



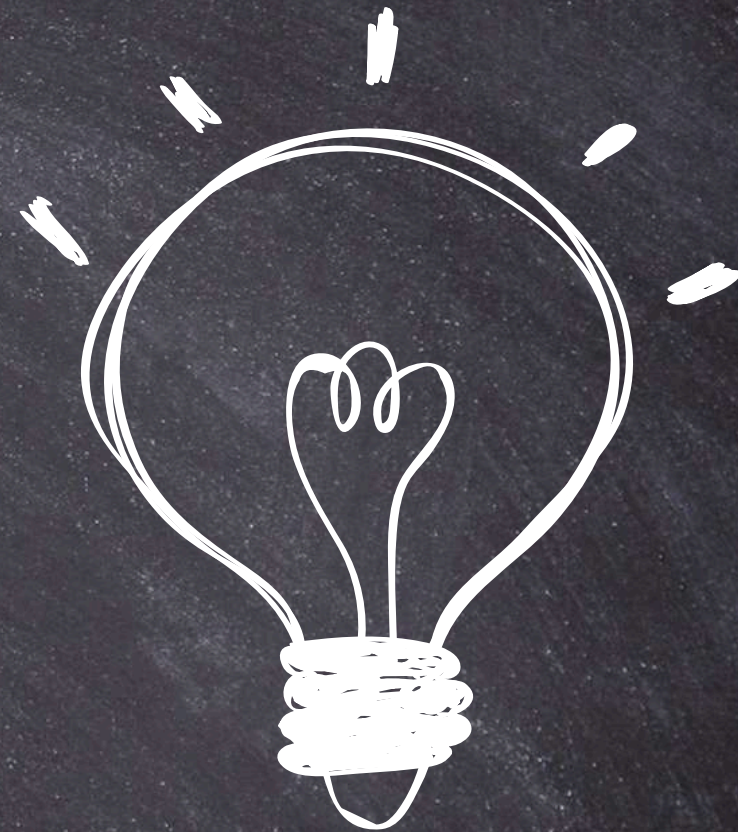
TELCO CUSTOMER CHURN



BY NABILA SETYA UTAMI



LATAR BELAKANG



- 
- Tantangan churn pelanggan di industri telekomunikasi.
 - Dampak churn terhadap pendapatan bisnis.
 - Pentingnya memahami faktor penyebab churn dan strategi berbasis data.

MENGAPA CHURN MENJADI MASALAH?

1. Penurunan Pendapatan

- Tingginya churn langsung berdampak pada penurunan pendapatan perusahaan.

2. Biaya Pelanggan Baru

- Biaya akuisisi pelanggan baru jauh lebih tinggi dibanding mempertahankan pelanggan lama.

3. Kekurangan Pelanggan Baru

- Ketidakmampuan menarik pelanggan baru memperburuk dampak negatif churn.

TUJUAN

- Memahami pelanggan berisiko churn.
- Mengidentifikasi penyebab churn.
- Merancang strategi retensi yang efektif.
- Mengoptimalkan anggaran retensi.

PENDEKATAN ANALITIK

Menggunakan model machine learning untuk memprediksi churn berdasarkan pola dan riwayat pelanggan, sehingga perusahaan dapat mengidentifikasi dan mencegah churn secara efektif.

PROSES ANALITIK CHURN

1

DATA PREPARATION

Bersihkan data dan tangani missing values serta outliers.

2

EDA (EXPLORATORY DATA ANALYSIS)

Identifikasi faktor utama penyebab terjadinya churn.

3

MODELLING

Bangun model ML untuk prediksi churn.

4

OPTIMIZATION

Lakukan hyperparameter tuning.

5

TEST MODEL

mengevaluasi performa model pada data test setelah dilakukan optimization

6

INSIGHTS

Analisis feature importance untuk faktor prediktor utama.

DATA PREPARATION

```
data.duplicated().sum()
```

```
np.int64(77)
```

Data Duplikasi

Ditemukan **77 baris** duplikasi.

Langkah: Hapus duplikasi untuk memastikan data bersih dan tidak berulang.

