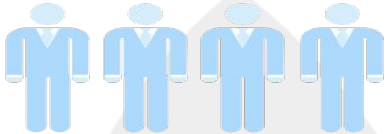
BINAR

Comparison of Logistic Regression and Random Forest in Application to Imbalance class and Multicollinearity using Stratified Cross-Validation to Predict Telecom Customer Churn

Table of Content

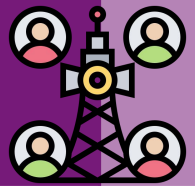
Business Use Case

BINAR



Telekomunikasi sebagai industri yang dinamis

Telekomunikasi merupakan industri yang terus berkembang seiring dengan perkembangan zaman



Persaingan yang Ketat di Industri Telekomunikasi

Persaingan ini memaksa perusahaan untuk terus berinovasi dan menyesuaikan diri dengan kebutuhan pelanggan

Pentingnya Prediksi Perilaku Pelanggan

Perusahaan perlu memiliki pemahaman yang baik tentang perilaku pelanggan dan membuat Model prediksi perilaku pelanggan untuk memetakan strategi bisnis yang efektif guna mempertahankan dan menarik pelanggan baru

DATA OVERVIEW

BINAR

Setelah mengimport data dan membuat data frame terdapat total 20 kolom yang terdiri dari :

- State : Kode Negara Customer
- Account_length : Lamanya berlangganan
- Area_code : Kode Area tempat tinggal
- International_plan : Rencana Panggilan Internasional
- Voice_mail_plan : Rencana kotak suara
- Number_vmail_messages : Frekuensi penggunaan kotak suara
- Total_day_minutes : Jumlah total menit panggilan siang hari
- Total_day_calls : Jumlah total panggilan siang hari
- Total_day_charge : Jumlah total biaya panggilan siang hari
- Total_eve_minutes : Jumlah total menit panggilan sore hari
- Total_eve_calls : Jumlah total panggilan sore hari
- Total_eve_charge : Jumlah total biaya panggilan sore hari
- Total_night_minutes : Jumlah total menit panggilan malam hari
- Total_night_calls : Jumlah total panggilan malam hari
- Total_night_charge : Jumlah total biaya panggilan malam hari
- Total_intl_minutes : Jumlah total menit panggilan international
- Total_intl_calls : Jumlah total panggilan international
- Total_intl_charge : jumlah total biaya panggilan malam har
- Number_customer_service_calls : Jumlah panggilan ke pusat layanan
- Churn : Apakah customer berhenti berlangganan

Variable churn merupakan target prediksi (Y) yang terdiri dari kategori yes, no (binary classification)

Variabel state sampai variabel number_customer_service_calls merupakan fitur (X) yang digunakan untuk memprediksi target

EXPLORATORY DATA ANALYSIS

BINAR

Data Information

```
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4250 entries, 0 to 4249
Data columns (total 20 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   state                         4250 non-null   object
 1   account_length                4250 non-null   int64
 2   area_code                     4250 non-null   object
 3   international_plan            4250 non-null   object
 4   voice_mail_plan               4250 non-null   object
 5   number_vmail_messages         4250 non-null   int64
 6   total_day_minutes             4250 non-null   float64
 7   total_day_calls               4250 non-null   int64
 8   total_day_charge              4250 non-null   float64
 9   total_eve_minutes             4250 non-null   float64
 10  total_eve_calls               4250 non-null   int64
 11  total_eve_charge              4250 non-null   float64
 12  total_night_minutes           4250 non-null   float64
 13  total_night_calls             4250 non-null   int64
 14  total_night_charge            4250 non-null   float64
 15  total_intl_minutes            4250 non-null   float64
 16  total_intl_calls              4250 non-null   int64
 17  total_intl_charge             4250 non-null   float64
 18  number_customer_service_calls 4250 non-null   int64
 19  churn                         4250 non-null   object
```

- Terdapat total 20 variabel dimana variabel churn merupakan target dan sisanya merupakan fitur.
- Data tidak memiliki nilai missing value yang ditandai dengan jumlah count baris data setiap kolom sama dengan 4250 record.
- Pada dataset ini 5 variabel diantaranya bertipe string, 8 variabel bertipe data float dan sisanya variabel dengan tipe data integer

