

Laporan Ujian Akhir Semester Analisis Big Data

Proyek Customer Churn Prediction



Oleh :

Vieco Syarifa Ambami	22031554008
Devina Sawitri	22031554036
Nur Halizah Amrita	22031554039
Siti Aida Hanun	22031554044

UNIVERSITAS NEGERI SURABAYA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

PROGRAM STUDI S1 SAINS DATA

2025

1. Ringkasan

Penelitian ini bertujuan untuk memprediksi *customer churn* pada layanan musik digital menggunakan machine learning. Data yang digunakan telah dilakukan EDA, tahap preprocessing dan feature engineering agar siap digunakan dalam pemodelan.

Dua model klasifikasi diterapkan, yaitu *Logistic Regression* sebagai model dasar dan *Random Forest* sebagai model lanjutan. Evaluasi kinerja model dilakukan menggunakan metrik klasifikasi untuk menilai kemampuan model dalam membedakan pelanggan *churn* dan tidak *churn*. Hasil pemodelan diharapkan dapat membantu perusahaan dalam memahami pola dan faktor yang memengaruhi *customer churn*, serta memberikan rekomendasi berbasis data guna menyusun strategi retensi pelanggan yang lebih efektif. Selain itu, penelitian ini juga memanfaatkan visualisasi data menggunakan *Elasticsearch* dan *Kibana* untuk menampilkan pola distribusi *churn* secara interaktif. Visualisasi ini diharapkan dapat memudahkan perusahaan dalam melakukan analisis eksploratif dan pengambilan keputusan secara lebih cepat dan informatif.

2. Analisis Data

Proyek ini menggunakan dataset KKBOX *Churn Prediction* www.kaggle.com/competitions/kkbox-churn-prediction-challenge/data yang bersumber dari Kaggle. Dataset ini merepresentasikan data perilaku pengguna layanan musik digital KKBOX, dengan tujuan memprediksi kemungkinan pelanggan melakukan *churn* (berhenti berlangganan). Dataset terdiri dari beberapa tabel utama yakni:

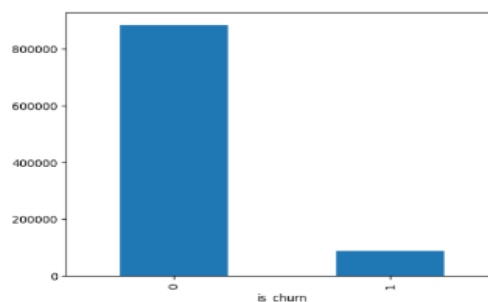
3. train.csv : label *churn* pengguna

- members.csv : informasi demografis pelanggan
- transactions.csv: riwayat transaksi dan status berlangganan
- user_logs.csv : aktivitas harian pengguna

Target variabel pada project ini adalah *is_churn*, yang menunjukkan apakah pelanggan berhenti berlangganan (1) atau tetap aktif (0). Dataset juga memiliki jutaan baris data dan ukuran file >300 MB, sehingga memenuhi kriteria dataset berskala menengah-besar.

2.1 Karakteristik Data

Dataset KKBox terdiri dari beberapa tabel berskala besar. Tabel target *churn* memiliki 970.960 baris dan 2 kolom tanpa missing value. Tabel transaksi berisi 1.431.009 baris dengan 9 kolom data pembayaran yang lengkap tanpa duplikasi. Tabel user logs merupakan tabel terbesar dengan 18.396.362 baris dan 9 kolom (± 1.2 GB) yang mencerminkan aktivitas pengguna secara detail dan tanpa missing value. Tabel member berjumlah 6.769.473 baris dengan 6 kolom, dengan missing value hanya pada kolom gender sebesar 65.43%, sementara kolom lainnya lengkap. Secara keseluruhan, dataset memiliki tipe data numerik, kategorikal, dan time-based dengan kualitas data yang baik dan tanpa data duplikat. Namun, distribusi target bersifat imbalanced pada Gambar 2.1.



Gambar 2.1 *Imbalance* pada Target

4. Metodologi

3.1 Data Preprocessing

Tahap data preprocessing dilakukan untuk memastikan kualitas dan konsistensi data sebelum analisis dan pemodelan.