	Nama Kolom	Tipe Data	Missing Value	% Missing Value	N Unique	Unique Sample
0	Dependents	object	0	0.0	2	[Yes, No]
1	tenure	int64	0	0.0	73	[9, 14, 64]
2	OnlineSecurity	object	0	0.0	3	[No, Yes, No internet service]
3	OnlineBackup	object	0	0.0	3	[No, Yes, No internet service]
4	InternetService	object	0	0.0	3	[DSL, Fiber optic, No]
5	DeviceProtection	object	0	0.0	3	[Yes, No internet service, No]
6	TechSupport	object	0	0.0	3	[Yes, No, No internet service]
7	Contract	object	0	0.0	3	[Month-to-month, Two year, One year]
8	PaperlessBilling	object	0	0.0	2	[Yes, No]
9	MonthlyCharges	float64	0	0.0	1422	[72.9, 82.65, 47.85]
10	Churn	object	0	0.0	2	[Yes, No]

Dataset Overview

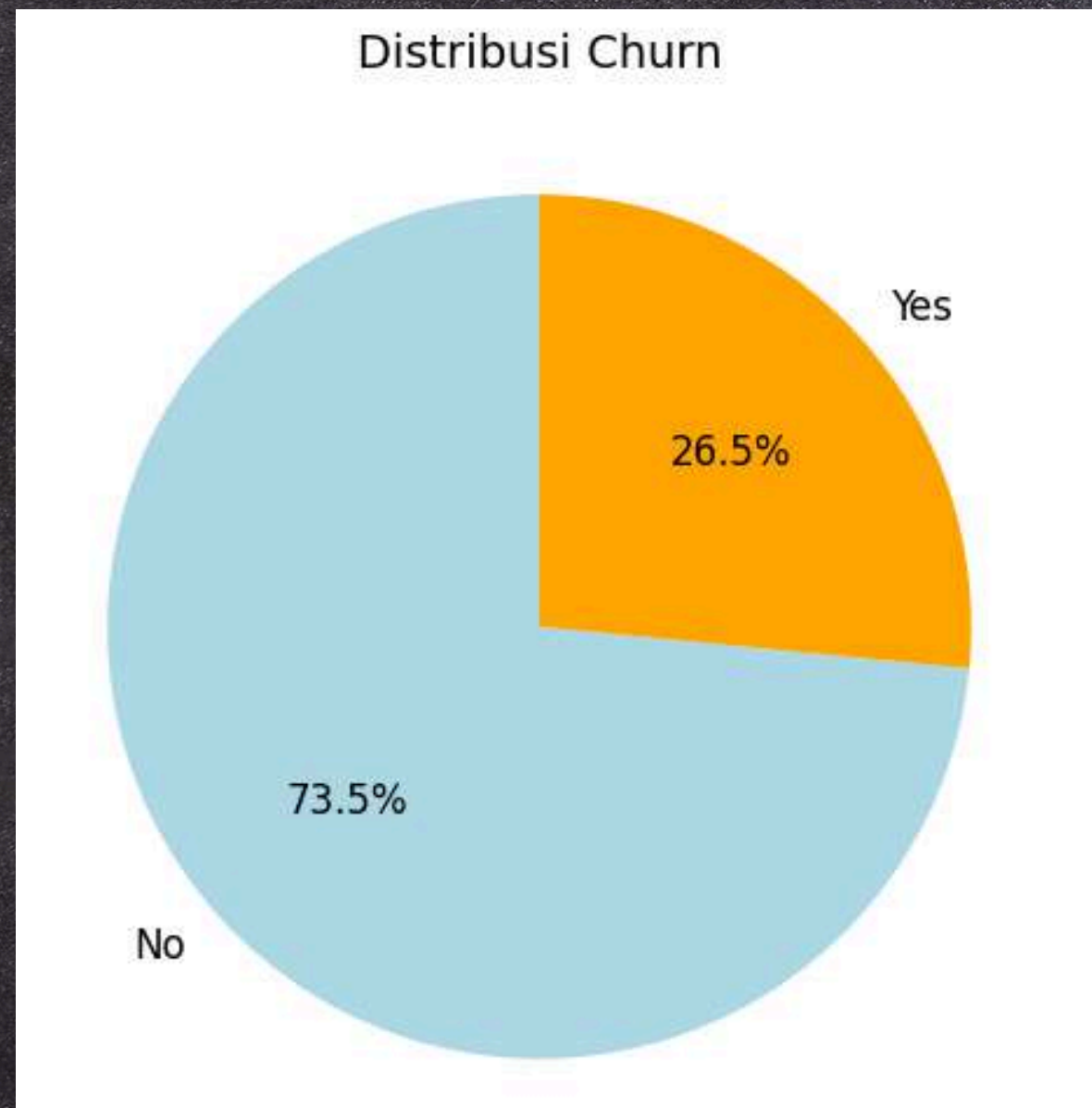
Tidak Ada Missing Value: Semua kolom memiliki data lengkap.

