Prediksi Customer Churn

Disebuah Internet Provider

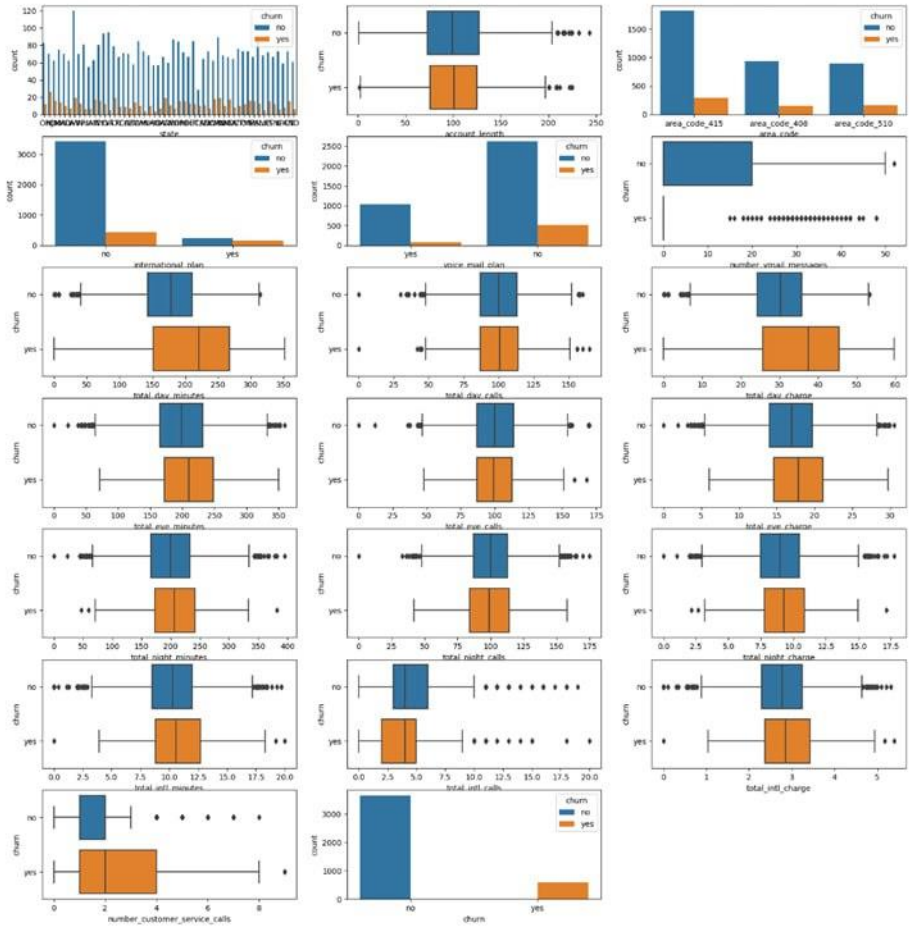
Kelompok 17 FGA DATA SCIENCE

• Tujuan Analisis

- Memahami Karakteristis Data
- Mempersiapkan Data Pemodelan
- Membangun model Decision Tree & Random Forest
- Menbandingkan kedua model dan memprediksi data Test