- Pada tabel *members*, atribut *bd* merepresentasikan usia pengguna. Nilai usia yang tidak realistis (≤ 10 tahun dan > 80 tahun) diubah menjadi *missing value* karena berpotensi menimbulkan bias pada model.
- Pada tabel *transactions*, beberapa atribut utama dianalisis dan dibersihkan. Atribut *payment_plan_days* merepresentasikan jumlah hari paket berlangganan yang dibeli oleh pengguna. Nilai nol atau negatif dianggap tidak valid karena tidak merefleksikan durasi langganan yang sebenarnya, sehingga diubah menjadi *missing value*. Atribut *transaction_date* menunjukkan tanggal terjadinya transaksi pembayaran, sedangkan *membership_expire_date* menunjukkan tanggal berakhirnya masa berlangganan pengguna. Ditemukan beberapa transaksi di mana *transaction_date* melebihi *membership_expire_date*, yang secara logika tidak mungkin terjadi, sehingga record tersebut dihapus. Selanjutnya, seluruh kolom tanggal pada tabel *transactions* dikonversi ke format *datetime* dan difilter menggunakan *cut-off date* 1 April 2017 untuk mencegah *data leakage*. Hanya transaksi yang terjadi sebelum tanggal tersebut yang digunakan dalam analisis.
- Pada tabel *user_logs*, atribut *date* merepresentasikan tanggal aktivitas pengguna dalam menggunakan layanan KKBox. Kolom ini dikonversi ke format *datetime* dan difilter menggunakan *cut-off date* yang sama agar periode observasi konsisten dengan data transaksi.

3.2 Feature Engineering

Data transaksi diolah untuk menangkap pola pembayaran, durasi keanggotaan, serta kebiasaan pelanggan dalam melakukan pembaruan langganan. Data transaksi diagregasi berdasarkan *msno* (user ID) untuk menghasilkan beberapa fitur utama sebagai berikut:

- *total_transactions*, jumlah total transaksi yang dilakukan oleh pengguna. Fitur ini merepresentasikan frekuensi pembayaran pelanggan.
- *total_payment*, total nominal pembayaran yang telah dilakukan pengguna selama periode pengamatan.
- *avg_plan_days*, rata-rata durasi paket langganan (dalam hari) yang dipilih pengguna.
- *cancel_rate*, rata-rata pembatalan transaksi (*is_cancel*). Nilai mendekati 1 menunjukkan pengguna sering membatalkan langganan.
- *auto_renew_rate*, rata-rata penggunaan fitur auto-renew (*is_auto_renew*), yang mencerminkan tingkat loyalitas pengguna.
- *avg_amt_per_day*, rata-rata biaya yang dibayarkan per hari, dihitung dari perbandingan antara jumlah pembayaran dan durasi paket langganan.
- *autorenew_and_not_cancel*, variabel biner yang bernilai 1 jika pada transaksi terakhir pengguna mengaktifkan *auto-renew* dan tidak melakukan pembatalan. Fitur ini digunakan sebagai indikator kuat retensi pelanggan.
- *days_since_last_transaction*, jumlah hari sejak transaksi terakhir hingga tanggal cutoff. Semakin besar nilainya, semakin tinggi potensi *churn*.
- *membership_duration*, rata-rata selisih hari antara tanggal transaksi dan tanggal kedaluwarsa keanggotaan. Fitur ini merepresentasikan kecenderungan pengguna dalam memilih paket jangka pendek atau panjang.
- *Discount*, selisih antara harga paket (*plan_list_price*) dan pembayaran aktual (*actual_amount_paid*).
- *pct_discount_transactions*, persentase transaksi yang menggunakan diskon untuk setiap pengguna. Nilai yang tinggi mengindikasikan ketergantungan pelanggan terhadap promosi.

Dari tabel *members*, dilakukan pengolahan data pendaftaran:

- `long_time_user`, variabel biner yang menunjukkan apakah pengguna telah terdaftar lebih dari satu tahun sejak tanggal pendaftaran awal. Fitur ini digunakan untuk membedakan pengguna lama dan pengguna baru.

Dari tabel user logs, data aktivitas pengguna diolah untuk merepresentasikan tingkat keterlibatan (*engagement*) pelanggan terhadap layanan KKBox. Agregasi dilakukan berdasarkan `msno` untuk menghasilkan fitur berikut:

- `total_listening_secs`, total durasi mendengarkan musik pengguna.
- `avg_listening_secs`, rata-rata durasi mendengarkan per aktivitas.
- `active_days`, jumlah hari unik pengguna aktif menggunakan layanan.
- `last_user_log_date`, tanggal terakhir pengguna melakukan aktivitas.