	Nama Kolom	Tipe Data	Missing Value	% Missing Value	N Unique	Unique Sample
0	Dependents	object	0	0.0	2	[Yes, No]
1	tenure	int64	0	0.0	73	[9, 14, 64]
2	OnlineSecurity	object	0	0.0	2	[No, Yes]
3	OnlineBackup	object	0	0.0	2	[No, Yes]
4	InternetService	object	0	0.0	3	[DSL, Fiber optic, No]
5	DeviceProtection	object	0	0.0	2	[Yes, No]
6	TechSupport	object	0	0.0	2	[Yes, No]
7	Contract	object	0	0.0	3	[Month-to-month, Two year, One year]
8	PaperlessBilling	object	0	0.0	2	[Yes, No]
9	MonthlyCharges	float64	0	0.0	1422	[72.9, 82.65, 47.85]
10	Churn	object	0	0.0	2	[Yes, No]

Dataset Overview (After Cleaning)

Nilai **"No internet service"** pada OnlineSecurity, OnlineBackup, DeviceProtection, dan TechSupport diubah menjadi **"No"** untuk konsistensi.

2 EDA (Exploratory Data Analysis)

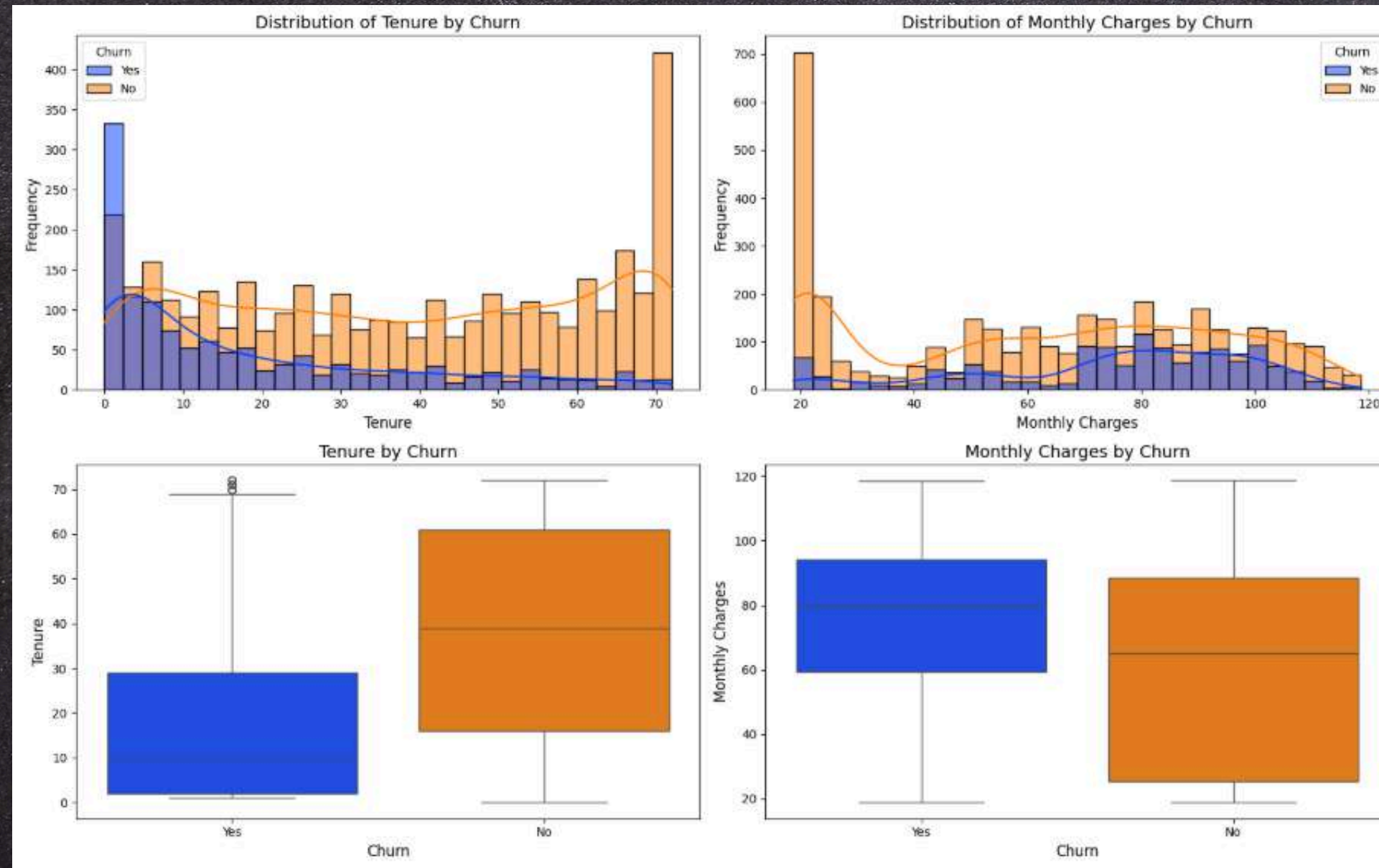


Yes: Pelanggan berhenti layanan.

No: Pelanggan tetap.

- Ketidakseimbangan Data: Non-Churn hampir 3x lebih banyak dari Churn.
- Catatan: Imbalance ini dapat memengaruhi performa model prediktif.

2 EDA (Exploratory Data Analysis)



- **Tenure & Churn:** Pelanggan dengan durasi langganan lebih lama cenderung tidak churn.
- **Monthly Charges & Churn:** Pelanggan dengan tagihan bulanan tinggi lebih rentan churn.
- **Outlier:** Terdapat outlier pada Tenure untuk pelanggan churn, memerlukan analisis lebih lanjut.