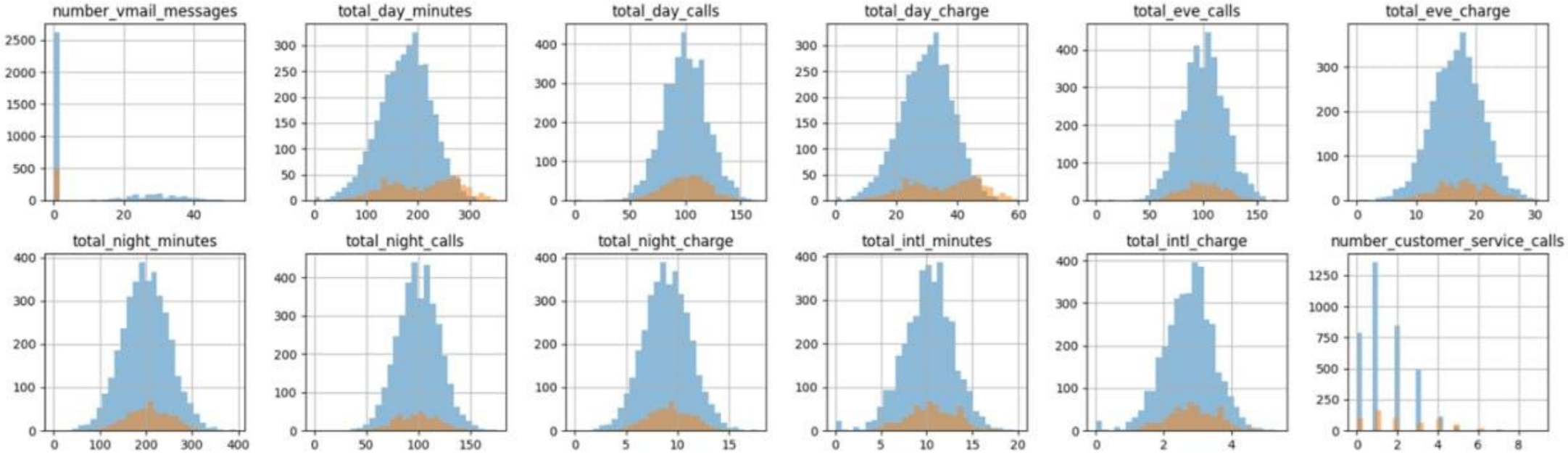
• DESKRIPSI DATA

- 20 fitur dan 4250 data
- Didata tersebut terdapat tiga type data yaitu float, integer dan string
- Dalam data tersebut terdapat 15 fitur numerik dan 5 fitur non numerik
- Pada data numerik terdapat dua data unik yaitu ordinal (3 hingga 10 nilia unik) dan kontinu (lebih dari dari 10 nilai unik
- Tidak terdapat data missing dan data duplicate

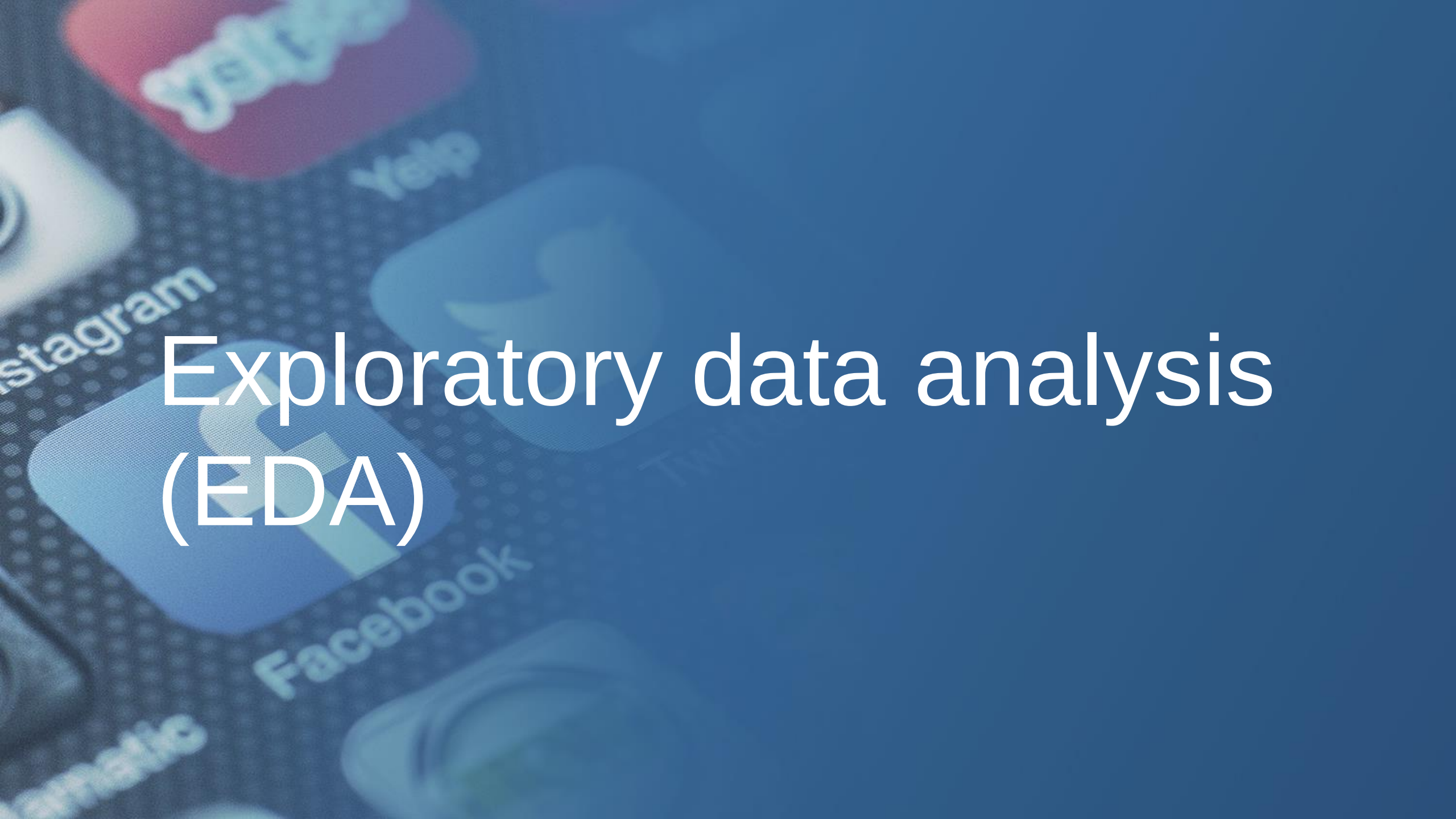


quality data :Tidak ada Duplicate
Tidak ada Missing Value
Terdapat Outliers
Imbalance

