



FINAL PROJECT

INSIGHT SEEKER TEAM

BANK CUSTOMER CHURN PREDICTION

OUR TEAM



SONI AGUNG W



KATON CAHYO A



DINDA THALIA F

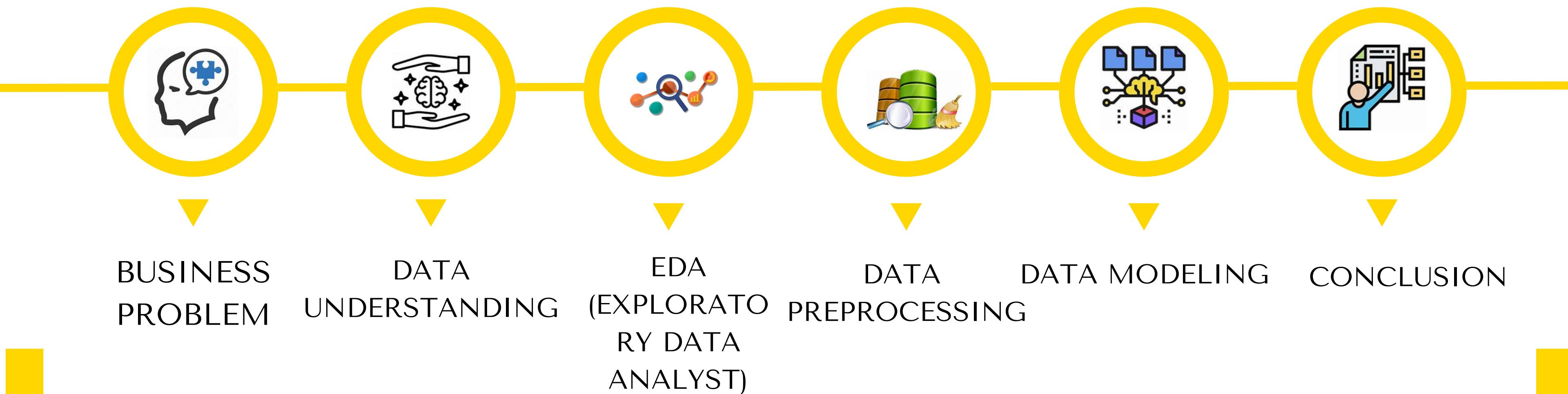


SUSI



AL-AZHARY P.S

TABLE OF CONTENT



BUSINESS PROBLEM



BUSINESS PROBLEM

- Industri perbankan saat ini menghadapi tantangan yang signifikan dalam mempertahankan pelanggan dan menghadapi persaingan yang ketat. Keberhasilan dalam memahami perilaku pelanggan dan memprediksi pelanggan yang berpotensi berhenti berlangganan layanan, yang dikenal sebagai "churn," sangat penting
- Projek Bank Customer Churn Prediction dari tim Insight Seeker bertujuan untuk mengidentifikasi faktor-faktor yang memengaruhi keputusan pelanggan untuk tetap menggunakan layanan bank atau beralih ke penyedia lain. Indikator utama keberhasilan proyek ini adalah nilai "churn."
- Proyek ini bertujuan memberikan pemahaman mendalam tentang perilaku pelanggan dalam industri perbankan dan faktor-faktor yang memengaruhi keberhasilan langganan layanan bank mereka. Hasil analisis akan membantu bank meningkatkan strategi retensi pelanggan dan pengalaman pelanggan secara keseluruhan.

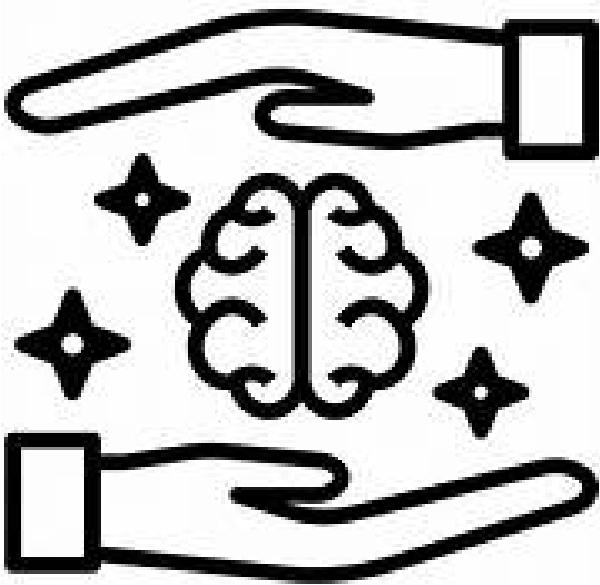


PURPOSE

Proyek ini bertujuan untuk memberikan wawasan komprehensif tentang perilaku pelanggan dalam industri perbankan dan bagaimana faktor-faktor tertentu dapat berkontribusi pada keberhasilan mereka dalam berlangganan layanan bank. Hasil analisis ini diharapkan akan membantu bank mengembangkan strategi retensi pelanggan yang lebih efektif dan meningkatkan pengalaman pelanggan secara keseluruhan.



