JSI (Jurnal Sistem Informasi) Universitas Dirgantara Marsekal Suryadarma
p-ISSN: 2355-9675
e-ISSN: 2541-3228

Klasifikasi Churn Pelanggan Telekomunikasi Menggunakan Logistic Regression dan Support Vector Machine

Lita Dwi Aryani[1,*] Hafiyyan Putra Pratama[2]
[1,2]Program Studi S1 Sistem Telekomunikasi, Universitas Pendidikan Indonesia
lita1411@upi.edu, hafiyyan@upi.edu

Article Info

Article history:

Received May 31, 2026
Accepted June 11, 2026
Published July 1, 2026

Kata Kunci:

Churn pelanggan,
klasifikasi,
logistic regression,
SVM,
machine learning,
telekomunikasi

ABSTRACT

Customer churn is a critical issue in the telecommunication industry as it directly affects a company's revenue. This study aims to develop and compare Logistic Regression and Support Vector Machine (SVM) models for predicting customer churn using the Telco Customer Churn dataset from IBM Watson Analytics, which consists of 7,043 customer records. The research process includes data exploration, data preprocessing, model training, and evaluation using Stratified K-Fold Cross-Validation (k = 5). The experimental results show that Logistic Regression achieved an accuracy of 80.70% with an average cross-validation score of 0.8043, while SVM achieved an accuracy of 79.28% with an average cross-validation score of 0.7954. Feature analysis indicates that tenure, MonthlyCharges, contract type, and internet service type are the most influential factors affecting customer churn. Based on these results, Logistic Regression demonstrates superior and more stable performance in predicting telecommunication customer churn.

Corresponding Author:

Lita Dwi Aryani,
Program Studi S1 Sistem Telekomunikasi,
Universitas Pendidikan Indonesia,
Email: *lita1411@upi.edu

1. PENDAHULUAN

Industri telekomunikasi merupakan salah satu sektor dengan tingkat persaingan yang sangat tinggi di era transformasi digital. Perkembangan teknologi informasi dan komunikasi yang pesat mendorong perusahaan telekomunikasi untuk terus berinovasi dalam menyediakan layanan yang berkualitas dan sesuai dengan kebutuhan pelanggan. Di sisi lain, kemudahan pelanggan dalam berpindah operator, rendahnya biaya perpindahan layanan, serta banyaknya pilihan produk yang tersedia menyebabkan perusahaan menghadapi tantangan besar dalam mempertahankan pelanggan yang sudah dimiliki (Desena Damanik & Ihsan Jambak, 2023). Salah satu permasalahan yang sering dihadapi perusahaan telekomunikasi adalah customer churn, yaitu kondisi ketika pelanggan memutuskan untuk menghentikan penggunaan layanan dan beralih ke penyedia layanan lain (Awaludin & Rehatalanit, 2026). Tingginya tingkat churn dapat memberikan dampak negatif terhadap perusahaan, seperti penurunan pendapatan, berkurangnya pangsa pasar, dan meningkatnya biaya pemasaran untuk memperoleh pelanggan baru (Rizki Kurniawan et al., 2023). Oleh karena itu, kemampuan perusahaan dalam mengidentifikasi pelanggan yang berpotensi melakukan churn menjadi faktor penting dalam mendukung keberlangsungan bisnis dan meningkatkan daya saing perusahaan (DAIPAH et al., 2025). Upaya mempertahankan pelanggan menjadi lebih penting karena biaya untuk memperoleh pelanggan baru umumnya lebih besar dibandingkan biaya mempertahankan pelanggan yang sudah ada (Rahayu, 2023). Selain itu, pelanggan yang loyal cenderung memberikan kontribusi yang lebih besar terhadap pendapatan perusahaan dalam jangka panjang. Dengan demikian, perusahaan memerlukan strategi yang efektif untuk mengurangi tingkat churn dan meningkatkan loyalitas pelanggan (Yeni et al., 2025). Salah

satu pendekatan yang dapat dilakukan adalah dengan memanfaatkan data pelanggan untuk memprediksi kemungkinan terjadinya churn sebelum pelanggan benar-benar meninggalkan layanan (Komang Dika Setiawan & Wayan Sudiarsa, 2026). Perkembangan teknologi machine learning memberikan peluang yang besar dalam menyelesaikan permasalahan prediksi churn. Machine learning memungkinkan sistem untuk mempelajari pola dari data historis dan menghasilkan model yang mampu memprediksi perilaku pelanggan di masa mendatang (Arya Renaldi et al., 2025). Data yang digunakan dapat mencakup informasi demografis pelanggan, lama berlangganan, jenis layanan yang digunakan, pola penggunaan layanan, metode pembayaran, serta riwayat transaksi pelanggan. Dengan memanfaatkan data tersebut, perusahaan dapat mengidentifikasi pelanggan yang memiliki risiko tinggi untuk melakukan churn dan mengambil tindakan preventif yang sesuai (Illah et al., 2024). Berbagai metode machine learning telah diterapkan untuk memprediksi churn pelanggan, mulai dari algoritma berbasis pohon keputusan, ensemble learning, artificial neural network, hingga algoritma klasifikasi lainnya (Awaludin, Yasin, et al., 2024). Setiap algoritma memiliki karakteristik, kelebihan, dan keterbatasan yang berbeda dalam mengolah data dan menghasilkan prediksi (Hulaifah Al Abrori & Subhiyakto, 2025). Oleh karena itu, pemilihan algoritma yang tepat menjadi salah satu faktor penting yang memengaruhi keberhasilan model prediksi churn.