Churn : orange
No Churn :Loyal Blue



Churn : orange
No Churn :Loyal Blue



Exploratory data analysis (EDA)

Area Code Vs Churn

Area Code 408
pada area code 408 terdapat churn False (no) 934 dan True (Yes) 152. shingga total data 1086. churn 14 %

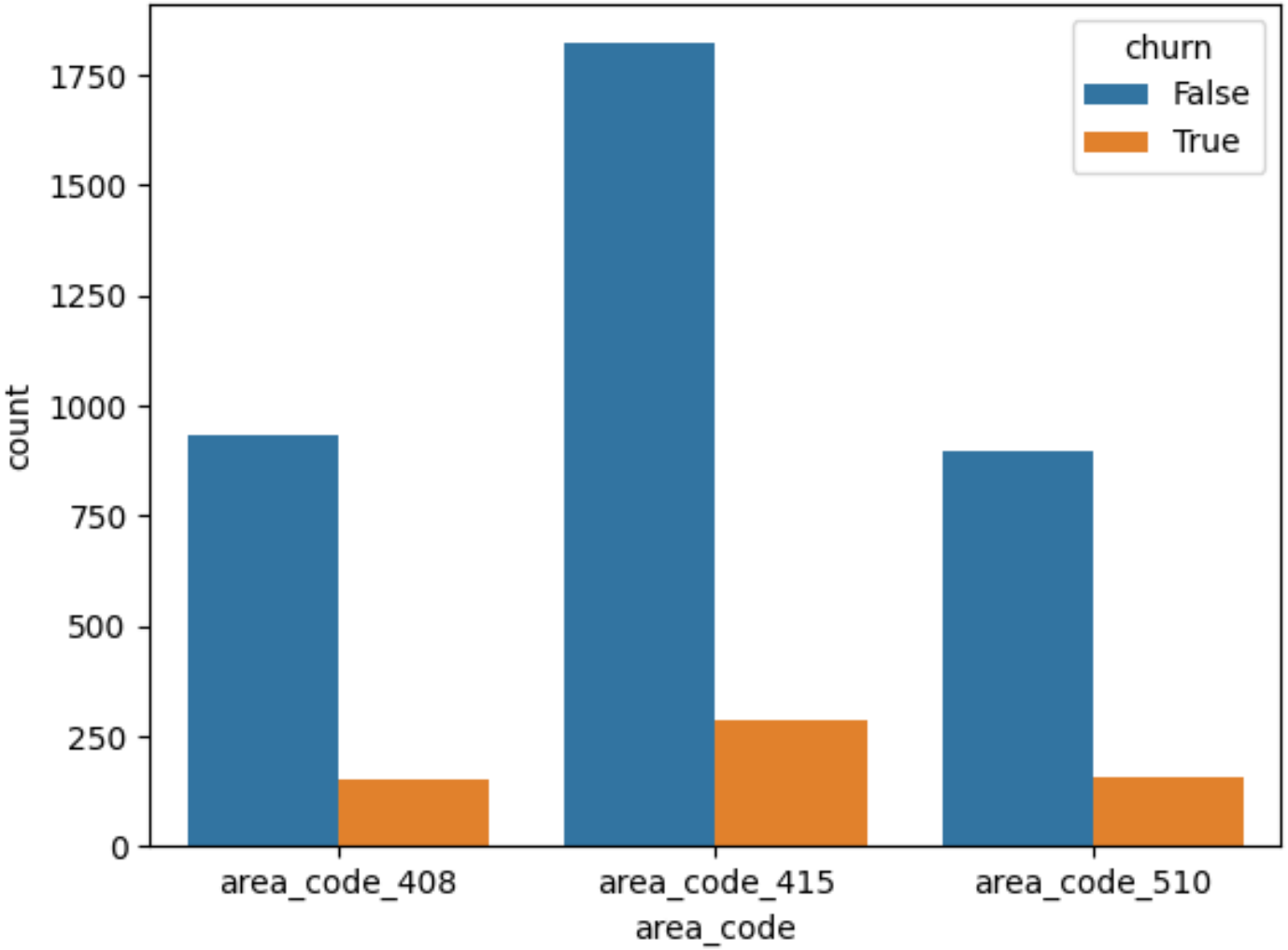
14%

Area Code 415
pada area code 408 terdapat churn False (no) 1821 dan True (Yes) 287. shingga total data 1086 atau sebanyak 13,6 %