DATA UNDERSTANDING



DATA UNDERSTANDING

data ini didapatkan dari database churn customer yang diambil dari website Kaggle.com berjumlah 10000 baris dan 12 kolom dengan variable seperti berikut:

customer_id	10000	[15634602, 15647311, 15619304, ..., 15584532, 15682355]
credit_score	460	Skor kredit pelanggan.
country	3	['France', 'Spain', 'Germany']
gender	2	['Female', 'Male']
age	70	['31-40', '21-30', '41-50', '51-60', '61+', '18-20']
tenure	11	[2, 1, 8, 7, 4, 6, 3, 10, 5, 9, 0]
balance	6382	Saldo uang di akun pelanggan.
products_number	4	[1, 3, 2, 4]
credit_card	2	[1, 0] // 1 = yes, 0=no
active_member	2	[1, 0] // 1 = yes, 0=no
estimated_salary	9999	Estimasi gaji atau pendapatan pelanggan.
churn	2	Indikator churn (1 = Ya, 0 = Tidak).

DATASET INFORMATION

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 13 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   customer_id      10000 non-null   int64  
 1   credit_score     10000 non-null   int64  
 2   country          10000 non-null   category
 3   gender           10000 non-null   category
 4   age              10000 non-null   int64  
 5   tenure           10000 non-null   int64  