Seluruh fitur hasil *feature engineering* kemudian digabungkan ke dalam satu dataset akhir dengan menggabungkan tabel train, members, transactions, dan user logs berdasarkan `msno`. Missing value yang muncul setelah proses merge ditangani dengan pengisian nilai nol (`fillna(0)`), sehingga dataset siap digunakan untuk tahap pemodelan. Dataset akhir disimpan dalam bentuk file `final_data.csv` dan digunakan sebagai input utama untuk proses training dan evaluasi model machine learning.

3.3 Model Machine Learning

3.3.1 Logistic Regression

Logistic Regression, merupakan metode klasifikasi biner untuk memprediksi apakah seorang pelanggan akan *churn* (1) atau tidak *churn* (0). Model ini dipilih karena sederhana, efisien, dan umum digunakan sebagai baseline model pada permasalahan klasifikasi.

Data yang telah diproses dibagi menjadi data latih (80%) dan data uji (20%) secara stratified agar distribusi kelas tetap seimbang. Sebelum *training*, dilakukan *preprocessing* berupa:

- Penghapusan kolom identitas (`msno`)
- Penanganan *missing value*
- Standardisasi fitur numerik
- *One-Hot Encoding* untuk fitur kategorikal

Evaluasi model dilakukan menggunakan data uji dengan beberapa metrik klasifikasi. Hasil evaluasi menunjukkan bahwa model mencapai *Accuracy* sebesar 0.935, *Precision* 0.595, *Recall* 0.885, *F1-Score* 0.711, serta *ROC-AUC* 0.964. Nilai *recall* yang tinggi menunjukkan bahwa sebagian besar pelanggan yang benar-benar *churn* berhasil terdeteksi oleh model, sedangkan nilai *ROC-AUC* yang mendekati 1 menandakan kemampuan model yang sangat baik dalam membedakan pelanggan *churn* dan tidak *churn*.

3.3.2 Random Forest

Random Forest merupakan metode *ensemble learning* berbasis pohon keputusan yang digunakan untuk memprediksi apakah seorang pelanggan akan *churn* (1) atau tidak *churn* (0). Model ini digunakan sebagai **model pembanding** karena mampu menangkap hubungan *non-linear* antar fitur dan cenderung memiliki performa prediksi yang lebih baik.

Data dibagi menjadi data latih (80%) dan data uji (20%) dengan menjaga proporsi kelas *churn* dan tidak *churn* tetap seimbang. *Preprocessing* meliputi penghapusan kolom identitas (`msno`), penanganan *missing value*, konversi fitur tanggal menjadi nilai numerik, serta *One-Hot Encoding* untuk fitur kategorikal. *Random Forest* tidak memerlukan standardisasi fitur numerik karena proses pemisahan data dilakukan berdasarkan nilai ambang pada fitur.

Hasil evaluasi menunjukkan bahwa *Random Forest* mencapai *Accuracy* 0.981, *Precision* 0.892, *Recall* 0.898, *F1-Score* 0.895, serta *ROC-AUC* 1.00, yang menunjukkan performa model sangat baik dan lebih unggul dibandingkan *Logistic Regression* dalam mendeteksi pelanggan *churn*.

5. Hasil dan Evaluasi

4.1 Hasil Final Data & Analisis Data Eksploratif

Dataset final diperoleh melalui proses penggabungan beberapa sumber data serta tahap *feature engineering*. Dataset ini terdiri dari 970.960 observasi dengan total 24 fitur, yang mencakup 19 fitur numerik serta 5 fitur kategorikal dan berbasis waktu. Variabel target yang digunakan dalam penelitian ini adalah *is_churn*, yang merepresentasikan status berhentinya pelanggan. Ukuran dataset mencapai sekitar 177,8 MB, sehingga memenuhi karakteristik data berskala besar. Selain itu, dataset final tidak mengandung *missing value*, sehingga siap digunakan untuk tahap analisis lanjutan dan pemodelan *machine learning*. Deskripsi setiap kolom terdapat pada tabel 4.1.