3 MODELLING

• Target

Nilai kategorikal pada kolom Churn diubah:

'Yes' → 1 (Pelanggan churn).

'No' → 0 (Pelanggan tidak churn).

Transformasi dilakukan untuk mempermudah analisis dan modeling.

• Fitur

	Feature	Chi-Square Statistics	P Value	Berpengaruh Terhadap Target(?)
0	Dependents	138.497851	5.671674e-32	True
1	OnlineSecurity	141.866889	1.039889e-32	True
2	OnlineBackup	48.493354	3.314102e-12	True
3	InternetService	505.302778	1.883198e-110	True
4	DeviceProtection	19.383346	1.069355e-05	True
5	TechSupport	130.040881	4.014244e-30	True
6	Contract	851.729986	1.119911e-185	True
7	PaperlessBilling	166.458582	4.392323e-38	True

Hasil Uji Chi-Square

Semua fitur memiliki pengaruh signifikan terhadap target ($P \text{ Value} < \text{ambang signifikan}$). Semua fitur akan digunakan dalam pengujian lanjutan.



3 MODELLING

• Data Splitting

Dataset dibagi menjadi:

- 80% Data Latih (train)
- 20% Data Uji (test)

• Preprocessing

- **One-Hot Encoding:** Mengubah data kategorikal menjadi numerik (0/1).
- **Fitur:** Dependents, PaperlessBilling, OnlineSecurity, OnlineBackup, InternetService, DeviceProtection, TechSupport, Contract
- **StandardScaler:** Normalisasi data numerik (rata-rata 0, standar deviasi 1).
- **Fitur:** Tenure dan MonthlyCharges