Di antara berbagai algoritma klasifikasi yang tersedia, Logistic Regression dan Support Vector Machine (SVM) merupakan dua metode yang banyak digunakan dalam berbagai kasus prediksi (Dedy, 2024). Logistic Regression dikenal sebagai algoritma klasifikasi yang sederhana, mudah diinterpretasikan, serta mampu menunjukkan hubungan antara variabel independen dan variabel target (Yunisia Rosari Bere & Fadhli Almu'iini Ahda, 2026) (Awaludin, Nuryadi, et al., 2024) Sementara itu, Support Vector Machine memiliki kemampuan yang baik dalam menangani data dengan pola yang kompleks melalui penggunaan fungsi kernel, sehingga berpotensi menghasilkan performa klasifikasi yang lebih baik pada kondisi tertentu (Sapaatullah & Darip, 2026). Perbedaan karakteristik kedua algoritma tersebut menjadikan keduanya menarik untuk dibandingkan dalam kasus prediksi churn pelanggan.

Penelitian ini menggunakan dataset Telco Customer Churn dari IBM Watson Analytics yang banyak digunakan sebagai benchmark dalam penelitian prediksi churn pelanggan. Dataset tersebut memuat berbagai informasi pelanggan yang dapat dimanfaatkan untuk membangun model klasifikasi dan menganalisis faktor-faktor yang memengaruhi keputusan churn. Penggunaan dataset benchmark juga memungkinkan hasil penelitian untuk dibandingkan dengan penelitian lain yang menggunakan data serupa (Yoga Pudya Ardhana et al., 2025). Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk membangun dan membandingkan model prediksi churn pelanggan menggunakan algoritma Logistic Regression dan Support Vector Machine (SVM). Evaluasi dilakukan menggunakan berbagai metrik performa klasifikasi untuk mengetahui model yang memiliki kinerja terbaik (Ridwan et al., 2025). Selain itu, penelitian ini juga bertujuan mengidentifikasi faktor-faktor yang berpengaruh terhadap churn pelanggan sehingga hasil yang diperoleh dapat memberikan informasi yang bermanfaat bagi perusahaan telekomunikasi dalam menyusun strategi retensi pelanggan yang lebih efektif (Arina & Ulfah, 2022).

2. METODE

A. State of the art

Terdapat research gap yang menjadi dasar penelitian ini, yang pertama penelitian-penelitian sebelumnya dominan menggunakan metode berbasis pohon (C4.5, LightGBM) sehingga perbandingan langsung antara Logistic Regression dan SVM pada dataset Telco IBM Watson masih terbatas. Kedua, penggunaan Stratified K-Fold Cross Validation sebagai mekanisme evaluasi yang robust belum banyak diterapkan secara eksplisit pada konteks prediksi churn telekomunikasi. Ketiga, analisis interpretabilitas fitur secara komparatif antara model linier (Logistic Regression) dan model berbasis kernel (SVM) belum banyak dieksplorasi. Tabel 1 berikut merangkum penelitian-penelitian terdahulu tersebut.