Tabel 4.1 Deskripsi Variabel Dataset

No	Nama Variabel	Tipe Data	Deskripsi Singkat
1	<i>msno</i>	Kategorikal	ID unik pengguna
2	<i>is_churn</i>	Biner	Status churn pelanggan (1 = churn, 0 = tidak)
3	<i>city</i>	Numerik	Kode kota pengguna
4	<i>bd</i>	Numerik	Usia pengguna
5	<i>gender</i>	Kategorikal	Jenis kelamin pengguna
6	<i>registered via</i>	Numerik	Kanal pendaftaran
7	<i>registration init time</i>	Waktu	Tanggal awal registrasi
8	<i>registration duration</i>	Numerik	Lama sejak registrasi (hari)
9	<i>long time user</i>	Biner	Indikator pengguna lama
10	<i>total transactions</i>	Numerik	Total transaksi
11	<i>total payment</i>	Numerik	Total pembayaran
12	<i>avg plan days</i>	Numerik	Rata-rata durasi paket
13	<i>cancel rate</i>	Numerik	Rasio pembatalan
14	<i>auto renew rate</i>	Numerik	Rasio auto-renew
15	<i>last membership expire</i>	Waktu	Tanggal akhir langganan
16	<i>avg amt per day</i>	Numerik	Rata-rata pembayaran per hari
17	<i>autorenew and not cancel</i>	Biner	Auto-renew tanpa pembatalan
18	<i>membership duration</i>	Numerik	Durasi keanggotaan
19	<i>days since last transaction</i>	Numerik	Hari sejak transaksi terakhir
20	<i>pct discount transactions</i>	Numerik	Persentase transaksi diskon
21	<i>total listening secs</i>	Numerik	Total waktu mendengarkan
22	<i>avg listening secs</i>	Numerik	Rata-rata waktu mendengarkan
23	<i>active days</i>	Numerik	Jumlah hari aktif
24	<i>last user log date</i>	Waktu	Aktivitas terakhir pengguna

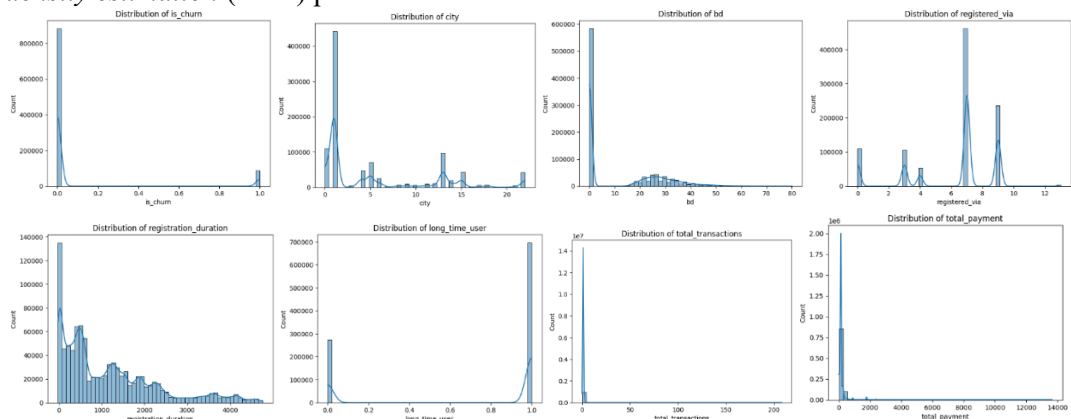
Selanjutnya dilakukan analisis terhadap persentase *imbalance* pada target Tabel 4.2.

