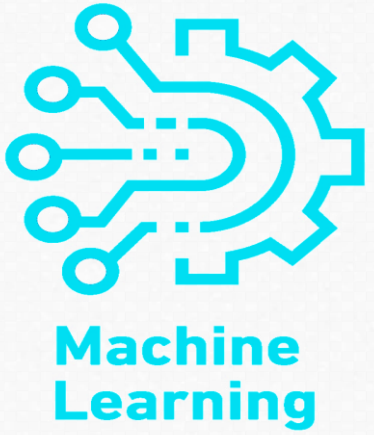
Challenge chapter 2
Model Machine Learning Supervised Learning
Prediksi Customer Chrun Pelanggan Internet Provider Telekomunikasi

Revo Faris Saifuddin

Latar Belakang

- Perkembangan industri telekomunikasi sangatlah cepat, hal ini dapat dilihat dari perilaku masyarakat yang menggunakan internet dalam berkomunikasi. Perilaku ini menyebabkan banyaknya perusahaan telekomunikasi dan meningkatnya internet service provider yang dapat menimbulkan persaingan antar provider. Pelanggan memiliki hak dalam memilih provider yang sesuai dan dapat beralih dari provider sebelumnya yang diartikan sebagai Customer Churn. Peralihan ini dapat menyebabkan berkurangnya pendapatan bagi perusahaan telekomunikasi sehingga penting untuk ditangani.



Permasalahan

Buat model machine learning dengan algoritma klasifikasi (supervised learning) menggunakan data train.csv 2. Lakukan prediksi customer yang churn dari hasil model (poin 1) menggunakan data test.csv 3. Kumpulkan code dalam bentuk file .ipynb/Google Colab dan hasil interpretasi dalam bentuk ppt / pdf mengenai step by step penyelesaian dan hasil yang sudah dilakukan menggunakan form submission yang disediakan oleh Tim Binar

Analisis Data

Memahami data

Import Data, Read Data, Info Data

Eksplor

Cleansing Data, Eksplorasi Data

Analisis Data

Scalling Data, Clustering, Hasil

Tujuan Analisis

- Membantu rekomendasi mengambil keputusan untuk memilih negara yang paling membutuhkan bantuan.
- Mengkategorikan negara menggunakan beberapa faktor sosial ekonomi dan kesehatan yang menentukan perkembangan negara secara keseluruhan.

Memahami Data

state	account_length	area_code	international_plan	voice_mail_plan	number_vmail_messages	total_day_minutes	total_day_calls	total_day_charge	total_eve_minutes	total_eve_calls	total_eve_charge	total_night_minutes
OH	107	area_code_415	no	yes	26	161.6	123	27.47	195.5	103	16.62	254.4
NJ	137	area_code_415	no	no	0	243.4	114	41.38	121.2	110	10.30	162.6
OH	84	area_code_408	yes	no	0	299.4	71	50.90	61.9	88	5.26	196.9
OK	75	area_code_415	yes	no	0	166.7	113	28.34	148.3	122	12.61	186.9
MA	121	area_code_510	no	yes	24	218.2	88	37.09	348.5	108	29.62	212.6
...	...	...	...	...	...	...	...	...	...	...	...	...
MT	83	area_code_415	no	no	0	188.3	70	32.01	243.8	88	20.72	213.7
WV	73	area_code_408	no	no	0	177.9	89	30.24	131.2	82	11.15	186.2
NC	75	area_code_408	no	no	0	170.7	101	29.02	193.1	126	16.41	129.1
HI	50	area_code_408	no	yes	40	235.7	127	40.07	223.0	126	18.96	297.5
VT	86	area_code_415	no	yes	34	129.4	102	22.00	267.1	104	22.70	154.8

Keterangan Data:

Data set yang digunakan memiliki 20 kolom data dan berisi 4250 baris data.

Memahami Data

Informasi Type Data

```
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
dtypes: float64(8), int64(7), object(5)
memory usage: 664.2+ KB
```

Cek Missing Value Data

```
Berdasarkan data set yang dibaca tidak ada missing value dari data set tersebut
state                           0
account_length                  0
area_code                       0
international_plan              0
voice_mail_plan                 0
number_vmail_messages           0
total_day_minutes               0
total_day_calls                 0
total_day_charge                0
total_eve_minutes               0
total_eve_calls                 0
total_eve_charge                0
total_night_minutes             0
total_night_calls               0
total_night_charge              0
total_intl_minutes              0
total_intl_calls                0
total_intl_charge               0
number_customer_service_calls   0
churn                           0
dtype: int64
```