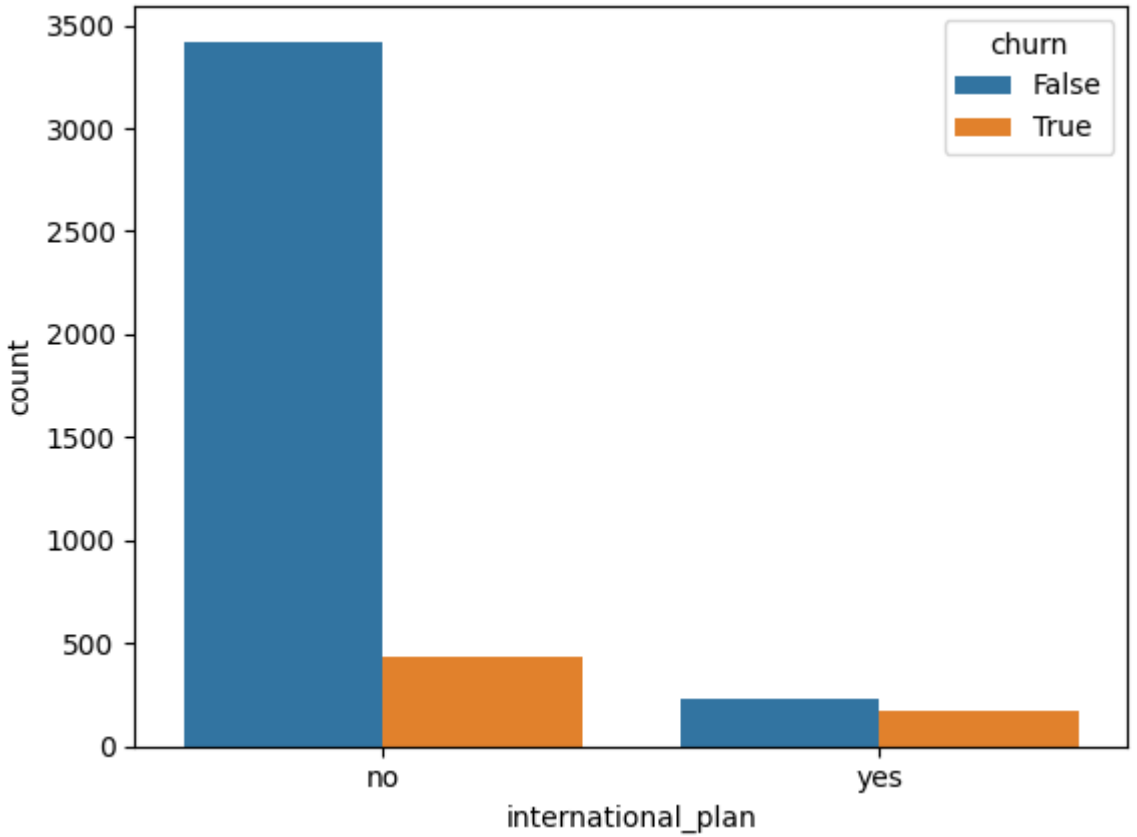
13,6%

Area Code 415
pada area code 408 terdapat churn False (no) 897 dan True (Yes) 159. shingga total data 1086 atau sebanyak 15,1%

15,1%



International Vs Churn



Tidak menggunakan International plan

Hubungan antara international plan dengan churn yang paling tinggi maka di customer yang tidak memakai international plan, dan banyak juga yang berali provider di custemor yang tidak menggunakan international plan. Banyak costumer yang berali maka 3423, sedangkan yang masih tetap sebesar 431. maka presntase churn 11 %