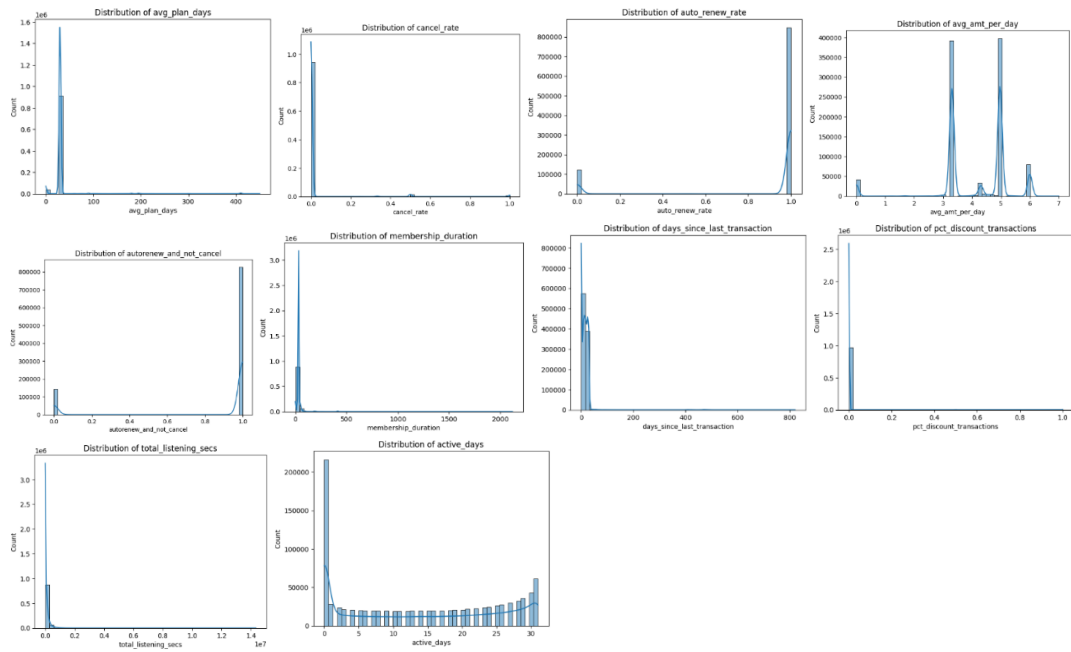
Tabel 4.2 Distribusi Target *Churn*

Status	Jumlah	Persentase
Tidak <i>Churn</i> (0)	883.719	91,01%
<i>Churn</i> (1)	87.241	8,99%

Dataset bersifat *imbalance*, sehingga evaluasi model tidak hanya bergantung pada akurasi.

Distribusi seluruh fitur numerik dianalisis menggunakan histogram dengan *kernel density estimation* (KDE) pada keseluruhan Gambar 4.1.

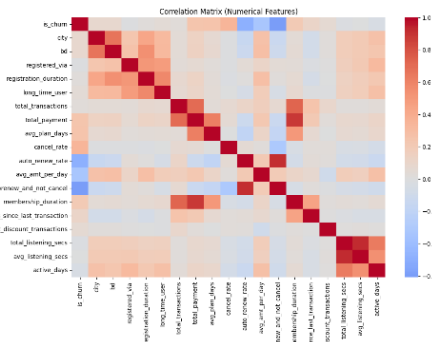




Gambar 4.1 Analisis Distribusi Seluruh Fitur Numerik.

Hasil visualisasi menunjukkan bahwa sebagian besar fitur memiliki distribusi yang tidak normal dan cenderung *skewed* ke kanan. Pola ini mencerminkan perbedaan perilaku pengguna yang signifikan dan menunjukkan bahwa pendekatan model non-linear berpotensi lebih sesuai untuk memodelkan data.

Analisis korelasi dilakukan menggunakan matriks korelasi Pearson yang divisualisasikan dalam bentuk *heatmap* ditampilkan pada Gambar 4.2.

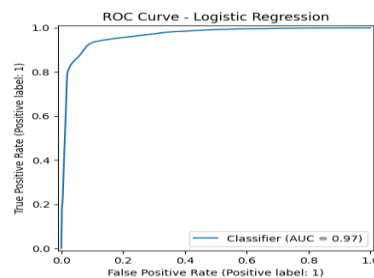


Gambar 4.2 *heatmap* Korelasi Antar Fitur

Hasil analisis menunjukkan bahwa sebagian besar fitur numerik tidak memiliki korelasi linear yang sangat tinggi satu sama lain, meskipun terdapat juga fitur yang memiliki korelasi cukup tinggi.

4.2 Evaluasi Model

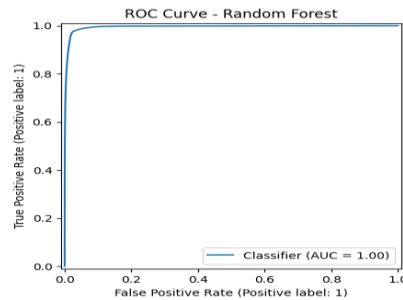
4.2.1 Logistic Regression



Model *Logistic Regression* menunjukkan performa yang cukup baik dalam memprediksi *churn* pelanggan. Model ini memiliki nilai *recall* yang tinggi pada kelas *churn*, yang berarti sebagian besar

pelanggan yang benar-benar churn berhasil terdeteksi. Namun, nilai *precision* yang lebih rendah menunjukkan masih adanya kesalahan prediksi pada pelanggan yang diperkirakan *churn*. Meskipun demikian, nilai ROC-AUC yang tinggi menandakan bahwa *Logistic Regression* mampu membedakan pelanggan *churn* dan tidak *churn* dengan baik, sehingga layak digunakan sebagai model *baseline*.