Tabel 1 Ringkasan State of the Art Penelitian Prediksi Churn Pelanggan

No	Nama Peneliti	Pembahasan (Masalah dan Solusi)	Hasil Penelitian
1	Desena Damanik & Ihsan Jambak (2023)	Masalah: Tingginya tingkat churn pelanggan telekomunikasi berdampak pada penurunan pendapatan perusahaan. Solusi: Klasifikasi churn menggunakan algoritma C4.5 untuk mendukung strategi retensi pelanggan.	Akurasi 84,3%. Algoritma C4.5 efektif dalam identifikasi fitur penting, namun rentan overfitting pada data tidak seimbang.
2	Rizki Kurniawan et al. (2023)	Masalah: Prediksi churn dengan algoritma tunggal C4.5 menghasilkan akurasi yang belum optimal. Solusi: Optimasi C4.5 berbasis Particle Swarm Optimization (PSO) untuk meningkatkan seleksi fitur dan akurasi klasifikasi.	Akurasi meningkat hingga 87,2%. Pendekatan hybrid efektif namun memiliki kompleksitas komputasi yang lebih tinggi.
3	Illah et al. (2024)	Masalah: Ketidakseimbangan kelas dan kebutuhan analisis waktu bertahan pelanggan (survival time) dalam prediksi churn. Solusi: Kombinasi LightGBM dengan analisis survival untuk prediksi churn yang lebih komprehensif.	AUC sebesar 0,89. LightGBM unggul menangani data tidak seimbang, namun proses tuning lebih kompleks.
4	Penelitian ini (2026)	Masalah: Perbandingan langsung Logistic Regression vs SVM pada dataset Telco IBM Watson dengan evaluasi cross-validation yang robust masih terbatas. Solusi: Komparasi Logistic Regression dan SVM menggunakan Stratified K-Fold CV (k=5) disertai analisis interpretabilitas fitur.	LR: akurasi 80,70%, CV Mean 0,8043. SVM: akurasi 79,28%, CV Mean 0,7954. LR lebih unggul dan stabil.

B. Dataset

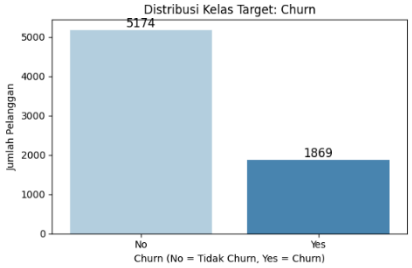
Dataset yang digunakan adalah Telco Customer Churn dari IBM Watson Analytics, tersedia secara publik di Kaggle, yang terdiri dari 7.043 entri pelanggan dengan 21 kolom (20 fitur + 1 variabel target). Rincian fitur mencakup informasi demografis (gender, SeniorCitizen, Partner, Dependents), detail akun (tenure, Contract, PaperlessBilling, PaymentMethod), tagihan (MonthlyCharges, TotalCharges), serta berbagai layanan yang digunakan pelanggan. Variabel target adalah Churn (Yes/No).

C. Kerangka Metodologi

Penelitian ini mengikuti pipeline data science standar yang terdiri dari lima tahapan terstruktur: (1) pengumpulan dan pemahaman data; (2) eksplorasi data awal (EDA); (3) praproses data; (4) pelatihan dan evaluasi model; serta (5) analisis hasil dan penarikan kesimpulan. Setelah itu, tahap EDA dilakukan dengan tujuan untuk memahami karakteristik dataset sebelum pemodelan menggunakan library pandas, numpy, matplotlib, dan seaborn. Pada pemeriksaan awal, dataset memiliki 7.043 baris dan 21 kolom dan ditemukan 11 nilai kosong tersembunyi pada kolom TotalCharges setelah konversi tipe data.

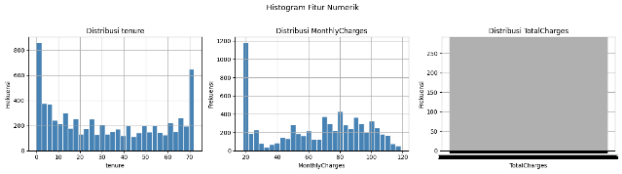
1) Analisis Distribusi Target

Analisis distribusi variabel target mengungkapkan ketidakseimbangan kelas yang signifikan: 5.174 pelanggan tidak churn (73,5%) dan 1.869 pelanggan churn (26,5%), dengan rasio ketidakseimbangan sekitar 2,77:1 yang dapat dilihat pada Gambar 1.



Gambar 1 Analisis Distribusi Target

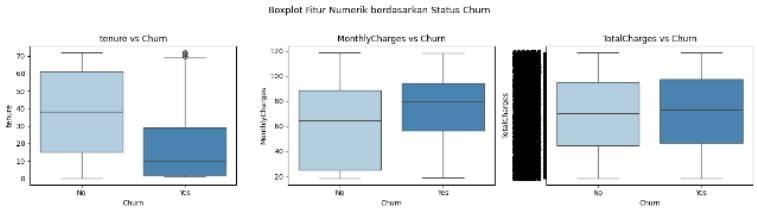
2) Analisis Univariat Fitur Numerik



Gambar 2 Analisis Univariat Fitur Numerik

Berdasarkan Gambar 2**,** variabel *tenure* menunjukkan pola distribusi bimodal. Sementara itu, *MonthlyCharges* terdistribusi relatif merata dan *TotalCharges* memiliki distribusi *right-skewed* yang mencerminkan dominasi pelanggan baru.