EXPLORATORY DATA ANALYSIS

BINAR

Check Duplicate and Unique value

```
[ ] print('Nilai duplicated data train:', df_train.duplicated().sum())
```

```
Nilai duplicated data train: 0
```

```
print('Nilai unique pada data train :\n', df_train.nunique())
```

```
Nilai unique pada data train :
 state                          51
account_length                215
area_code                       3
international_plan              2
voice_mail_plan                 2
number_vmail_messages          46
total_day_minutes            1843
total_day_calls               120
total_day_charge             1843
total_eve_minutes            1773
total_eve_calls               123
total_eve_charge             1572
total_night_minutes          1757
total_night_calls             128
total_night_charge            992
total_intl_minutes            168
total_intl_calls               21
total_intl_charge             168
number_customer_service_calls  10
churn                           2
dtype: int64
```

Tidak terdapat data yang duplikat

Data unique pada setiap variabel menunjukkan bahwa data sudah konsisten dan valid

EXPLORATORY DATA ANALYSIS

BINAR

Statistics Descriptive

```
df_train.describe().T
```

	count	mean	std	min	25%	50%	75%	max
account_length	4250.0	100.236235	39.698401	1.0	73.0000	100.00	127.0000	243.00
number_vmail_messages	4250.0	7.631765	13.439882	0.0	0.0000	0.00	16.0000	52.00
total_day_minutes	4250.0	180.259600	54.012373	0.0	143.3250	180.45	216.2000	351.50
total_day_calls	4250.0	99.907294	19.850817	0.0	87.0000	100.00	113.0000	165.00
total_day_charge	4250.0	30.644682	9.182096	0.0	24.3650	30.68	36.7500	59.76
total_eve_minutes	4250.0	200.173906	50.249518	0.0	165.9250	200.70	233.7750	359.30
total_eve_calls	4250.0	100.176471	19.908591	0.0	87.0000	100.00	114.0000	170.00
total_eve_charge	4250.0	17.015012	4.271212	0.0	14.1025	17.06	19.8675	30.54
total_night_minutes	4250.0	200.527882	50.353548	0.0	167.2250	200.45	234.7000	395.00
total_night_calls	4250.0	99.839529	20.093220	0.0	86.0000	100.00	113.0000	175.00
total_night_charge	4250.0	9.023892	2.265922	0.0	7.5225	9.02	10.5600	17.77
total_intl_minutes	4250.0	10.256071	2.760102	0.0	8.5000	10.30	12.0000	20.00
total_intl_calls	4250.0	4.426353	2.463069	0.0	3.0000	4.00	6.0000	20.00
total_intl_charge	4250.0	2.769654	0.745204	0.0	2.3000	2.78	3.2400	5.40
number_customer_service_calls	4250.0	1.559059	1.311434	0.0	1.0000	1.00	2.0000	9.00

Total panggilan terbanyak customer terjadi pada sore hari dengan rerata 100 panggilan/hari dibandingkan siang dan malam hari

Namun, Rata-rata **total lama panggilan** tertinggi customer terjadi pada malam hari sebanyak 201 menit/hari dibandingkan siang dan sore hari

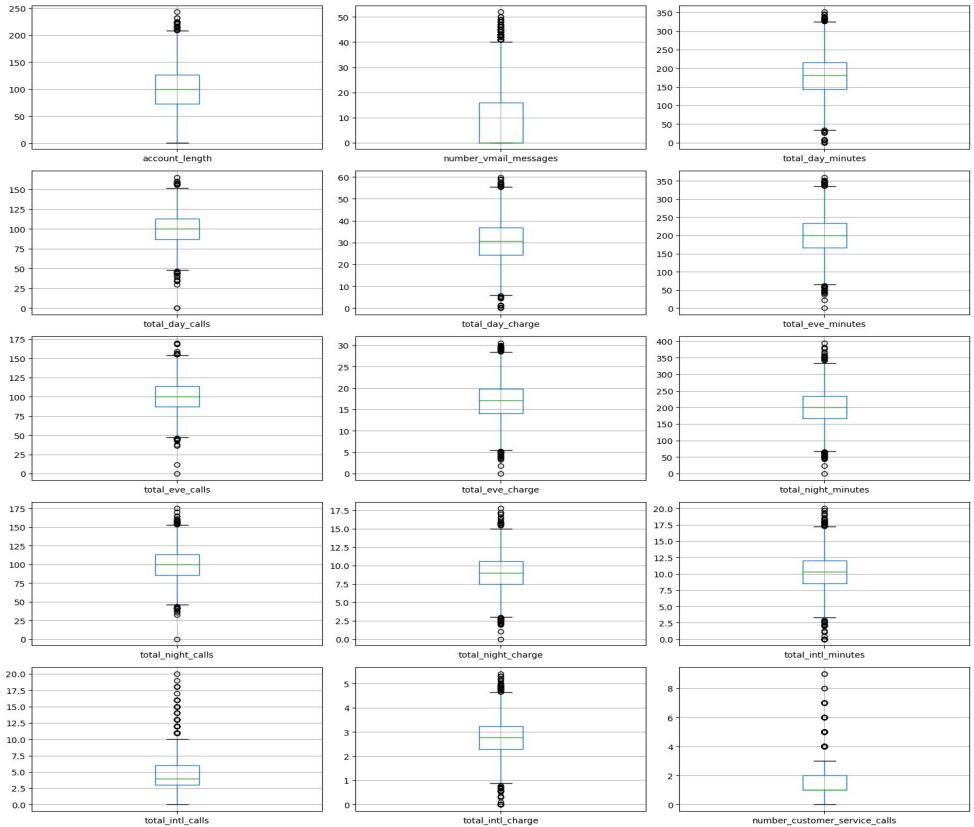
Sedangkan **total biaya panggilan** yang dikenakan termahal terjadi pada siang hari dengan rerata sebesar 31 dollar/hari dibandingkan sore dan malam hari