4.2.2 Random Forest



Model *random forest* menunjukkan performa yang sangat baik dalam memprediksi *churn* pelanggan. Model ini mampu menyeimbangkan *precision* dan *recall* pada kelas *churn* dengan *F1-score* yang tinggi, sehingga efektif dalam mendeteksi pelanggan yang berpotensi *churn* tanpa menghasilkan terlalu banyak kesalahan prediksi. Nilai ROC-AUC yang mendekati 1 menunjukkan kemampuan model dalam membedakan pelanggan *churn* dan tidak *churn* secara sangat baik, menjadikan *random forest* sebagai model pembandingan yang unggul dibandingkan model *baseline*.

4.3 Analisis Distribusi Statistik

Untuk memahami perbedaan karakteristik pelanggan *churn* dan tidak *churn*, dilakukan analisis distribusi nilai rata-rata fitur numerik berdasarkan status *churn*. Hasil analisis menunjukkan adanya perbedaan yang cukup signifikan pada beberapa fitur utama.

Tabel 4.3 Distribusi Statistik

<i>is churn</i>	0	1	diff
<i>total payment</i>	150.7752	345.5465	194.7712
<i>registration duration</i>	1142.4459	1214.4328	71.9869
<i>membership duration</i>	34.9119	80.2982	45.3862
<i>avg plan days</i>	29.7527	58.5166	28.7639
<i>days since last transaction</i>	13.9367	23.6882	9.7515
<i>bd</i>	11.4541	16.2932	4.8390
<i>city</i>	5.0882	7.0506	1.9625
<i>is churn</i>	0.0000	1.0000	1.0000
<i>cancel rate</i>	0.0049	0.1376	0.1328
<i>total transactions</i>	1.1498	1.2755	0.1257
<i>long time user</i>	0.7113	0.7757	0.0644
<i>pct discount transactions</i>	0.0018	0.0113	0.0095
<i>registered via</i>	6.1252	5.9776	-0.1476
<i>auto renew rate</i>	0.9258	0.3537	-0.5721
<i>autorenew and not cancel</i>	0.9205	0.1675	-0.7531

Untuk memahami perbedaan karakteristik antara pelanggan yang *churn* dan tidak *churn*, dilakukan perbandingan nilai rata-rata fitur numerik berdasarkan status *churn*. Hasil analisis menunjukkan perbedaan yang cukup signifikan pada beberapa fitur utama.

Pelanggan yang mengalami *churn* memiliki:

6. Nilai *total payment* yang lebih tinggi
7. *Membership duration* dan *registration duration* yang lebih panjang
8. *Days since last transaction* yang lebih besar
9. *Cancel rate* yang jauh lebih tinggi
10. *Auto-renew rate* dan *autorenew and not cancel* yang jauh lebih rendah

Temuan ini mengindikasikan bahwa *churn* tidak hanya dipengaruhi oleh tingkat penggunaan layanan, tetapi juga oleh pola transaksi dan kebiasaan pembaruan langganan.

4.4 Fitur Importance

Setelah dilakukan pemodelan menggunakan dua algoritma, yaitu Logistic Regression dan Random Forest, diperoleh daftar fitur penting (*feature importance*) dari masing-masing model. Selanjutnya, dilakukan perbandingan fitur penting dari kedua model untuk mengidentifikasi fitur yang muncul secara konsisten dan memiliki kontribusi signifikan dalam prediksi *churn*. Didapatkan 7 fitur penting yang dapat mempengaruhi model dalam mengklasifikasikan resiko *churn* atau tidak :

Tabel 4.4 Fitur Importance

RF	RegLog	Konsisten
<i>membership_duration</i>	<i>avg_plan_days</i>	<i>last_membership_expire</i>
<i>autorenew_and_not_cancel</i> ,	<i>total_transactions</i>	<i>total_payment</i>
<i>last_membership_expire</i>	<i>last_membership_expire</i>	<i>autorenew_and_not_cancel</i>
<i>auto_renew_rate</i>	<i>days_since_last_transaction</i>	<i>avg_plan_days</i>
<i>days_since_last_transaction</i>	<i>membership_duration</i>	<i>total_transactions</i>
<i>avg_plan_days</i>	<i>autorenew_and_not_cancel</i>	<i>membership_duration</i>
<i>total_transactions</i>	<i>total_payment</i>	<i>days_since_last_transaction</i>
<i>total_payment</i>		

Hasil analisis menunjukkan terdapat 7 fitur yang paling konsisten dan valid dalam mempengaruhi prediksi *churn*, yaitu:

- *last_membership_expire* = Tanggal berakhirnya masa langganan terakhir
- *total_payment* = Total pembayaran pelanggan
- *autorenew_and_not_cancel* = Status perpanjangan otomatis (auto-renew) dan tidak pernah membatalkan langganan
- *avg_plan_days* = Rata-rata durasi paket langganan (dalam hari)
- *total_transactions* = Total jumlah transaksi pelanggan
- *days_since_last_transaction* = Jumlah hari sejak transaksi terakhir
- *membership_duration* = Lama waktu pelanggan menjadi anggota

Fitur - fitur tersebut kemudian dianalisis lebih lanjut untuk memperoleh insight terkait perilaku *churn* pelanggan.