3) Analisis Bivariat



Gambar 3 Boxplot Fitur Numerik Berdasarkan Status Churn

Berdasarkan Gambar 3, pelanggan churn memiliki median *tenure* sekitar 10 bulan, lebih rendah dibandingkan pelanggan non-churn yang memiliki median sekitar 38 bulan. Selain itu, pelanggan churn cenderung memiliki nilai *MonthlyCharges* yang lebih tinggi. Hasil analisis juga menunjukkan bahwa tingkat churn tertinggi ditemukan pada pelanggan dengan kontrak *Month-to-month*, pengguna layanan *Fiber optic*, dan pengguna metode pembayaran *Electronic check*.

D. Praproses Data

1) Pembersihan Data

Berdasarkan Gambar 4, kolom *customerID* dihapus karena tidak memiliki nilai prediktif dalam proses klasifikasi. Selain itu, kolom *TotalCharges* dikonversi ke tipe numerik menggunakan fungsi pd.to_numeric() dengan parameter errors='coerce', sehingga berhasil mengidentifikasi 11 nilai yang hilang (*missing values*).

```
# Salin dataset agar data asli tidak berubah
df_prep = df.copy()

# 1. Hapus kolom customerID karena tidak relevan untuk prediksi
df_prep.drop(columns=['customerID'], inplace=True)

# 2. Konversi TotalCharges ke numerik (ada spasi/nilai kosong yang masuk sebagai string)
df_prep['TotalCharges'] = pd.to_numeric(df_prep['TotalCharges'], errors='coerce')

print("Nilai kosong setelah konversi TotalCharges:")
print(df_prep.isnull().sum()[df_prep.isnull().sum() > 0])
```

```
Nilai kosong setelah konversi TotalCharges:
TotalCharges    11
dtype: int64
```

Gambar 4 Penghapusan Kolom & Konversi

2) Imputasi

Berdasarkan Gambar 5, sebanyak 11 *missing values* pada variabel *TotalCharges* diimputasi menggunakan nilai median sebesar 1.397,47. Metode ini dipilih karena lebih robust terhadap *outlier* dibandingkan *mean*, sehingga lebih sesuai untuk distribusi data yang cenderung *right-skewed*.

```
# 3. Imputasi nilai kosong pada TotalCharges dengan nilai median
median_tc = df_prep['TotalCharges'].median()
df_prep['TotalCharges'].fillna(median_tc, inplace=True)

print(f"Imputasi TotalCharges dengan median: {median_tc:.2f}")
print("Nilai kosong setelah imputasi:", df_prep.isnull().sum().sum())
```

```
Imputasi TotalCharges dengan median: 1397.47
Nilai kosong setelah imputasi: 0
```

Gambar 5 Imputasi Median

3) Encoding Variabel Kategorikal

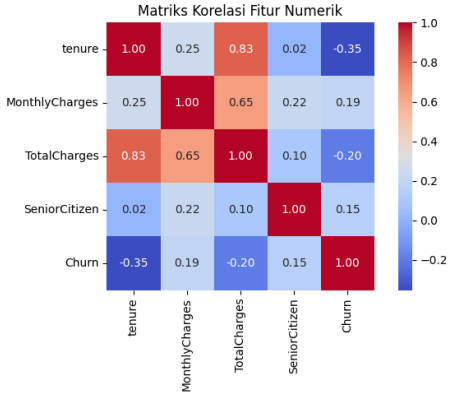
Sebelum tahap pemodelan, transformasi variabel kategorikal dilakukan agar mengubah menjadi bentuk numerik. *Label Encoding* diterapkan pada fitur biner, sedangkan *One-Hot Encoding* dengan parameter drop_first=True digunakan pada fitur multi-kelas. Variabel target diencode menjadi No = 0 dan Yes = 1. Setelah proses encoding, dimensi dataset berubah dari 7.043 × 21 menjadi 7.043 × 31.

4) Normalisasi

Fitur numerik dinormalisasi menggunakan *StandardScaler* sehingga memiliki *mean* 0 dan standar deviasi 1. Tahapan ini penting untuk memastikan seluruh fitur berada pada skala yang sebanding, mengingat algoritma Logistic Regression dan SVM sensitif terhadap perbedaan rentang nilai antarfitur.

5) Analisis Korelasi

Berdasarkan Gambar 6, matriks korelasi Pearson menunjukkan bahwa variabel *tenure* memiliki korelasi negatif yang cukup kuat terhadap *Churn* (r = −0,35), yang mengindikasikan bahwa pelanggan dengan masa berlangganan lebih lama cenderung memiliki risiko churn yang lebih rendah. Selain itu, *tenure* memiliki korelasi positif yang sangat kuat dengan *TotalCharges* (r = 0,83), sedangkan *MonthlyCharges* menunjukkan korelasi positif moderat terhadap *Churn* (r = 0,19).