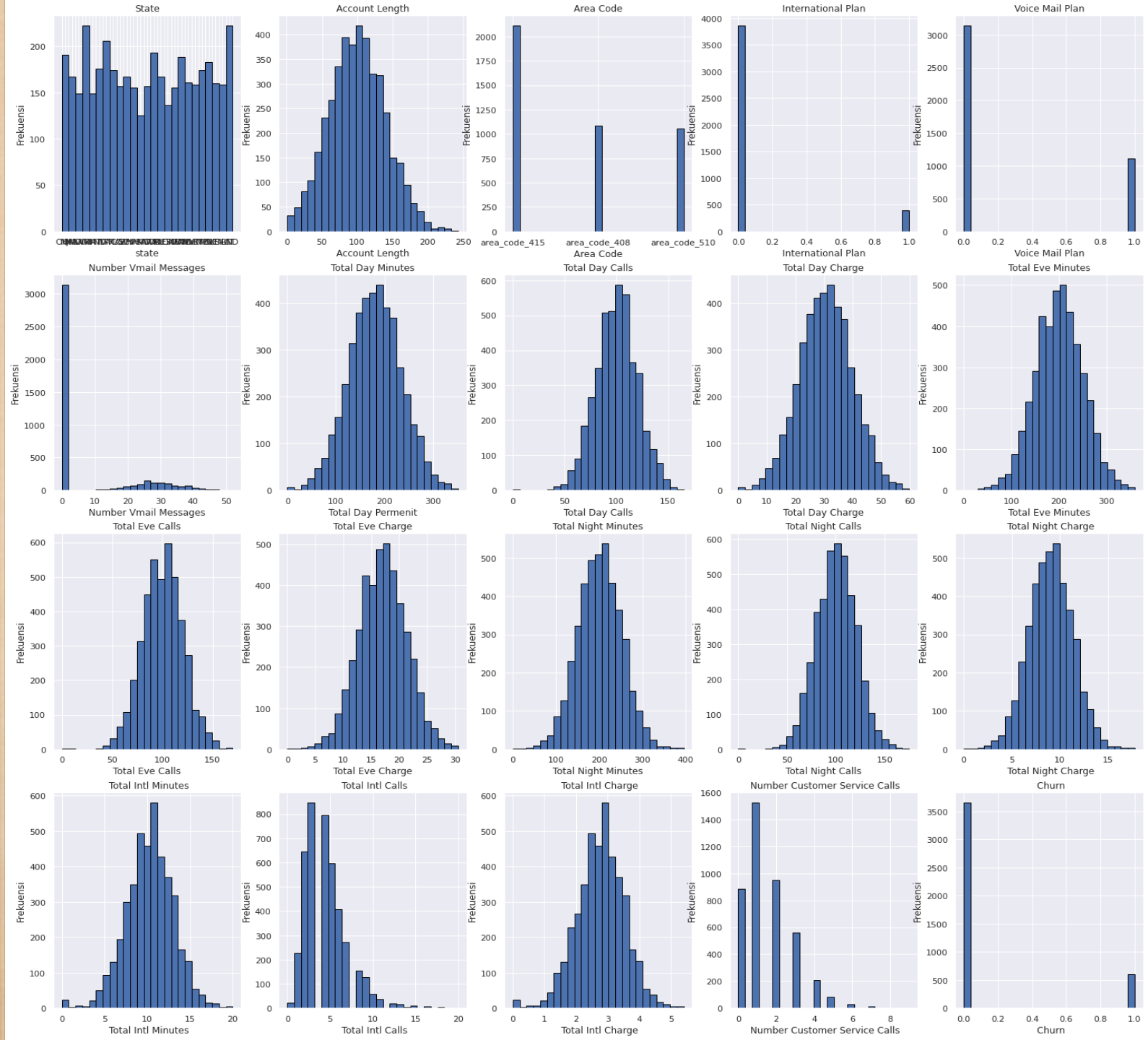
Statistika data mentah

	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

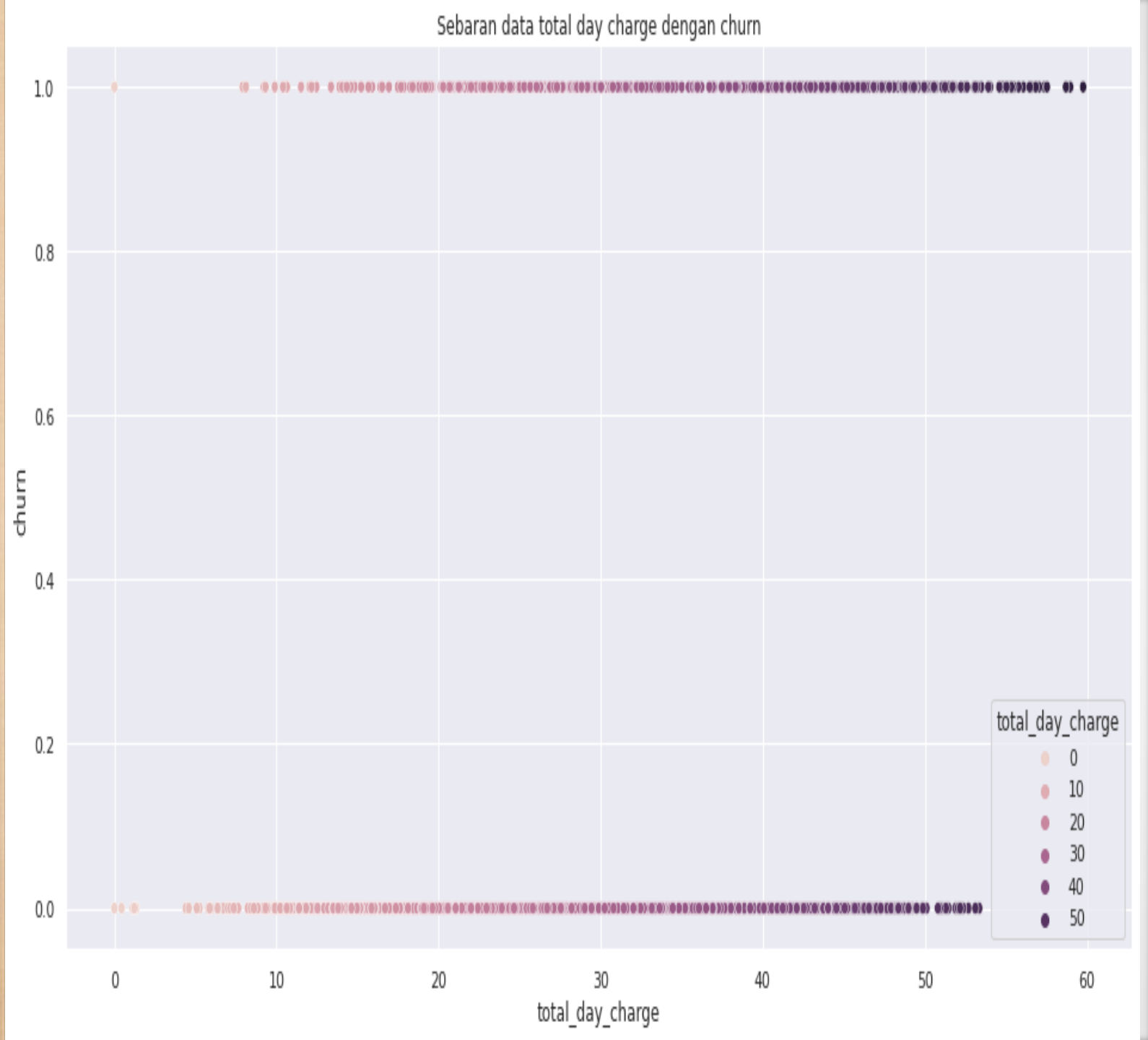
Univariate, Bivariate, and Multivariate Analysis

Univariate

Dari hasil visualisasi dengan barchart atau histogram kita mendapatkan informasi mengenai informasi mengenai jumlah frekuensi dari setiap kolom data pada dataset dengan bentuk visual