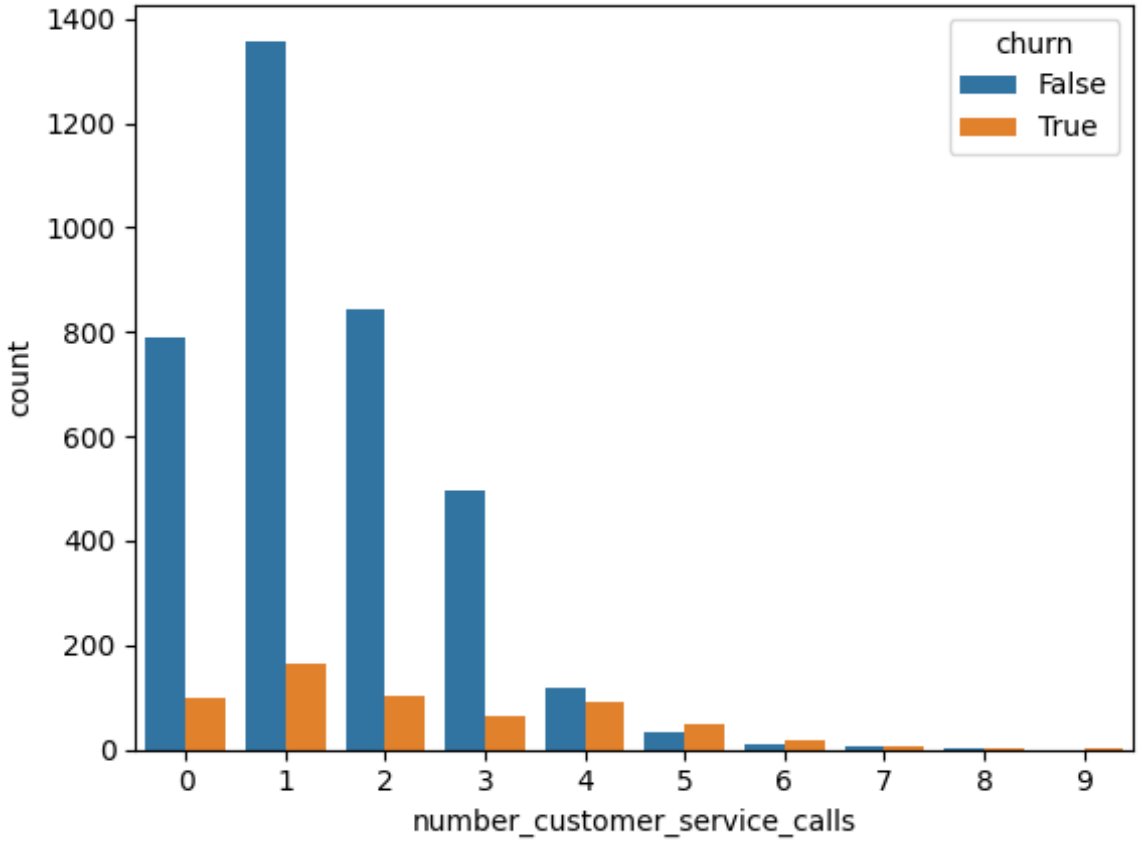
11%

Menggunakan International plan

Customer yang menggunakan international plan sangat sedikit. Customer penguna international plan yang berali provider itu sebanyak 229 sedangkan yang masih tetap sebanyak 167. maka presentase churn 42 %