Gambar 6 Matriks korelasi fitur

E. Pembagian Data dan Pemodelan

Setelahnya, data dibagi menggunakan train_test_split dengan rasio 80:20 sehingga diperoleh 5.634 data pelatihan dan 1.409 data pengujian. Parameter stratify=y digunakan untuk mempertahankan proporsi kelas pada kedua subset data, sedangkan random_state=42 diterapkan untuk menjamin reproduktibilitas hasil. Tahapan pembagian data dan proses pelatihan model menggunakan dua algoritma machine learning ditunjukkan pada Gambar 7.

```
# Split data: 80% training, 20% testing
X_train, X_test, y_train, y_test = train_test_split(
    X_scaled, y, test_size=0.2, random_state=42, stratify=y
)

print(f"Data Training : {X_train.shape[0]} sampel")
print(f"Data Testing  : {X_test.shape[0]} sampel")
print(f"Distribusi Churn (train): {pd.Series(y_train).value_counts().to_dict()}")
print(f"Distribusi Churn (test) : {pd.Series(y_test).value_counts().to_dict()}")
```

```
Data Training : 5634 sampel
Data Testing  : 1409 sampel
Distribusi Churn (train): {0: 4139, 1: 1495}
Distribusi Churn (test) : {0: 1035, 1: 374}
```

Gambar 7 Pelatihan menggunakan dua algoritma machine learning

1. Logistic Regression

Berdasarkan Gambar 8, model prediksi churn dibangun menggunakan algoritma *Logistic Regression* dari library *scikit-learn*. Parameter max_iter=1000 digunakan untuk memastikan proses optimasi mencapai konvergensi, random_state=42 diterapkan untuk menjaga reproduktibilitas hasil, dan *solver* lbfgs dipilih sebagai metode optimasi dalam proses pelatihan model.

```
=== Logistic Regression ===
Akurasi: 0.8070

              precision    recall  f1-score   support

    No Churn       0.85      0.89      0.87      1035
       Churn       0.66      0.57      0.61       374

    accuracy                           0.81      1409
   macro avg       0.75      0.73      0.74      1409
weighted avg       0.80      0.81      0.80      1409
```

Gambar 8 Model Logistic Regression

2. Support Vector Machine

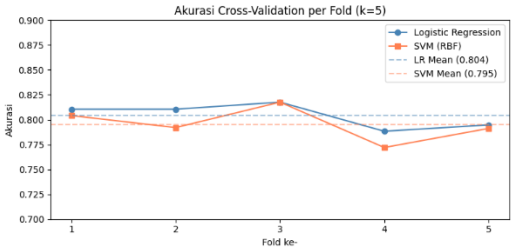
Berdasarkan Gambar 9, model prediksi churn dikembangkan menggunakan algoritma *Support Vector Machine (SVM)* melalui kelas SVC pada library *scikit-learn*. Parameter kernel='rbf' digunakan untuk memodelkan hubungan nonlinier antarfitur, C=1.0 berfungsi sebagai parameter regularisasi, sedangkan gamma='scale' digunakan untuk menyesuaikan pengaruh masing-masing data pelatihan terhadap pembentukan batas keputusan. Selain itu, random_state=42 diterapkan untuk menjaga konsistensi hasil selama proses pengujian.

```
=== Support Vector Machine (RBF) ===
Akurasi: 0.7928

              precision    recall  f1-score   support

    No Churn       0.83      0.90      0.86      1035
       Churn       0.64      0.49      0.56       374

    accuracy                           0.79      1409
   macro avg       0.74      0.70      0.71      1409
weighted avg       0.78      0.79      0.78      1409
```

Gambar 9 Model Support Vector Machine (SVM)

F. Evaluasi

Evaluasi model dilakukan melalui beberapa metrik pengukuran. Pertama, accuracy_score digunakan untuk mengukur tingkat ketepatan prediksi model secara keseluruhan. Kedua, classification_report digunakan untuk memperoleh nilai *precision*, *recall*, dan *F1-score* pada setiap kelas sehingga performa model dapat dianalisis secara lebih rinci. Selanjutnya, confusion_matrix digunakan untuk melihat distribusi hasil prediksi dan mengidentifikasi kesalahan klasifikasi yang terjadi. Selain itu, berdasarkan Gambar 10, metode Stratified K-Fold Cross Validation dengan nilai k = 5 diterapkan untuk memperoleh estimasi performa model yang lebih stabil dan representatif dengan tetap mempertahankan proporsi kelas pada setiap pembagian data.



Gambar 10 Visualisasi Cross Validation

3. HASIL DAN PEMBAHASAN

A. Hasil Logistic Regression

Model Logistic Regression menghasilkan akurasi sebesar 80,70% pada data testing. Berdasarkan classification report, model menunjukkan performa yang cukup baik dalam memprediksi kelas Non-Churn (precision=0,85, recall=0,89, F1=0,87), namun masih terbatas pada kelas Churn (precision=0,66, recall=0,57, F1=0,61). Nilai recall Churn sebesar 0,57 mengindikasikan bahwa sekitar 43% pelanggan yang benar-benar akan churn tidak berhasil diidentifikasi oleh model.