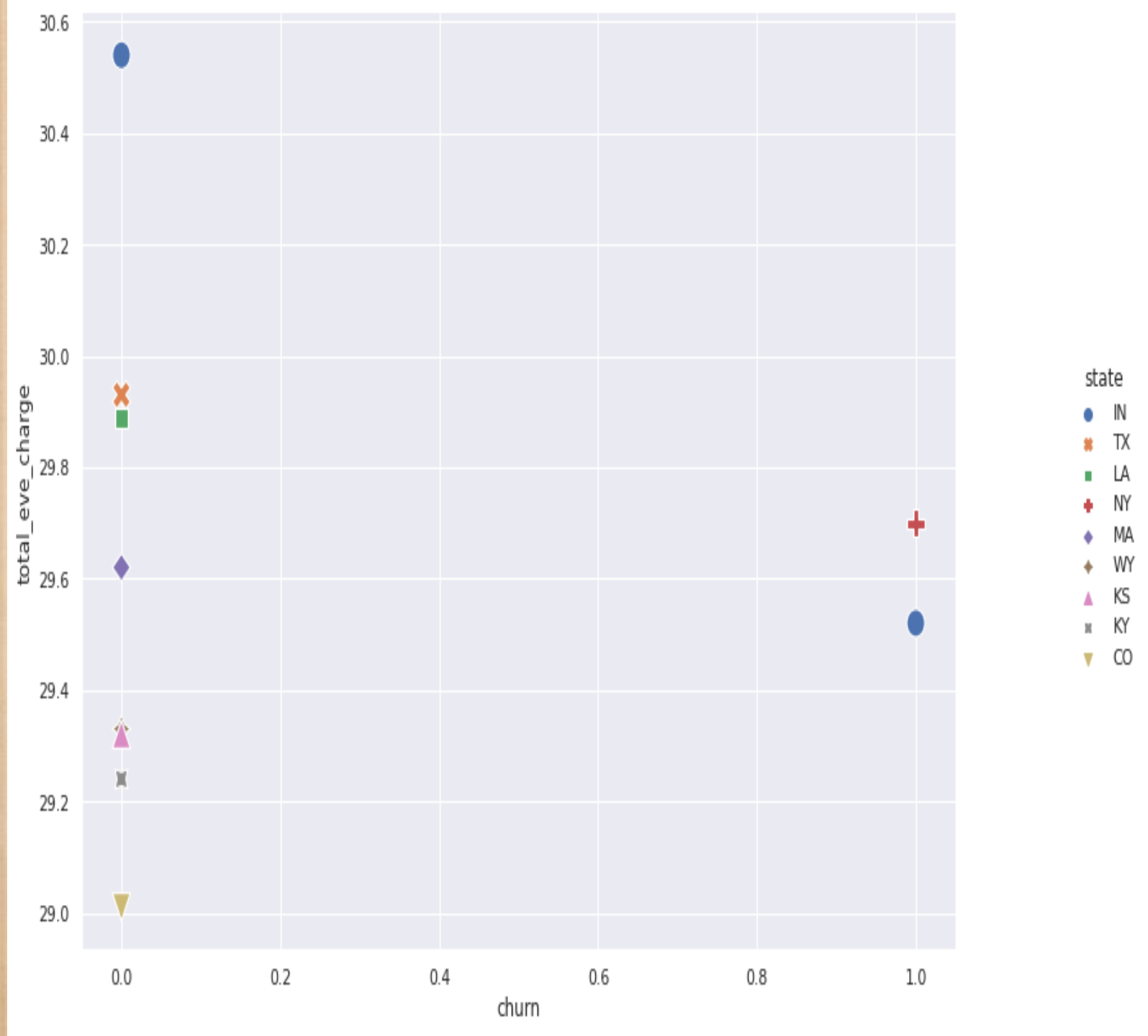


Bivariate


Sebaran data total day charge dengan churn
churn
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
total_day_charge
total_day_charge
0
10
20
30
40
50

Multivariate

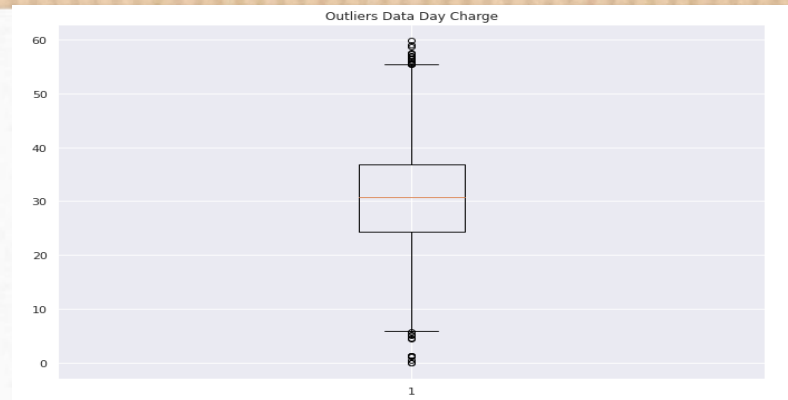
Dari hasil visualisasi menghasilkan informasi state yang memiliki hubungan antara eve charge dan churn