3

MODELLING

- **Benchmarking**

proses membandingkan performa beberapa model machine learning untuk menentukan model terbaik dalam menyelesaikan masalah klasifikasi

Model yang Dibandingkan

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Decision Tree
- Random Forest
- XGBoost
- LightGBM
- Support Vector Classifier (SVC)

3 MODELLING

- Cross Validation

	mean recall	sdev
model		
Logistic Regression	0.531068	0.009513
LGBM	0.526214	0.034408
XGB	0.518447	0.042917
Random Forest	0.490291	0.036714
Decision Tree	0.489320	0.027869
KNN	0.482524	0.042386
SVC	0.462136	0.029029

Hasil Cross Validation Tiap Model

- Logistic Regression: Terbaik, mean recall tertinggi (0.531) dan stabilitas tinggi (sdev: 0.009).
- LGBM: Alternatif kompetitif, recall 0.526 dengan variabilitas rendah.
- SVC: Terendah, recall 0.462.

3 MODELLING

- Cross Validation

	mean recall	sdev
model		
Logistic Regression	0.531068	0.009513
LGBM	0.526214	0.034408
XGB	0.518447	0.042917
Random Forest	0.490291	0.036714
Decision Tree	0.489320	0.027869
KNN	0.482524	0.042386
SVC	0.462136	0.029029

Hasil Cross Validation Tiap Model

- Logistic Regression: Terbaik, mean recall tertinggi (0.531) dan stabilitas tinggi (sdev: 0.009).
- LGBM: Alternatif kompetitif, recall 0.526 dengan variabilitas rendah.
- SVC: Terendah, recall 0.462.

METODE BALANCING

- **Oversampling:** Tambah data kelas minoritas (SMOTE, ADASYN, RandomOverSampler).
- **Undersampling:** Kurangi data kelas mayoritas (RandomUnderSampler, NearMiss, CondensedNearestNeighbour).
- **Model Berbobot:** Terapkan bobot untuk kelas minoritas dalam model.

	Method	Metode Balancing	Recall
0	Logistic Regression	ADASYN	0.82
1	Logistic Regression	SMOTE	0.79
2	Logistic Regression	RandomUnderSampler	0.79
3	Logistic Regression	RandomOverSampler	0.78
4	KNN	RandomUnderSampler	0.75
5	Logistic Regression	CondensedNearestNeighbour	0.74
6	Random Forest	RandomUnderSampler	0.74
7	KNN	ADASYN	0.73
8	KNN	RandomOverSampler	0.70
9	KNN	SMOTE	0.68
10	KNN	CondensedNearestNeighbour	0.68
11	Logistic Regression	NearMiss	0.64
12	Random Forest	CondensedNearestNeighbour	0.59
13	Random Forest	ADASYN	0.57
14	Random Forest	SMOTE	0.56
15	KNN	NearMiss	0.55
16	Random Forest	NearMiss	0.53
17	Random Forest	RandomOverSampler	0.52

Hasil Balancing dan Model

- **Logistic Regression + ADASYN:** Recall tertinggi (0.82), terbaik untuk kelas minoritas.
- **KNN + RandomUnderSampler:** Recall 0.75, terbaik untuk KNN.
- **Random Forest + RandomUnderSampler:** Recall 0.74, optimal untuk Random Forest.

HYPERPARAMETER TUNING

4

Proses mengoptimalkan parameter model untuk performa terbaik.

Tujuan:

- Tingkatkan performa model.
- Kurangi overfitting/underfitting.
- Sesuaikan model dengan dataset.
- Efisiensi tanpa mengganti algoritma.

Hasil Hyperparameter Tuning

Logistic Regression:

	model_C	model_max_iter	model_penalty	model_solver
0	1	100	l1	liblinear

- Kombinasi terbaik:
 - C: 1
 - max_iter: 100
 - penalty: l1
 - solver: liblinear

KNN (K-Nearest Neighbors):

	model_metric	model_n_neighbors	model_weights
0	euclidean	9	uniform

- Kombinasi terbaik:
 - metric: euclidean
 - n_neighbors: 9
 - weights: uniform

Random Forest:

	model_class_weight	model_max_depth	model_min_samples_leaf	model_min_samples_split	model_n_estimators
0	None	10	4	2	50

- Kombinasi terbaik:
 - class_weight: None
 - max_depth: 10
 - min_samples_leaf: 4
 - min_samples_split: 2
 - n_estimators: 50

5

TEST MODEL

Logistic Regression:

	Train Recall	Test Recall
model		
Logistic Regression	0.828155	0.825581

- Train Recall: 0.828
- Test Recall: 0.826 (stabil dan terbaik untuk generalisasi).