B. Hasil Support Vector Machine

Model SVM dengan kernel RBF menghasilkan akurasi 79,28%, lebih rendah dari Logistic Regression. Meskipun recall untuk kelas Non-Churn sangat tinggi (0,90), kemampuan mendeteksi kelas Churn lebih rendah dengan recall hanya 0,49. Hal ini menunjukkan bahwa SVM dengan parameter default cenderung bias terhadap kelas mayoritas.

C. Analisis Confusion Matrix

Tabel 2 Hasil Confusion Matrix pada Data Testing

Model	TP	TN	FP	FN
Logistic Regression	212	925	110	162
SVM(RBF)	183	934	101	191

Untuk memperoleh gambaran yang lebih rinci mengenai kinerja model, dilakukan analisis menggunakan confusion matrix pada data testing. Hasil confusion matrix dari kedua model ditunjukkan pada Tabel 2. Model Logistic Regression menghasilkan nilai True Positive (TP) yang lebih tinggi dibandingkan SVM, yaitu 212 berbanding 183. Selain itu, Logistic Regression memiliki jumlah False Negative (FN) yang lebih rendah, yaitu 162 dibandingkan 191 pada model SVM. Hasil ini menunjukkan bahwa Logistic Regression lebih baik dalam mengidentifikasi pelanggan yang berpotensi melakukan churn. Logistic Regression memperoleh akurasi sebesar 80,70%, sedangkan SVM memperoleh akurasi sebesar 79,28%. Nilai akurasi tersebut sesuai dengan hasil evaluasi pada data testing yang akan disajikan pada Tabel 3. Dengan demikian, hasil confusion matrix dan metrik evaluasi yang diperoleh menunjukkan konsistensi dalam pengukuran performa kedua model.

D. Hasil Validasi Silang

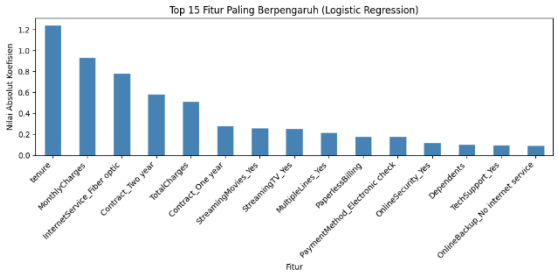
Selain menggunakan data testing, evaluasi model juga dilakukan menggunakan metode Stratified 5-Fold Cross Validation untuk mengukur kestabilan dan kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Hasil evaluasi ditunjukkan pada Tabel 3.

Tabel 3 Perbandingan Performa Model pada Data Testing dan Cross Validation

Model	Akurasi	Prec.	Recall	F1-Score	Mean CV Accuracy	CV Std
Logistic Regression	80,70%	0,66	0,57	0,61	0,8043	0,0110
SVM (RBF)	79,28%	0,64	0,49	0,56	0,7954	0,0151

Berdasarkan Tabel 3, metrik akurasi, precision, recall, dan F1-score diperoleh dari data testing, sedangkan Mean CV Accuracy dan CV Std berasal dari Stratified 5-Fold Cross Validation. Logistic Regression menghasilkan Mean CV Accuracy sebesar 0,8043 dengan standar deviasi 0,0110, lebih baik dibandingkan SVM yang memperoleh 0,7954 dengan standar deviasi 0,0151. Kedekatan antara akurasi data testing dan Mean CV Accuracy pada kedua model menunjukkan performa yang stabil serta kemampuan generalisasi yang baik tanpa indikasi overfitting. Meskipun demikian, Logistic Regression tetap unggul dibandingkan SVM berdasarkan seluruh metrik evaluasi yang digunakan.

E. Hasil Analisis Fitur yang Berpengaruh



Gambar 11 Fitur yang mempengaruhi hasil

Berdasarkan Gambar 11, terdapat 15 fitur yang paling berpengaruh terhadap prediksi churn pelanggan. Fitur tenure memiliki pengaruh terbesar dengan koefisien negatif, yang menunjukkan bahwa semakin lama pelanggan berlangganan, semakin kecil kemungkinan mereka melakukan churn. Sebaliknya,

MonthlyCharges dan penggunaan layanan Fiber Optic memiliki koefisien positif, sehingga pelanggan dengan biaya bulanan tinggi dan pengguna Fiber Optic cenderung memiliki risiko churn yang lebih besar. Di sisi lain, fitur Contract_Two year, TechSupport, dan OnlineSecurity memiliki koefisien negatif, yang mengindikasikan bahwa kontrak jangka panjang serta layanan dukungan dan keamanan tambahan dapat meningkatkan loyalitas pelanggan dan menurunkan kemungkinan churn.