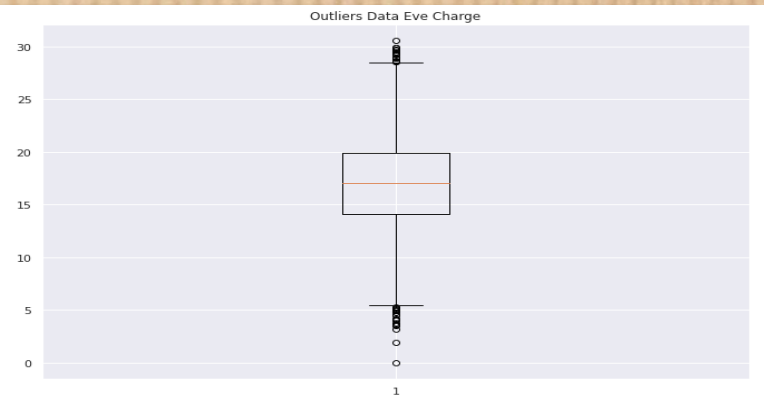


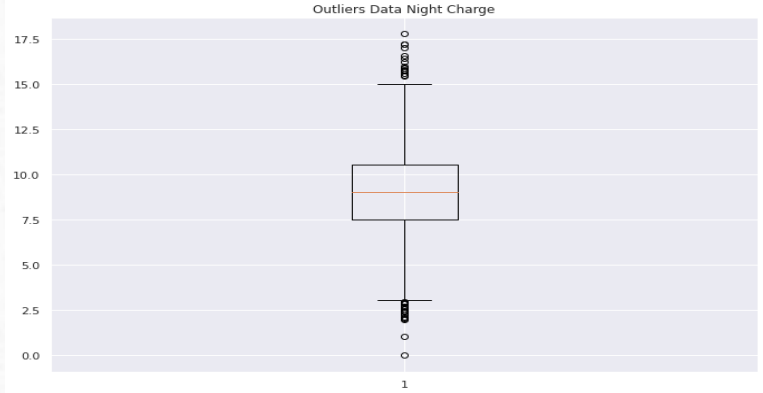
Analisis Data

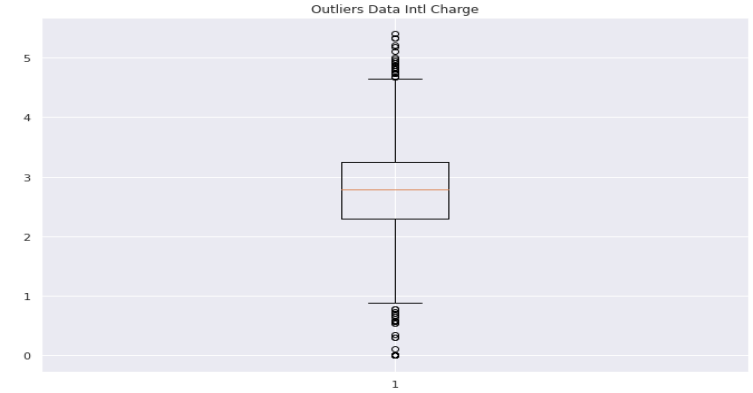
Handling outliers

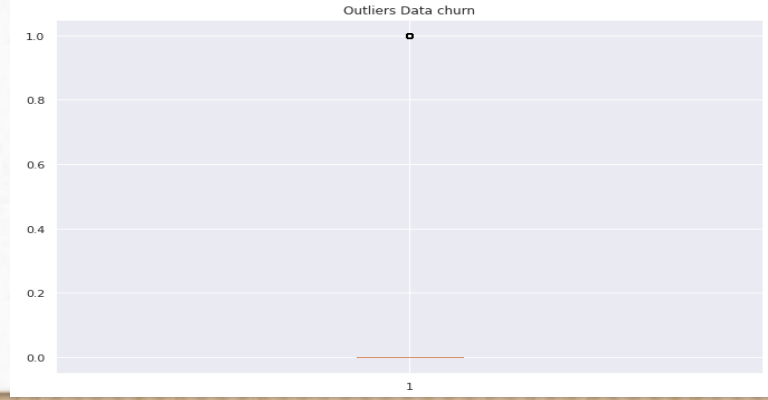
Terdapat data outliners dari semua kolom data dengan nilai jauh di atas range nilai normal