4.5 Visualisasi Kibana (Elastic)

Analisis dilakukan dengan memanfaatkan Elasticsearch dan Kibana untuk memvisualisasikan persebaran data setiap fitur penting terhadap label *churn*, sehingga pola hubungan antara karakteristik pelanggan dan risiko *churn* dapat diamati secara lebih jelas.

Tabel 4.5 Visualisasi Data Importance terhadap Churn

	<p><i>Last membership expire + autorenew and not cancel</i> Heatmap menunjukkan bahwa pelanggan yang tidak menggunakan fitur auto-renew memiliki rata-rata <i>churn</i> yang lebih tinggi, khususnya pada periode berakhirnya langganan antara tahun 2017 hingga 2022. Sebaliknya, pelanggan dengan auto-renew cenderung memiliki risiko <i>churn</i> yang lebih rendah secara konsisten.</p>
	<p><i>Total Payment + Day Last Transaction</i> Heatmap menunjukkan bahwa <i>days_since_last_transaction</i> memiliki pengaruh yang sangat kuat terhadap risiko <i>churn</i>.</p>

	<p>Terlepas dari besarnya total pembayaran, pelanggan yang tidak melakukan transaksi selama lebih dari 60 hari cenderung memiliki tingkat <i>churn</i> yang tinggi. Sebaliknya, pelanggan yang masih aktif bertransaksi dalam 30 hari terakhir menunjukkan risiko <i>churn</i> yang lebih rendah, bahkan pada kelompok dengan total pembayaran yang relatif kecil. Hal ini mengindikasikan bahwa aktivitas transaksi terbaru merupakan indikator yang kuat terhadap <i>churn</i> dibandingkan nilai pembayaran semata.</p>
	<p>Avg Plan Days</p> <p>Bar chard menunjukkan pelanggan dengan durasi paket yang lebih panjang justru memiliki tingkat <i>churn</i> yang lebih tinggi. Pada kelompok pelanggan dengan <i>avg_plan_days</i> di atas 150 hari, hampir seluruh pelanggan mengalami <i>churn</i>. Sebaliknya, pelanggan dengan paket berjangka pendek (0–50 hari) cenderung tidak <i>churn</i>. Temuan ini mengindikasikan bahwa paket jangka panjang tidak selalu menjamin retensi pelanggan, karena pelanggan cenderung berhenti setelah paket berakhir tanpa melakukan perpanjangan dan sebaliknya jika paket jangka pendek memungkinkan pelanggan akan terus berlangganan.</p>
	<p>Membership Duration</p> <p>Data menunjukkan bahwa pelanggan dengan durasi keanggotaan 0–90 hari memiliki risiko <i>churn</i> yang relatif rendah karena masih berada pada fase awal penggunaan layanan. Risiko <i>churn</i> tertinggi terjadi pada pelanggan dengan durasi keanggotaan 90–365 hari, yang menunjukkan fase evaluasi antara ekspektasi dan manfaat layanan. Sementara itu, pelanggan dengan durasi keanggotaan lebih dari satu tahun cenderung memiliki tingkat <i>churn</i> yang lebih stabil, mencerminkan terbentuknya loyalitas pelanggan dalam jangka panjang</p>
	<p>Total Transactions</p> <p>Visualisasi menunjukkan bahwa risiko <i>churn</i> meningkat seiring bertambahnya jumlah transaksi pelanggan. Pelanggan dengan total transaksi rendah (1–20) cenderung memiliki tingkat <i>churn</i> yang rendah karena masih berada pada fase awal penggunaan layanan. Namun, pada kelompok pelanggan dengan transaksi menengah (20–50), risiko <i>churn</i> mulai meningkat, yang mengindikasikan fase evaluasi layanan. Risiko <i>churn</i> tertinggi ditemukan pada pelanggan dengan jumlah transaksi sangat tinggi (>50), menunjukkan bahwa pelanggan yang sudah sangat aktif justru lebih berpotensi menghentikan langganan setelah kebutuhan mereka terpenuhi atau mengalami kekecewaan terhadap layanan.</p>