F. Pembahasan

Berdasarkan hasil evaluasi, Logistic Regression menunjukkan performa terbaik dibandingkan Support Vector Machine (SVM), baik dari sisi akurasi, recall, maupun stabilitas cross-validation. Temuan ini mengindikasikan bahwa pola churn pada data telekomunikasi cenderung dapat dimodelkan secara efektif menggunakan pendekatan linier. Sementara itu, performa SVM yang lebih rendah kemungkinan dipengaruhi oleh belum optimalnya pengaturan parameter, ketidakseimbangan kelas, serta karakteristik data yang relatif linier. Tantangan utama pada kedua model adalah distribusi kelas yang tidak seimbang, yang menyebabkan nilai recall kelas churn masih terbatas sehingga sebagian pelanggan churn terklasifikasi sebagai non-churn. Mengingat kesalahan tersebut berpotensi menimbulkan kerugian bisnis, peningkatan recall perlu menjadi fokus pengembangan model selanjutnya. Analisis fitur menunjukkan bahwa tenure dan MonthlyCharges merupakan faktor paling berpengaruh terhadap churn, di mana pelanggan dengan masa berlangganan singkat, biaya layanan tinggi, kontrak bulanan, dan pengguna Fiber Optic memiliki risiko churn lebih besar. Temuan ini dapat menjadi dasar bagi perusahaan untuk menerapkan strategi retensi pelanggan yang lebih tepat sasaran dan berbasis data.

4. KESIMPULAN

Penelitian ini berhasil membangun dan membandingkan model Logistic Regression dan Support Vector Machine (SVM) untuk prediksi customer churn menggunakan dataset IBM Watson Telco Customer Churn. Hasil evaluasi menunjukkan bahwa Logistic Regression memberikan performa terbaik dengan akurasi 80,70% dan rata-rata cross-validation 0,8043 ± 0,0110, lebih unggul dibandingkan SVM dengan akurasi 79,28% dan rata-rata cross-validation 0,7954 ± 0,0151. Kedua model menunjukkan kemampuan generalisasi yang baik tanpa indikasi overfitting. Analisis fitur mengidentifikasi tenure, MonthlyCharges, jenis kontrak, dan jenis layanan internet sebagai faktor utama yang memengaruhi churn. Temuan ini dapat dimanfaatkan perusahaan untuk menyusun strategi retensi pelanggan yang lebih tepat sasaran. Namun, penelitian ini masih menghadapi keterbatasan berupa ketidakseimbangan kelas pada dataset. Oleh karena itu, penelitian selanjutnya disarankan menerapkan teknik penyeimbangan data, melakukan hyperparameter tuning, serta mengeksplorasi algoritma lain seperti Random Forest, XGBoost, dan LightGBM guna meningkatkan performa prediksi churn.

DAFTAR PUSTAKA

Arina, F., & Ulfah, M. (2022). Analisa survival untuk mengurangi customer churn pada perusahaan telekomunikasi. *Journal Industrial Servicess*, *8*(1), 59. https://doi.org/10.36055/jiss.v8i1.14313

Arya Renaldi, R., Rais Rahmat Razak, M., Yakub, R., Ekonomi dan Bisnis, F., & Muhammadiyah Sidenreng Rappang, U. (2025). *Jurnal Maneksi (Management Ekonomi Dan Akuntansi) ANALISIS PENGGUNAAN MACHINE LEARNING TERHADAP PREDIKSI JUMLAH NASABAH PADA PRODUK AMANAH DI PT. PEGADAIAN CPS PANGKAJENE*. https://doi.org/10.31959/jm.v15i1

Awaludin, M., Nuryadi, H., & Pribadi, G. N. (2024). *Sistem Otomatisasi Laporan untuk Optimalisasi Pelaporan Data Penelitian dan Pengabdian kepada Masyarakat di Universitas Dirgantara Marsekal Suryadarma*. *9675*, 1–7.

Awaludin, M., & Rehatalanit, Y. L. R. (2026). Optimizing YOLOv8 Architecture and Augmentation for Efficient License Plate Detection. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *15*(2), 99–105. https://doi.org/10.22146/jnteti.v1512.24886

Awaludin, M., Yasin, V., & Risyda, F. (2024). The Influence of Artificial Intelligence Technology, Infrastructure and Human Resource Competence on Internet Access Networks. *Inform : Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *9*(2), 111–120. https://doi.org/10.25139/inform.v9i2.8109

DAIPAH, I. I., ASTUTI, R., & PRIHARTONO, W. (2025). PREDIKSI CHURN PELANGGAN PADA LAYANAN DESAIN GRAFIS HOME DESAIN MENGGUNAKAN ALGORITMA NAÏVE BAYES. *Jurnal Informatika Dan Teknik Elektro Terapan*, *13*(1). https://doi.org/10.23960/jitet.v13i1.5811