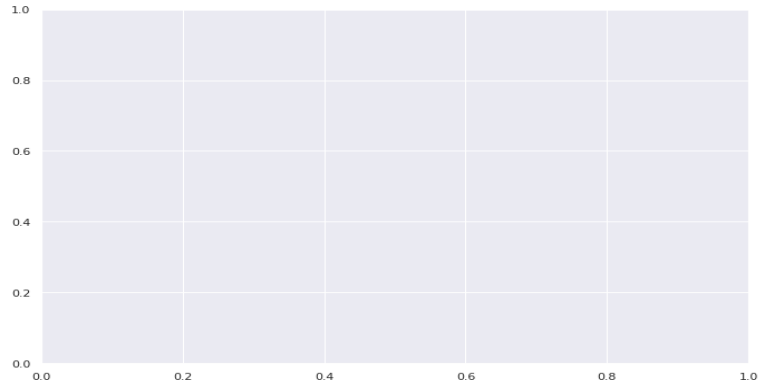








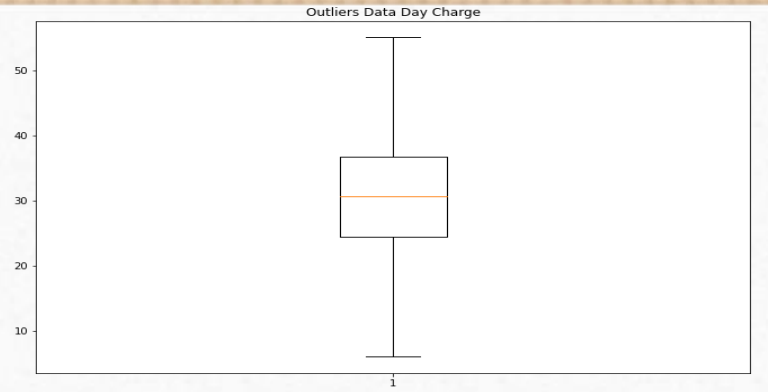


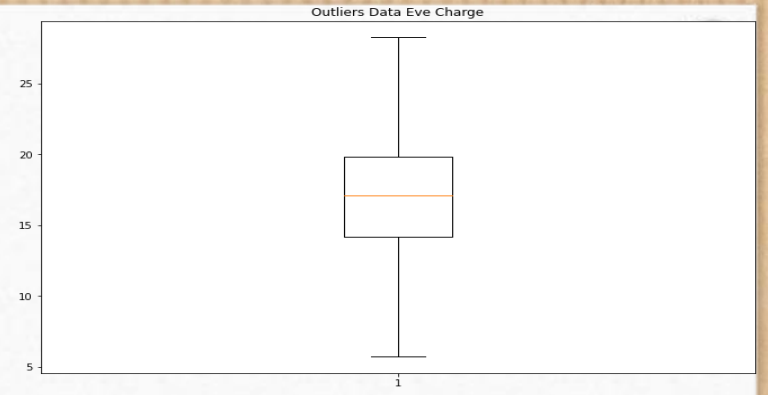


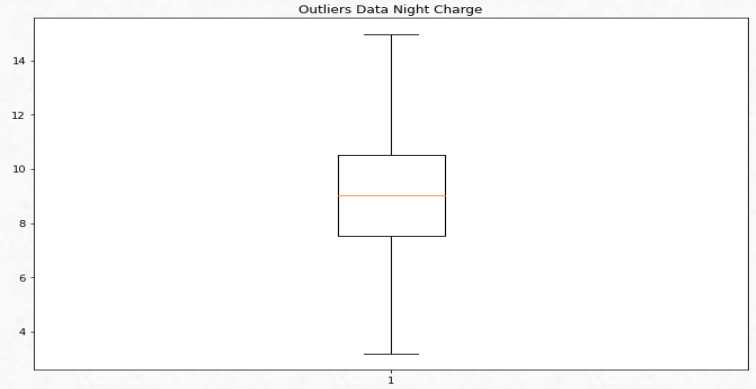
Analisis Data

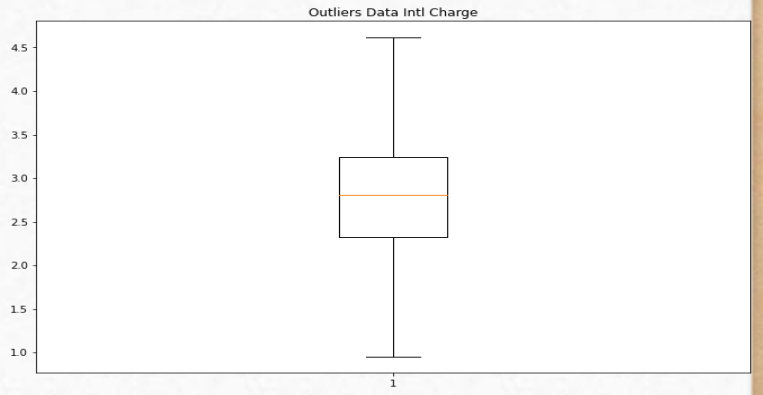
Remove outliers

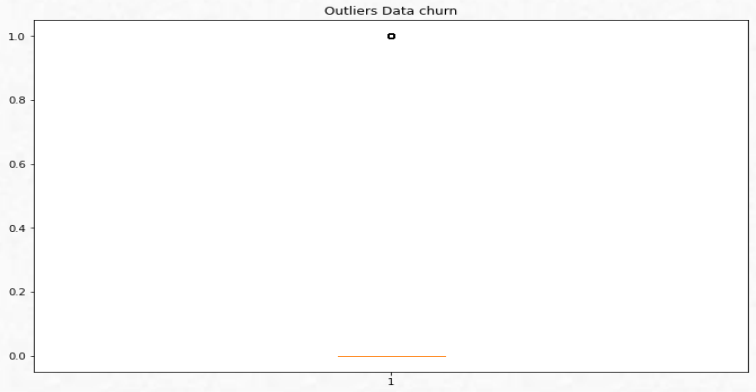
Data yang sudah di proses dengan menggunakan batas atas dan bawah pada nilai quartil