Tingginya angka panggilan customer service

EXPLORATORY DATA ANALYSIS

BINAR

Boxplot

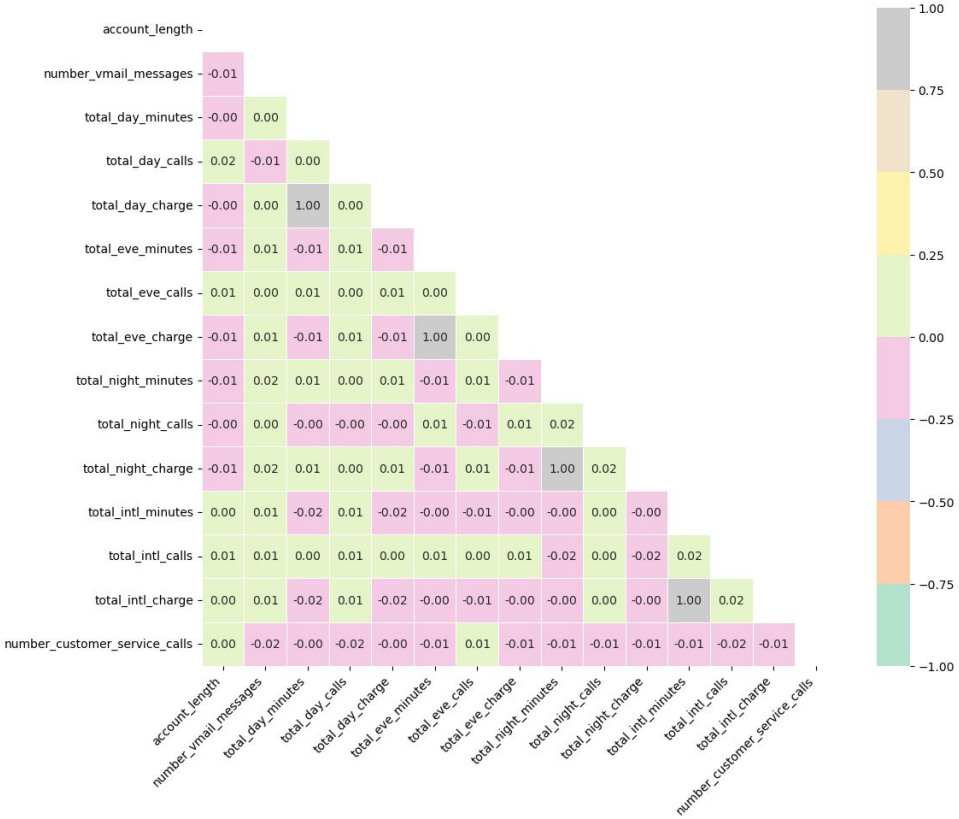


Terdapat outlier pada semua fitur numerik namun outlier ini masih tergolong valid dan bukan merupakan kesalahan input

EXPLORATORY DATA ANALYSIS

BINAR

Correlation Heatmap



Terdapat empat pasangan variabel bebas yang memiliki tingkat hubungan sempurna dan searah yang menunjukkan bahwa semakin lama durasi panggilan dikenakan biaya yang semakin besar