Dedy, D. (2024). Analisis Algoritma Logistic Regression dan Support Vector Machine pada Kasus Klasifikasi Citra Hewan Rawa dengan Dataset yang tidak Seimbang. *Data Sciences Indonesia (DSI)*, *4*(1), 69–77. https://doi.org/10.47709/dsi.v4i1.4433

Desena Damanik, S., & Ihsan Jambak, M. (2023). KLIK: Kajian Ilmiah Informatika dan Komputer Klasifikasi Customer Churn pada Telekomunikasi Industri Untuk Retensi Pelanggan Menggunakan Algoritma C4.5. *Media Online)*, *3*(6), 1303–1309. https://doi.org/10.30865/klik.v3i6.829

Hulaifah Al Abrori, Z. Z., & Subhiyakto, E. R. (2025). Analisis Komparatif Akurasi Prediksi Kanker Payudara Menggunakan Algoritma Random Forest dan Logistic Regression. *Jurnal Algoritma*, *22*(1), 300–311. https://doi.org/10.33364/algoritma/v.22-1.2164

Illah, I. Z. A., Jauharis Sapu, W. S., & Damaliana, A. T. (2024). Implementasi Metode Klasifikasi LightGBM dan Analisis Survival dalam Memprediksi Pelanggan Churn. *Jurnal Komtika (Komputasi Dan Informatika)*, *8*(1), 43–53. https://doi.org/10.31603/komtika.v8i1.11194

Komang Dika Setiawan, I., & Wayan Sudiarsa, I. (2026). PT. Media Akademik Publisher ANALISIS KLASIFIKASI PERILAKU PENGGUNA TERHADAP CUSTOMER CHURN PADA LAYANAN MUSIK SPOTIFY MENGGUNAKAN METODE RANDOM FOREST. *JMA)*, *4*(1), 3031–5220. https://doi.org/10.62281

Rahayu, S. (2023). Strategi Pemasaran Produk Dalam Meningkatkan Kepuasan Pelanggan. *Jurnal Penelitian Dan Pengkajian Ilmiah Sosial Budaya*, *2*(1), 109–113. https://doi.org/10.47233/jppisb.v2i1.705

Ridwan, R., Handayani, H. H., Lestari, S. A. P., & Cahyana, Y. (2025). Evaluasi Kinerja Algoritma Random Forest Dan Gradient Boosting Untuk Klasifikasi Penyakit Jantung. *Jurnal Komtika (Komputasi Dan Informatika)*, *9*(1), 112–124. https://doi.org/10.31603/komtika.v9i1.13450

Rizki Kurniawan, M., Nurul Sabrina, P., Ilyas Teknik Informatika, R., Jenderal Achmad Yani Jl Terusan Jend Sudirman, U., Cimahi Sel, K., Cimahi, K., & Barat, J. (2023). PREDIKSI CUSTOMER CHURN PADA PERUSAHAAN TELEKOMUNIKASI MENGGUNAKAN ALGORITMA C4.5 BERBASIS PARTICLE SWARM OPTIMIZATION. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 7, Number 5). https://doi.org/https://doi.org/10.36040/jati.v7i5.7476

Sapaatullah, A., & Darip, M. (2026). Analisis Performa Support Vector Machine untuk Klasifikasi Risiko Kredit Nasabah pada Perbankan Daerah. *Bulletin of Information Technology (BIT)*, *7*(1), 84–91. https://doi.org/10.47065/bit.v5i2.2603

Yeni, Y., Koli, D. Y., Irwansyah, R., & Sjioen, A. E. (2025). STRATEGI PEMASARAN PERSONALIZED RECOMMENDATION DALAM MENINGKATKAN RETENSI PELANGGAN. *JURNAL LENTERA BISNIS*, *14*(2), 1722–1735. https://doi.org/10.34127/jrlab.v14i2.1527

Yoga Pudya Ardhana, V., Lonang, S., Tejo Kumoro, D., Dermawan Mulyodiputro, M., & Qamarul Huda Badaruddin, U. (2025). Benchmarking Model Machine Learning untuk Prediksi Data Berdasarkan Akurasi dan Error Benchmarking Machine Learning Models for Data Prediction Based on Accuracy and Error. In *SIJ* (Vol. 2, Number 8). https://doi.org/https://doi.org/10.37824/sij.v8i2.2025.1141

Yunisia Rosari Bere, & Fadhli Almu'iini Ahda. (2026). *Perbandingan Metode Decision Tree dan Logistic Regression dalam Klasifikasi Tingkat Obesitas Berdasarkan Gaya Hidup*. https://doi.org/http://dx.doi.org/10.35889/jutisi.v15i2.3591