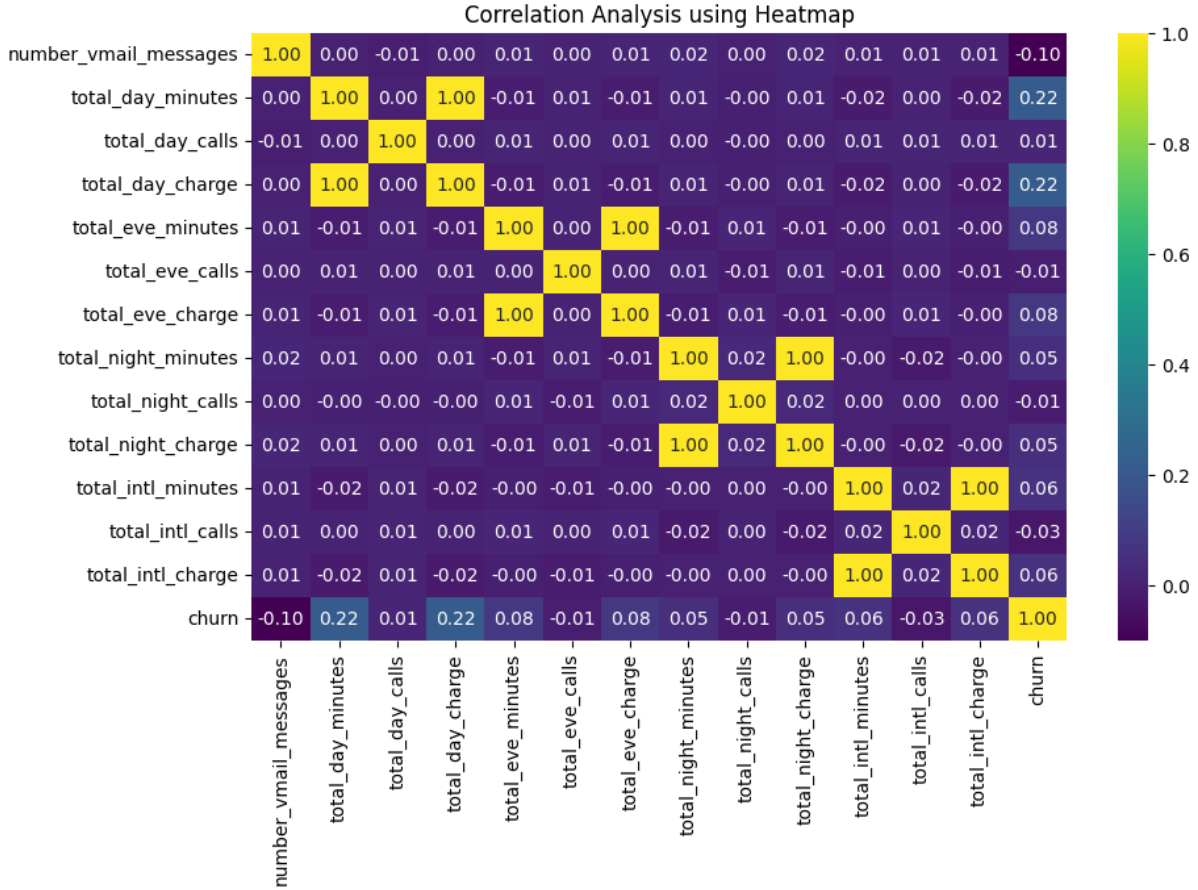
42%

Number Customer Servis calls Vs Churn



	number_customer_service_calls	churn_no	churn_yes	total	churn Percentage
0	0	789	97	886	10.948081
1	1	1358	166	1524	10.892388
2	2	845	102	947	10.770855
3	3	495	63	558	11.290323
4	4	117	92	209	44.019139
5	5	32	49	81	60.493827
6	6	9	19	28	67.857143
7	7	6	7	13	53.846154
8	8	1	1	2	50.000000
9	9	0	2	2	100.000000
10	All	3652	598	4250	14.070588

Feature Correlations



total_night_charge dan total_night_minutes

total_day_charge dan total_day_minutes

total_intl_charge dan total_intl_minutes

Ini sudah diduga karena orang harus menagih lebih banyak jika mereka menggunakan lebih banyak menit.

PREPROCESSING

Preprocessing	Input	Output
Handle Outliers	4250	3899
Feature Selection	Data_train	remove state
Convert To Numeric	String	Numeric
Handle Imbalance	Data Imbalance 3899	Balance 50:50 6734 baris, 24 kolom
Feature Encoding	Data balance Cateorical	Data dengan skala yang sama (numeric)

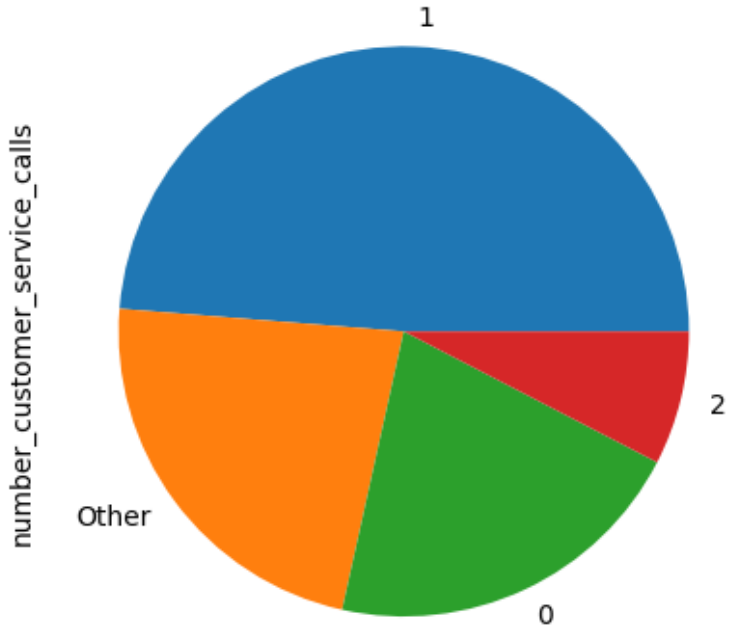
```
le = LabelEncoder()

data_train['international_plan'] = le.fit_transform(data_train['international_plan'])
data_train['voice_mail_plan'] = le.fit_transform(data_train['voice_mail_plan'])
data_train['churn'] = le.fit_transform(data_train['churn'])
```

- Columns international plan, voice mail plan, churn diganti ke angka supaya bisa di analisis yaitu yes =1 dan No = 0

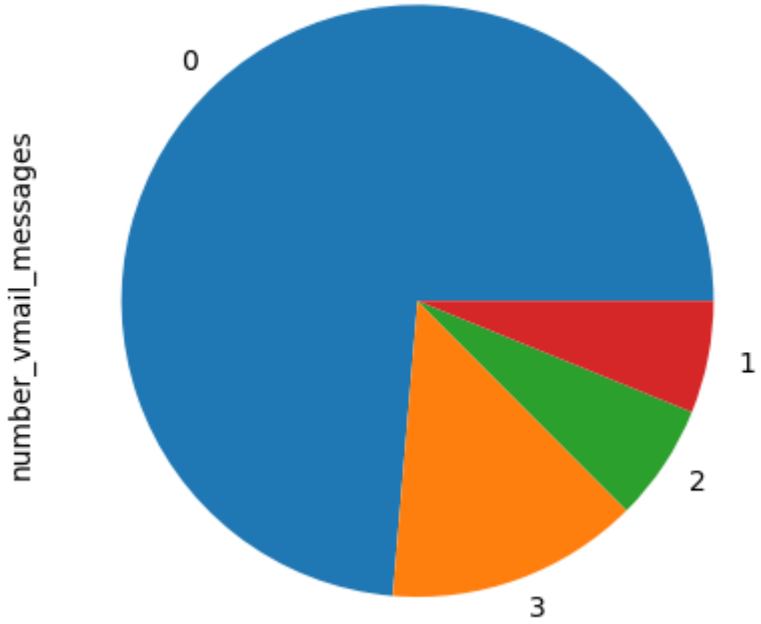
```
def groupcat(x):
    if x == 0:
        return '0'
    elif x >= 1 | x <= 3:
        return '1'
    elif x >= 4 | x <= 6:
        return '2'
    else:
        return 'Other'
```