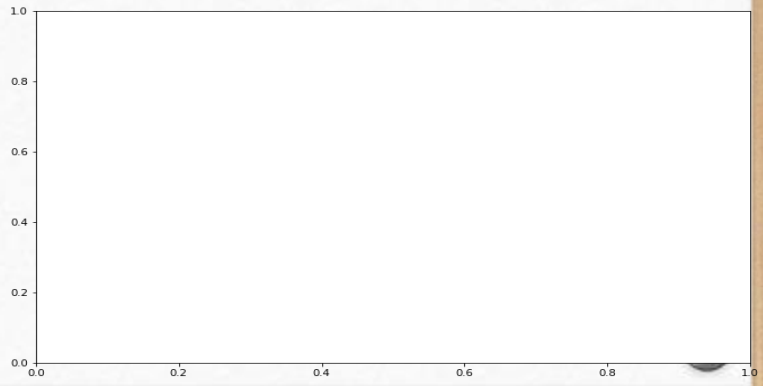






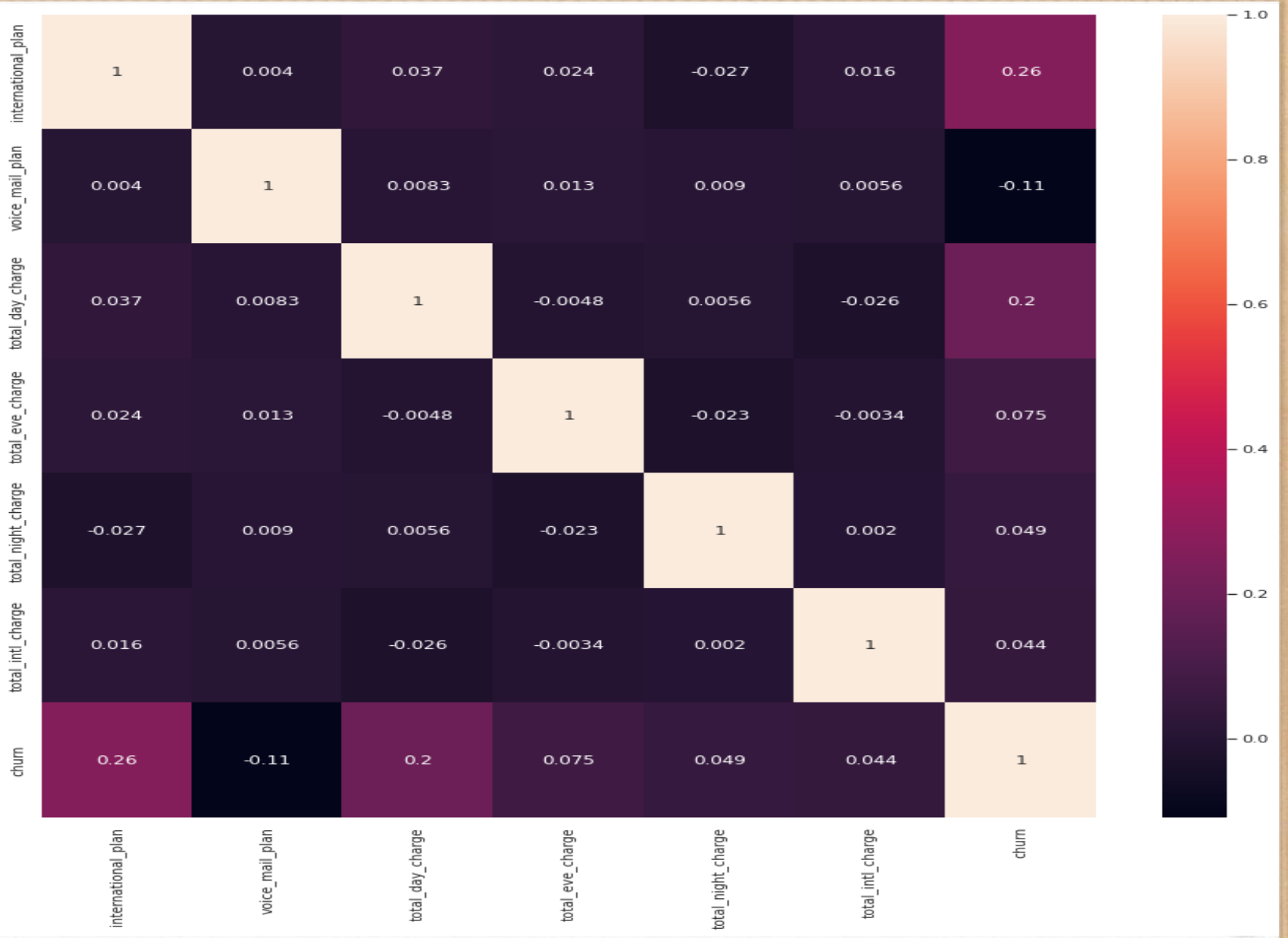






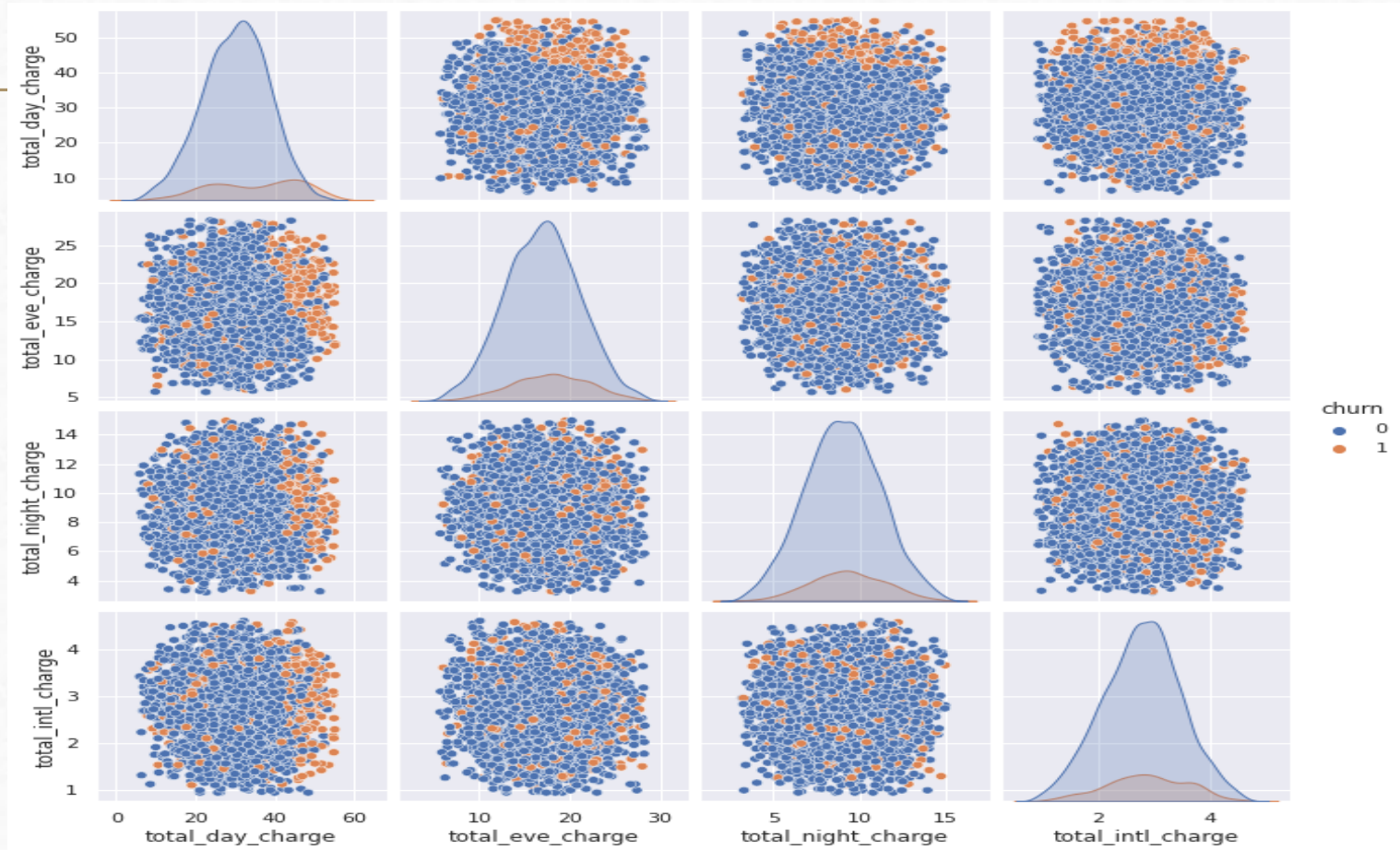
Korelasi Data

Dari hasil visualisasi menghasilkan informasi korelasi data frame yang akan digunakan modeling machine learning