Tingkat hubungan yang sempurna ini menandakan bahwa adanya **multikolinearitas** dalam data yang dapat menyebabkan kesalahan interpretasi dan mengurangi kualitas dari hasil prediksi. Oleh karena itu, dibutuhkan metode untuk menangani hal ini salah satunya dengan melakukan feature extraction menggunakan PCA

EXPLORATORY DATA ANALYSIS

BINAR

Chi-square Test

Jika p-value <= 0.05, **tolak H0;** Jika tidak maka **gagal tolak H0**

Berikut merupakan tabel kontingensi

churn	area_code	area_code_408	area_code_415	area_code_510
no		934	1821	897
yes		152	287	159

Nilai P-Value = 0.5442605842955197
Gagal Tolak H0

H0 : Tidak terdapat hubungan yang signifikan antara variabel area code dengan variabel churn
H1 : Terdapat hubungan yang signifikan antara variabel area code dengan variabel churn

Karena **Gagal Tolak H0** berarti bahwa kedua variabel **tidak** memiliki hubungan yang berarti atau signifikan

H0 : Tidak terdapat hubungan yang signifikan antara variabel international plan dengan variabel churn
H1 : Terdapat hubungan yang signifikan antara variabel international plan dengan variabel churn

Karena **Tolak H0** berarti bahwa kedua variabel **memiliki** hubungan yang signifikan atau berarti

Berikut merupakan tabel kontingensi

churn	international_plan	no	yes
no		3423	229
yes		431	167

Nilai P-Value = 1.9831895448817517e-63
Tolak H0

Berikut merupakan tabel kontingensi

churn	voice_mail_plan	no	yes
no		2622	1030
yes		516	82

Nilai P-Value = 1.139803854851859e-13
Tolak H0

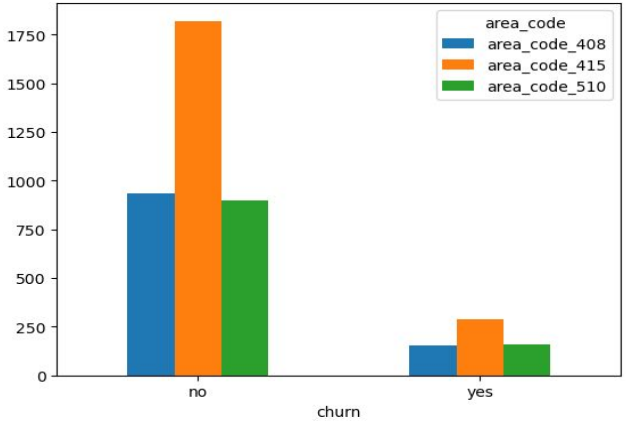
H0 : Tidak terdapat hubungan yang signifikan antara variabel voice_mail_plan dengan variabel churn
H1 : Terdapat hubungan yang signifikan antara variabel voice_mail_plan dengan variabel churn

Karena **Tolak H0** berarti bahwa kedua variabel **memiliki** hubungan yang siginifikan atau berarti

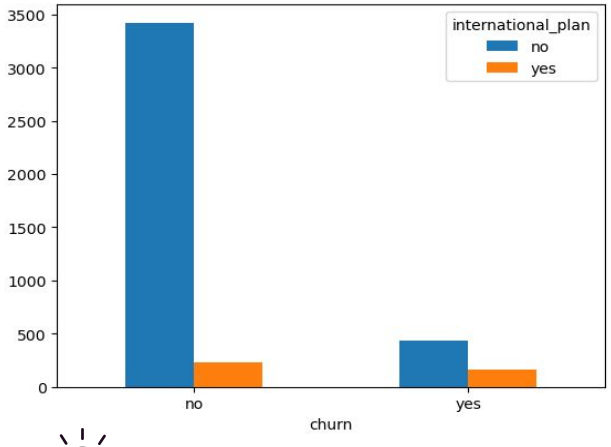
EXPLORATORY DATA ANALYSIS

BINAR

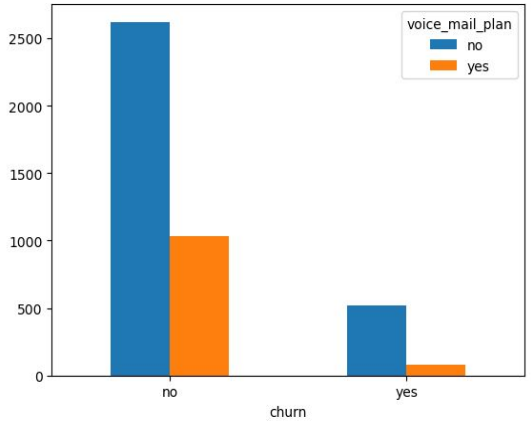
Bar chart



Customer pada area code 415 lebih banyak memutuskan untuk churn dibandingkan area code lainnya sehingga customer pada area ini perlu diperhatikan