KNN:

	Train Recall	Test Recall
model		
KNN	0.840777	0.782946

- Train Recall: 0.841
- Test Recall: 0.783 (sedikit menurun, masih baik).

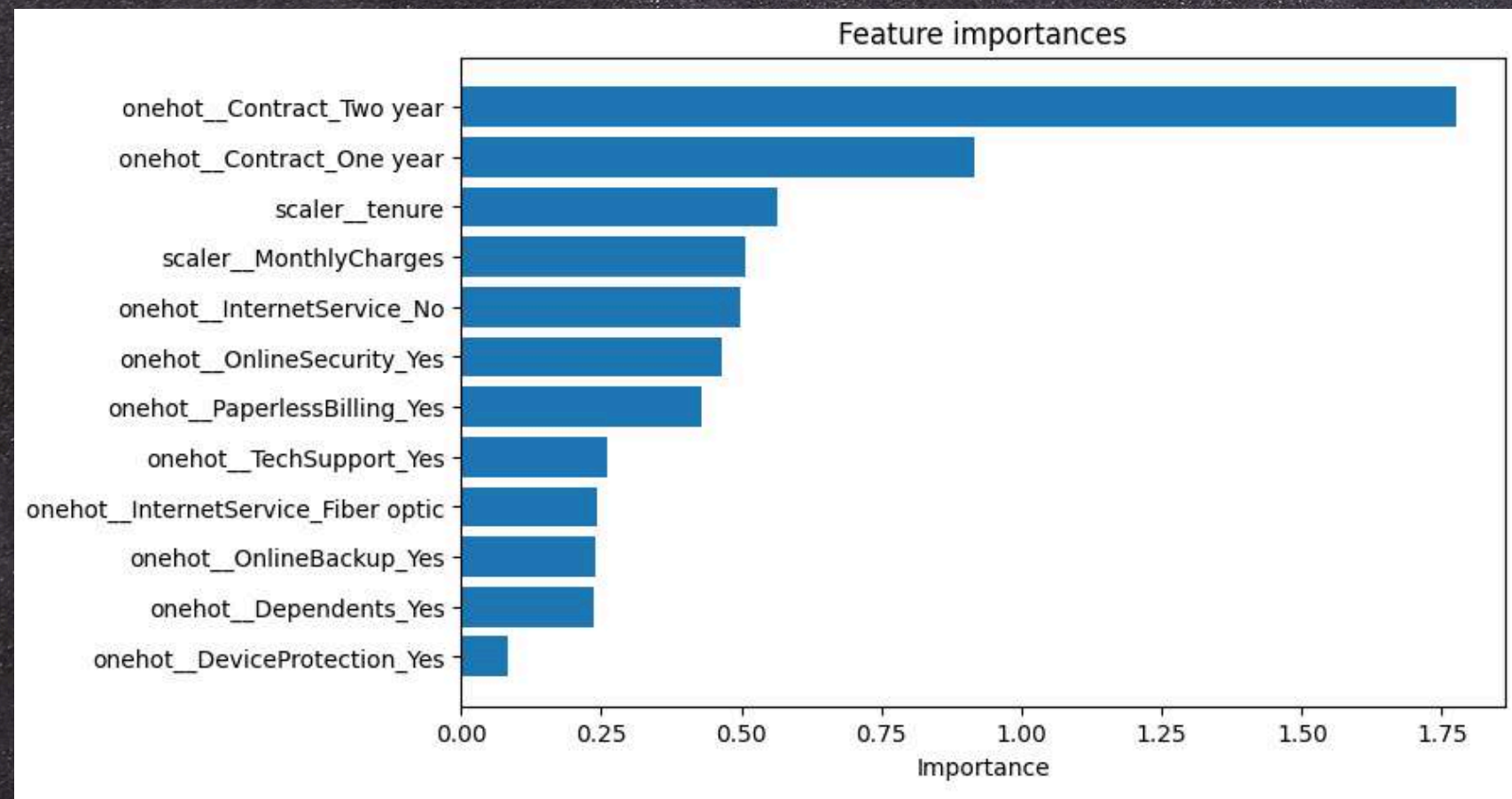
Random Forest:

	Train Recall	Test Recall
model		
Random Forest	0.881553	0.775194

- Train Recall: 0.882
- Test Recall: 0.775 (potensi overfitting lebih tinggi dibandingkan Logistic Regression).

6

FEATURE IMPORTANCE



- **Kontrak:** Fitur terpenting, kontrak 2 tahun dan 1 tahun memiliki pengaruh terbesar pada prediksi churn.
- **Tenure & MonthlyCharges:** Lama berlangganan dan biaya bulanan juga signifikan.
- **Layanan Internet:** Penggunaan layanan dan fitur keamanan seperti OnlineSecurity turut memengaruhi churn.
- **Fitur Tambahan:** PaperlessBilling, TechSupport, dan DeviceProtection memiliki dampak lebih kecil.



KESIMPULAN

DOMAIN TEKNIS

Hasil Model Logistic Regression

- Tuning Model: Recall kelas minoritas meningkat dari 0.82 ke 0.83.
- Accuracy: Tetap di 72%, namun macro average recall naik dari 0.75 ke 0.76.
- Performa Kelas Minoritas: Precision & F1-score meningkat.

Tuning berhasil meningkatkan kemampuan model mendeteksi churn, terutama pada kelas minoritas.

KESIMPULAN


DOMAIN BISNIS

- **Jenis Kontrak:** Pelanggan kontrak bulanan lebih rentan churn dibanding kontrak tahunan, tetapi potensi retensi mereka tinggi dengan strategi yang tepat.
- **Durasi Berlangganan (Tenure):** Pelanggan lama lebih cenderung churn, menekankan pentingnya inovasi untuk menjaga kepuasan.
- **Layanan Internet:** Pelanggan Fiber Optic lebih bernilai dibanding DSL.
- **Biaya Bulanan:** Biaya tinggi meningkatkan risiko churn, tetapi memberikan dampak bisnis lebih besar.
- **Dukungan Teknis:** Pelanggan tanpa dukungan teknis berkontribusi lebih positif, menunjukkan perlunya evaluasi efektivitas layanan tersebut.
- **Faktor Lain:** Online Security, Dependents, dan Online Backup juga memengaruhi retensi pelanggan.

REKOMENDASI



- **Promosi Kontrak Jangka Panjang:** Tawarkan diskon atau manfaat untuk pelanggan kontrak bulanan beralih ke kontrak tahunan.
- **Keterlibatan Pelanggan Lama:** Tingkatkan loyalitas dengan program eksklusif dan inovasi layanan.
- **Ekspansi Fiber Optic:** Perluas jangkauan dan tingkatkan kualitas layanan Fiber Optic.
- **Strategi Harga:** Evaluasi harga dengan menambahkan manfaat untuk pelanggan biaya tinggi.
- **Perbaikan Dukungan Teknis:** Tingkatkan kualitas dan kecepatan layanan untuk meningkatkan kepuasan.
- **Penguatan Layanan Tambahan:** Optimalkan Online Security, Backup, dan Device Protection untuk nilai lebih.
- **Paket Keluarga:** Sasar pelanggan dengan tanggungan (Dependents) untuk retensi lebih baik.

The background is a dark, textured surface resembling a chalkboard. It is decorated with several hand-drawn elements: a light blue wavy line in the top left corner, a yellow wavy line in the top right corner, a yellow smiley face in the upper right quadrant, three pink diagonal lines to the left of the text, three pink diagonal lines to the right of the text, a purple oval around the word 'KASIH', a purple wavy line in the bottom left corner, and a yellow wavy line in the bottom right corner.

TERIMA
KASIH