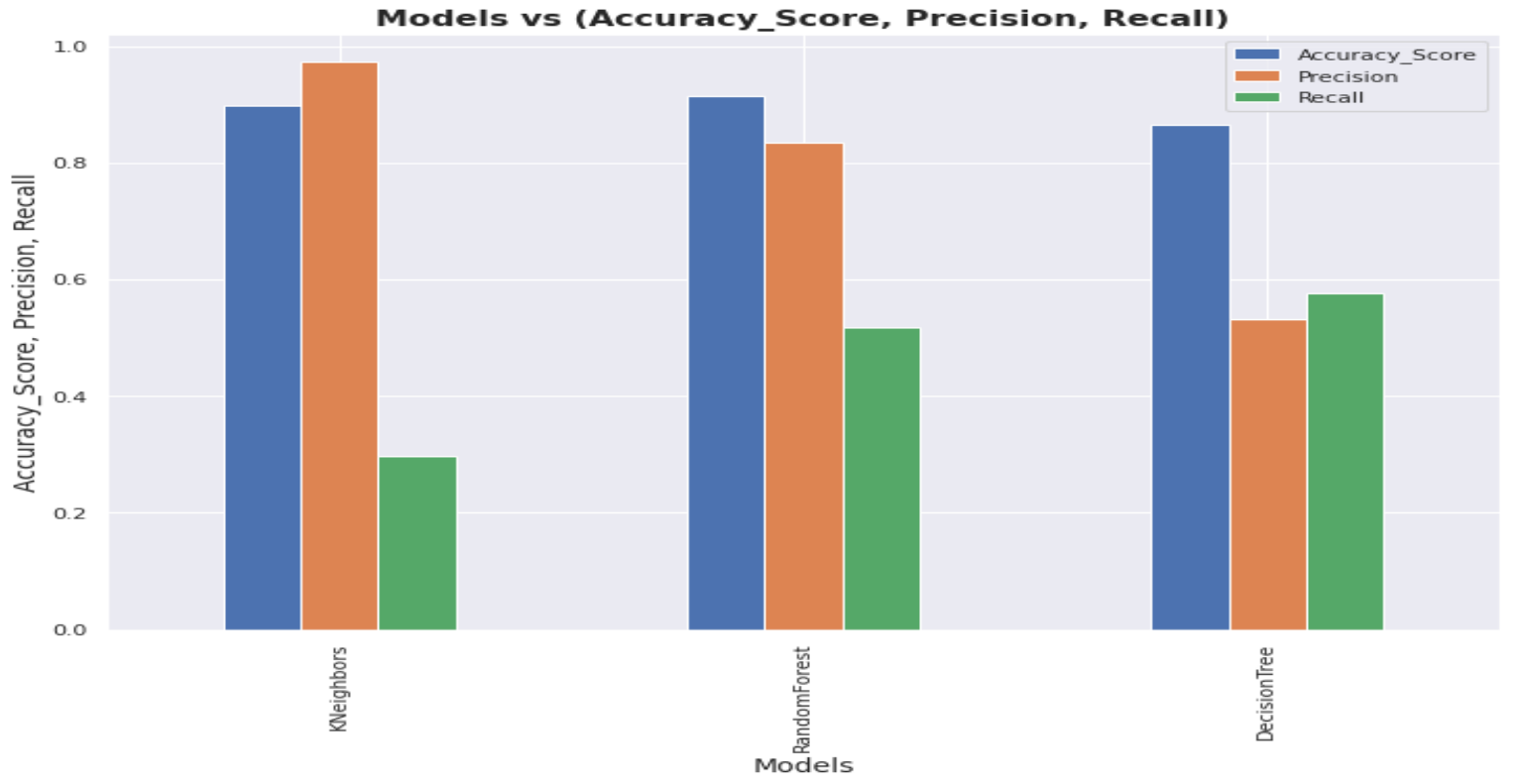


Persebaran Data

Visualisasi persebaran data terhadap customer churn



Model Machine Learning



	Accuracy_Score	Precision	Recall
KNeighbors	0.896933	0.972222	0.296610
RandomForest	0.915337	0.835616	0.516949
DecisionTree	0.865031	0.531250	0.576271

Kesimpulan

- Data yang digunakan telah dilakukan preprocessing data yang akan digunakan untuk analisis
- Model Machine learning yang akan digunakan Adalah random forest pada model klasifikasi, karena penggunaan model tersebut pada data frame yang digunkan menghasilkan nilai akurasi, presisi, dan recall lebih tinggi dari DecisionTree dan KNN yang digunakan dalam analisa ini


CONTACT US
someone@example.com