunication companies. *9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 161–166. https://doi.org/10.23919/EECSI56542.2022.9946535

Salunkhe, U. R., & Mali, S. N. (2021). A hybrid approach for class imbalance problem in customer churn prediction: A novel extension to under-sampling. *International Journal of Intelligent Systems and Applications*, 10, 71–81. https://doi.org/10.5281/zenodo.5090570

Sharma, T., Gupta, P., Nigam, V., & Goel, M. (2020). Customer churn prediction in telecommunications using gradient boosted trees. In *International Conference on Innovative Computing and Communications*. Springer, Singapore. https://doi.org/10.1007/978-981-15-0324-5_20

Smith, J., & Johnson, P. (2020). Data Quality in Predictive Modeling: Techniques and Applications. *Journal of Data Analytics*, 10(3), 45-60. https://doi.org/10.1016/j.jda.2020.02.014

Tavassoli, S., & Koosha, H. (2022). Hybrid ensemble learning approaches to customer churn prediction. *Kybernetes*, 51(3), 1062-1088. https://doi.org/10.1108/K-04-2020-0214

Wang, T., Davis, K., & Brown, A. (2019). Outlier Detection Methods in Predictive Models: A Comparative Study. *International Journal of Data Science*, 6(2), 200-215. https://doi.org/10.1016/j.ijds.2019.04.017