- Di columns nomor customer service(NCS) diganti berdasarkakn logic di atas yaitu apabila NCS = 0 maka 0 NCS>=1 & NCS<=3 maka 1, apabila NCS>=4 & NCS <=6 maka 2, selain itu other



```
def categorizing(x):
    if x == 0:
        return 0
    elif x < 24:
        return 1
    elif x < 29:
        return 2
    else:
        return 3
```

- Di columns nomor customer service(NCS) diganti berdasarkakn logic di atas yaitu apabila NCS = 0 maka 0 NCS<24 maka 1, apabila NCS<29 maka 2, NCS >=29 maka 3



MODELING

Decision Tree

Decision tree, yang diterjemahkan menjadi "pohon keputusan" dalam bahasa Indonesia, adalah sebuah metode yang digunakan untuk mempermudah pengambilan keputusan.

Random forest

Random forest, yang secara harfiah diterjemahkan menjadi "hutan acak", adalah algoritma machine learning yang digunakan untuk **klasifikasi** dan **regresi** data. Ia bekerja dengan menggabungkan prediksi dari banyak **decision tree** (pohon keputusan) untuk menghasilkan prediksi akhir yang lebih akurat dan stabil.

Hasil Modeling

	Algorithm	ROC AUC Mean	ROC AUC STD	Accuracy Mean	Accuracy STD
1	Random Forest	98.89	0.26	95.14	0.82
0	Decision Tree Classifier	91.96	0.61	91.96	0.62

Dari hasil perbandingan didapatkan performa model terbaik yaitu Random forest (berdasarkan ROC dan accuracy).