Customer yang churn maupun tidak menunjukkan kecilnya pengguna rencana panggilan internasional yang menandakan bahwa kebanyakan customer hanya menggunakan panggilan domestik saja

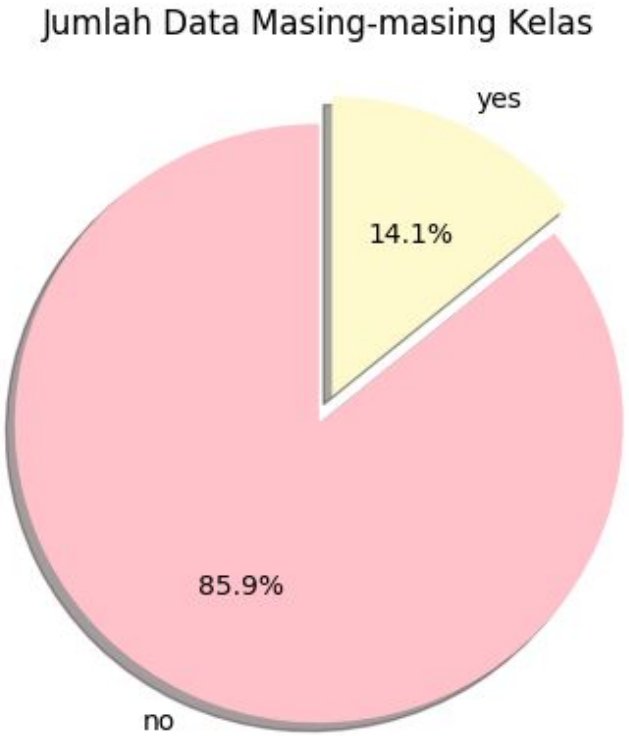


Tingkat churn lebih rendah pada customer yang memiliki rencana kotak suara sehingga perusahaan membutuhkan strategi untuk dapat menarik customer berencana menggunakan paket ini

EXPLORATORY DATA ANALYSIS

BINAR

Pie chart



Sebesar 14% customer menghentikan layanan namun angka ini masih lebih kecil dibandingkan customer yang masih berlangganan

Variabel target yaitu churn memiliki ketimpangan jumlah kelas yang cukup besar atau yang dikenal dengan **imbalance** class. Hal ini dapat menyebabkan model machine learning cenderung akan memprediksi kelas mayoritas lebih baik dibandingkan kelas minoritas.

PRE-PROCESSING DATA

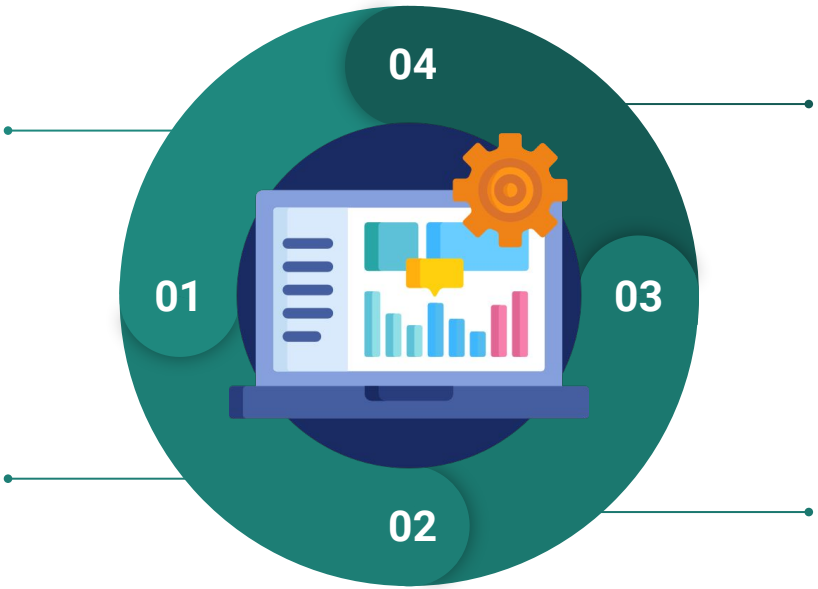
BINAR

Convert categorical into numeric one

Melakukan konversi data kategorik pada fitur area_code, international_plan, voice_mail_plan dan target yaitu churn