5. Insight dan Rekomendasi

5.1 Insight

- Masa akhir langganan merupakan titik paling rawan *churn***
last_membership_expire menjadi faktor paling berpengaruh terhadap *churn*. Risiko *churn* meningkat ketika pelanggan berada pada periode mendekati atau setelah masa langganan berakhir, sehingga fase ini menjadi titik kritis dalam mempertahankan pelanggan.
- Auto-renew berperan penting dalam menekan *churn***
Pelanggan yang mengaktifkan *autorenew* and *not cancel* memiliki risiko *churn* yang jauh lebih rendah. Hal ini menunjukkan bahwa mekanisme perpanjangan otomatis mampu menjaga keberlanjutan langganan tanpa ketergantungan pada tindakan manual pelanggan.
- Aktivitas transaksi lebih menentukan dibandingkan nilai pembayaran**
Fitur *days_since_last_transaction* menunjukkan bahwa pelanggan yang lama tidak bertransaksi

memiliki risiko *churn* lebih tinggi, hal ini mengindikasikan bahwa keterlibatan terbaru pelanggan merupakan indikator *churn* yang kuat.

4. **Risiko *churn* meningkat pada fase evaluasi pelanggan**
Analisis *membership_duration* menunjukkan bahwa *churn* paling tinggi terjadi pada pelanggan dengan durasi keanggotaan menengah (90–365 hari), yang mencerminkan fase evaluasi antara ekspektasi dan manfaat layanan.
5. **Paket langganan jangka panjang tidak selalu meningkatkan retensi**
Fitur *avg_plan_days* menunjukkan bahwa pelanggan dengan durasi paket yang lebih panjang cenderung memiliki tingkat *churn* lebih tinggi setelah paket berakhir, sehingga paket jangka panjang tidak selalu menjamin keberlanjutan pelanggan.
6. ***Churn* juga terjadi pada pelanggan sangat aktif**
Berdasarkan *total_transactions*, pelanggan dengan jumlah transaksi yang sangat tinggi menunjukkan risiko *churn* yang meningkat, yang mengindikasikan kemungkinan kejenuhan atau telah terpenuhinya kebutuhan pelanggan.
7. ***Churn* pelanggan bernilai tinggi berdampak signifikan secara finansial**
Fitur *total_payment* menunjukkan bahwa *churn* pada pelanggan dengan nilai pembayaran tinggi berpotensi memberikan dampak finansial besar, sehingga kelompok ini perlu menjadi prioritas dalam strategi retensi.

5.2 Rekomendasi

1. *Last Membership Expire* dan *Auto-renew and Not Cancel*
Perusahaan disarankan untuk melakukan intervensi sebelum masa langganan pelanggan berakhir dengan mengirimkan pengingat perpanjangan secara berkala dan perusahaan perlu mendorong aktivasi fitur perpanjangan otomatis (*auto-renew*) dan melalui pemberitahuan manfaat serta bonus untuk pembeli yang mengaktifkan *auto renew*
2. *Days Since Last Transaction* dan *Total Payment*
Perusahaan disarankan untuk memantau pelanggan yang memiliki jeda waktu transaksi yang panjang, termasuk pelanggan dengan nilai pembayaran yang tinggi. Kampanye re-engagement melalui notifikasi personal atau penawaran promosi dapat dilakukan untuk meningkatkan kembali aktivitas transaksi pelanggan.
3. *Membership Duration*
Perusahaan disarankan untuk menerapkan program loyalitas serta memberikan manfaat eksklusif berdasarkan lama keanggotaan pelanggan, khususnya bagi pelanggan dengan durasi keanggotaan menengah, guna meningkatkan retensi pelanggan.
4. *Avg Plan Days*
Perusahaan disarankan untuk mengevaluasi dan menyesuaikan durasi paket langganan agar lebih fleksibel dan sesuai dengan pola penggunaan pelanggan, sehingga dapat mendukung keberlanjutan langganan dalam jangka panjang.
5. *Total Transactions* Perusahaan disarankan untuk memberikan layanan atau penawaran yang bersifat personal kepada pelanggan dengan tingkat transaksi yang tinggi guna menjaga keterlibatan dan mengurangi risiko *churn* akibat kejenuhan layanan.