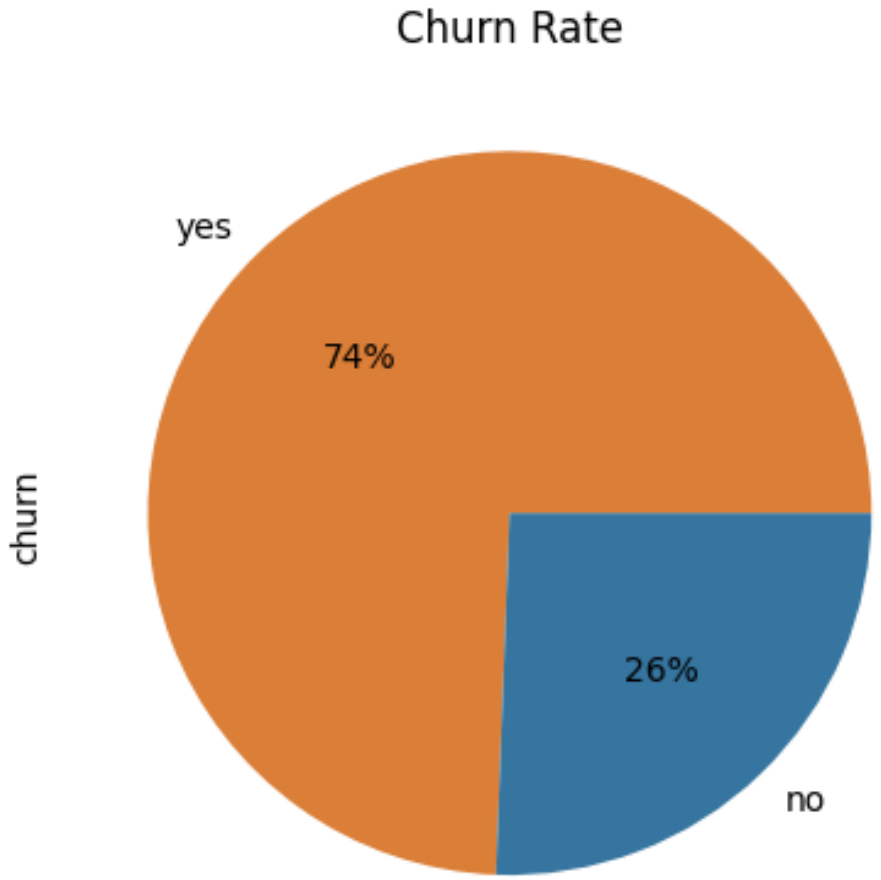
PREDICTION

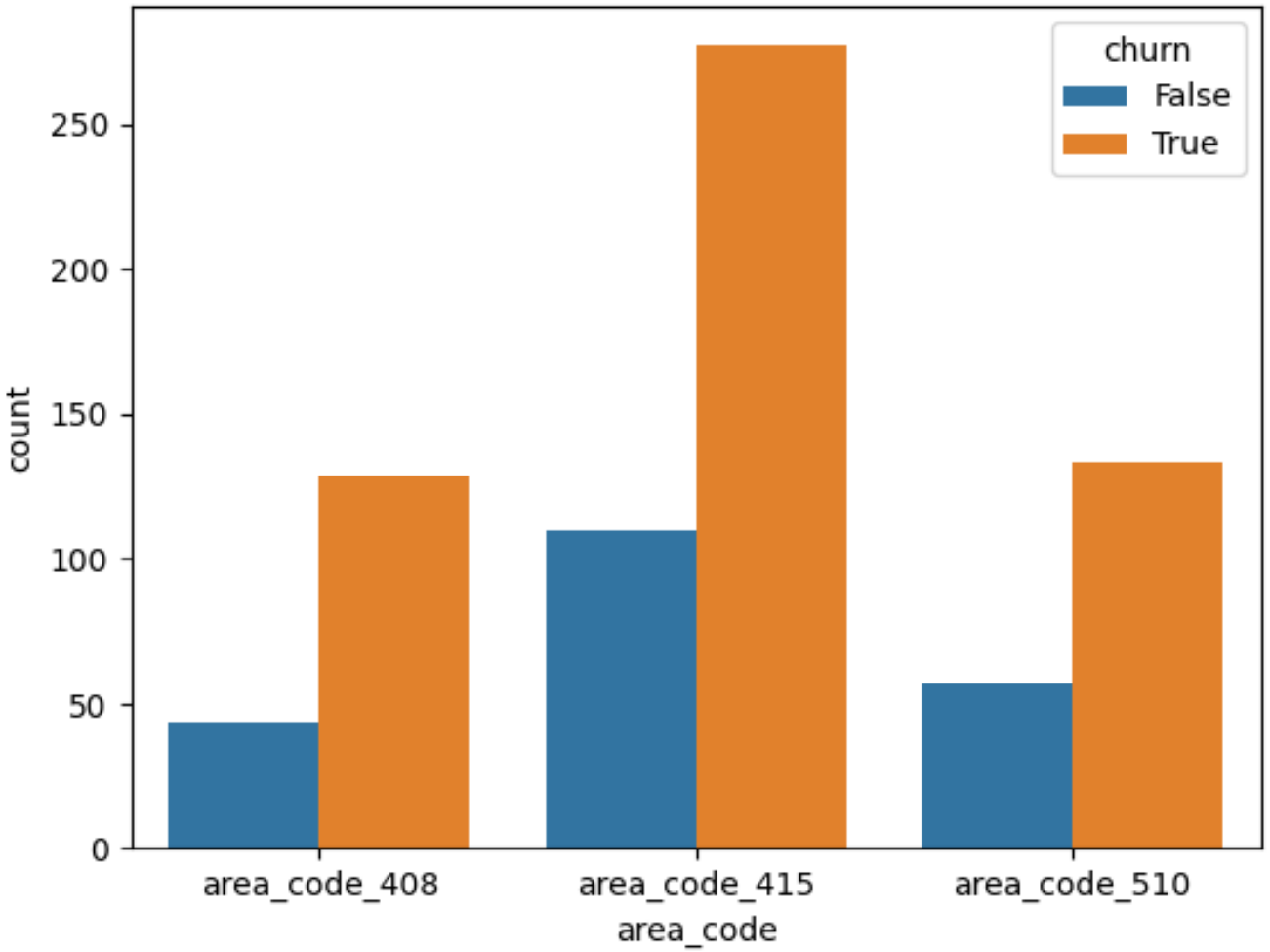
Tabel Feature Important

	importance
total_day_charge	0.112514
total_day_minutes	0.111636

Berdasarkan nilai importance feature total day charge dan minute adalah yang tertinggi, dan sesuai hasil eda keduanya

Churn Rate mencapai 72% dari 750 pelanggan. sehigga sebelum terjadi churn disarankan untuk meningkatan layanan, atau penawaran terbaru.




churn
False
True
250
200
150
100
50
0
count
area_code_408
area_code_415
area_code_510
area_code

Kesimpulan & Saran

- Berdasarkan nilai importance feature total day charge dan minute adalah yang tertinggi, dan sesuai hasil eda keduanya
- Churn Rate mencapai 72% dari 750 pelanggan. sehigga sebelum terjadi churn disarankan untuk meningkatan layanan, atau penawaran terbaru.

Thank You