Drop unnecessary columns and splitting data

Fitur state dikeluarkan karena dianggap tidak memberikan informasi yang relevan terhadap pembuatan model



Feature Extraction

Ekstraksi fitur menggunakan PCA untuk mengurangi dampak dari multikolinearitas terhadap modelling dengan pemilihan komponen utama 9 yang dapat menjelaskan kumulatif varians sebesar 91%

Rescaling data

Rescaling semua fitur menggunakan normalisasi dengan rentang data berkisar 0 sampai 1

**Pre-process pada data test juga sama seperti data train

MODELLING & EVALUATION

BINAR

Data Processed

Logistic Regression

- Menggunakan parameter class_weight balanced karena kelas tidak seimbang

Without PCA

- DIlanjutkan fitting model tanpa tambahan perlakuan

With PCA

- Menggunakan pipeline untuk menyalurkan hasil pca ke untuk fitting model

Random Forest

- Menggunakan parameter class_weight balanced karena kelas tidak seimbang

Without PCA

- DIlanjutkan fitting model tanpa tambahan perlakuan

With PCA

- Menggunakan pipeline untuk menyalurkan hasil pca ke dalam model

Evaluation

- Menggunakan metode Stratified Cross-Validation karena Imbalance class

MODEL COMPARISON

BINAR

Model / Evaluation	Logistic Regression **without** PCA	Logistic Regression **with** PCA	Random Forest **without** PCA	Random Forest **with** PCA
Precision	95.29%	95.45%	95.96%	93.91%
Recall	77.66%	77.66%	99.23%	98.74%
F1 Score	85.55%	85.62%	97.56%	96.26%
ROC AUC	82.98%	82.94%	91.76%	90.65%

Precision mengukur seberapa akurat model dalam memastikan bahwa **pelanggan yang di prediksi churn adalah benar-benar churn**

Recall mengukur seberapa akurat model dalam **menemukan semua pelanggan yang churn**

F1 score merupakan keseimbangan antara precision dan recall yang digunakan ketika imbalance class

ROC AUC mengukur seberapa baik model dalam **membedakan** pelanggan yang churn dan tidak churn dan merupakan alternatif dari metrik akurasi pada klasifikasi imbalance class

PREDICT THE DATA

BINAR

Random Forest without PCA

```python
prediksi_modelRF = RF.predict(X_test_norm)
print(prediksi_modelRF)
```

```
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 0 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1
 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1
 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 0 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 0 1 1 1 1 0 1 1
 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1
 1 1 0 1 1 0 1 1 1 0 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1
 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1
 1 1 1 1 1 1 1 1 1 1]
```

```python
[ ] len(prediksi_modelRF)
```

```
750
```

```python
[ ] df_test['churn'].value_counts()
```

```
0:  130
1:  620
Name: churn, dtype: int64
```

Hasil prediksi menggunakan model terbaik menunjukkan bahwa 750 total customer 130 diantaranya “churn” dimana customer akan berhenti menggunakan layanan telecom sedangkan 620 customer “tidak churn” yakni tetap berlangganan.

CONCLUSION & RECOMMENDATION

Conclusion

- Semakin lama **durasi panggilan** semakin besar pula **biaya yang dikenakan**
- Rencana customer dalam melakukan **panggilan internasional** dan menggunakan **kotak suara** memiliki **hubungan yang signifikan** terhadap keputusan **churn**
- PCA mereduksi dimensi yang menghasilkan 9 Komponen Utama dengan jumlah kumulatif varians sebesar 91% serta **meningkatkan performa evaluasi** precision dan f1 score pada regresi logistik
- Model terbaik dengan metrik evaluasi tertinggi yaitu **Random Forest tanpa PCA** dengan f1 score sebesar 97.56% dan ROC AUC sebesar 91.76%. Random Forest merupakan algoritma yang **robust terhadap klasifikasi imbalance class sekaligus adanya multikolinearitas** dalam dataset yang dapat dilihat dari performa model yang lebih baik jika dibandingkan dengan model yang menggunakan PCA

Recommendation

- Mengenakan **charge** yang sama pada panggilan di siang, sore maupun malam hari
- **Cepat tanggap** dalam memperbaiki kualitas pelayanan serta memberikan solusi terkait keluhan/masalah dari customer
- Meningkatkan strategi untuk **menarik customer** dalam berencana menggunakan kotak suara dan panggilan internasional

BINAR

Thank You