 6   balance          10000 non-null   float64
 7   products_number  10000 non-null   int64  
 8   credit_card      10000 non-null   int64  
 9   active_member    10000 non-null   int64  
 10  estimated_salary 10000 non-null   float64
 11  churn            10000 non-null   int64  
 12  Age Group        9867 non-null   category
dtypes: category(3), float64(2), int64(8)
memory usage: 811.1 KB
```

menampilkan informasi detail tentang dataframe, seperti jumlah baris data, nama-nama kolom beserta jumlah data dan type datanya.

DATASET DESCRIBE

	customer_id	credit_score	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	churn
count	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

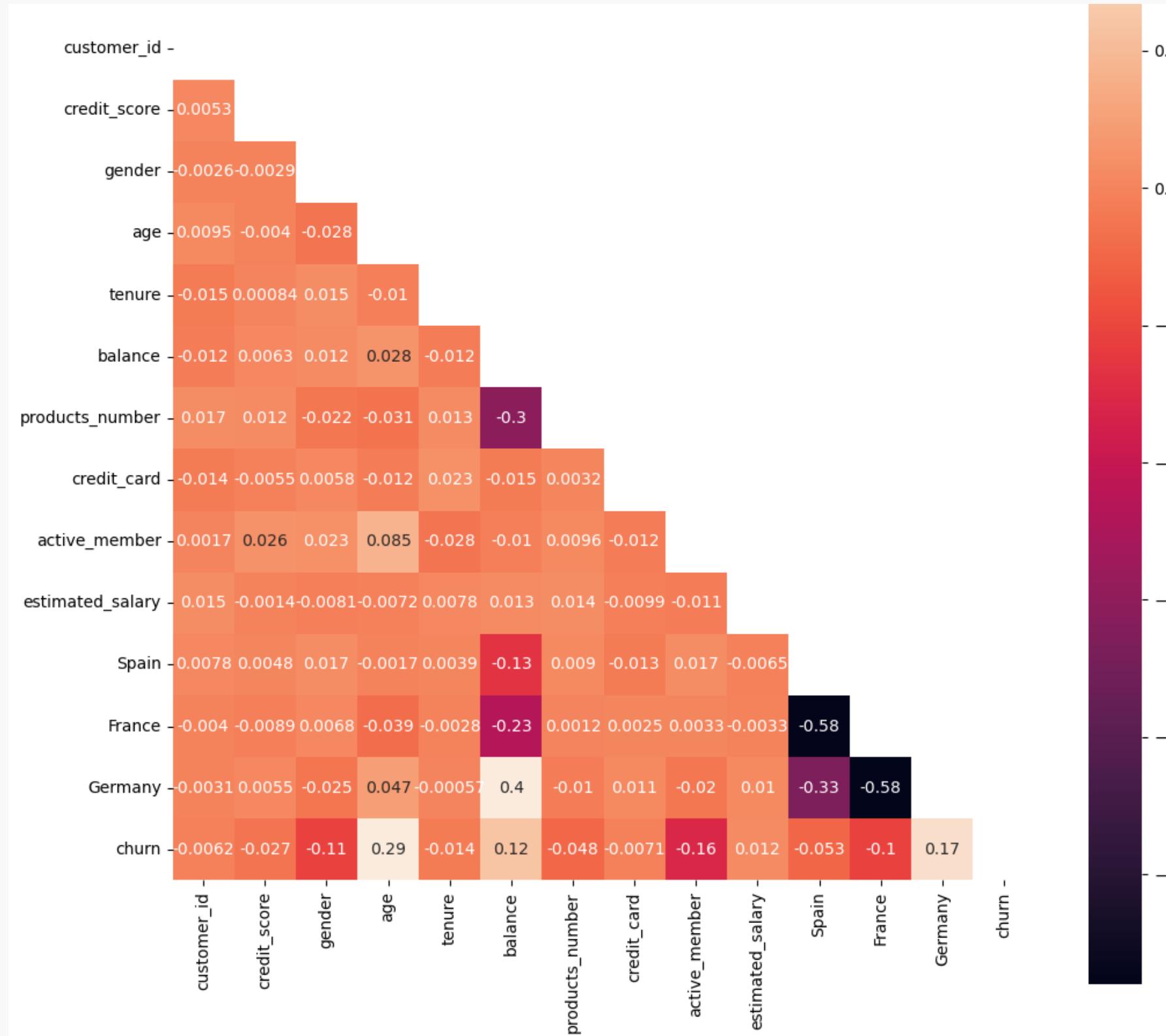
Describe berfungsi untuk mengetahui statistika data untuk data numeric seperti **Count, Mean, Standard deviation, Minimum dan Quartile**

EDA (EXPLORATORY DATA ANALYST)



FEATUR ANALYSIS

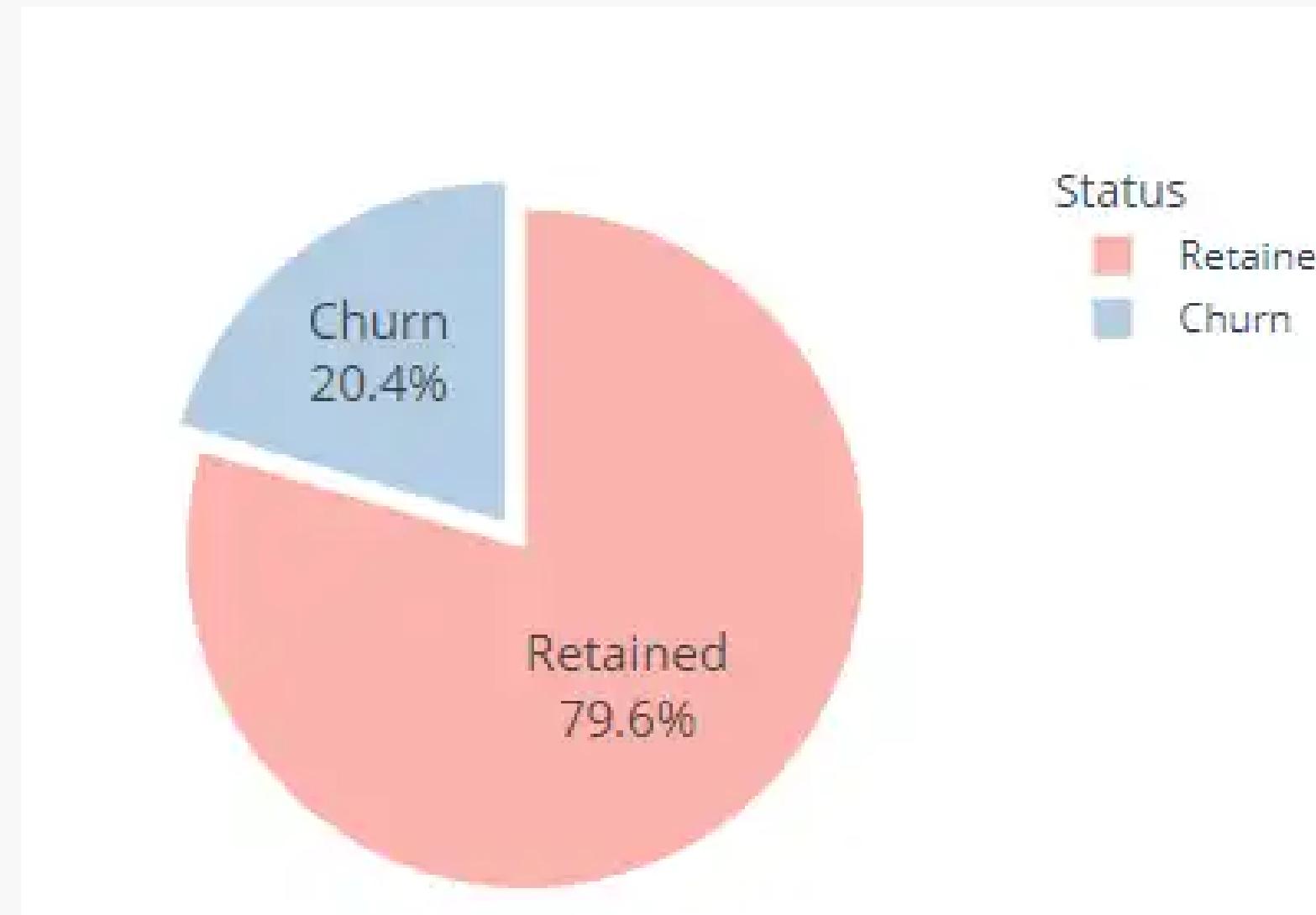
menggunakan korelasi untuk mencari feature - feature yang mempengaruhi Churn



Lima faktor (gender, usia, saldo akun, status anggota aktif, negara) menunjukkan korelasi yang relatif kuat dengan hasil "churn."

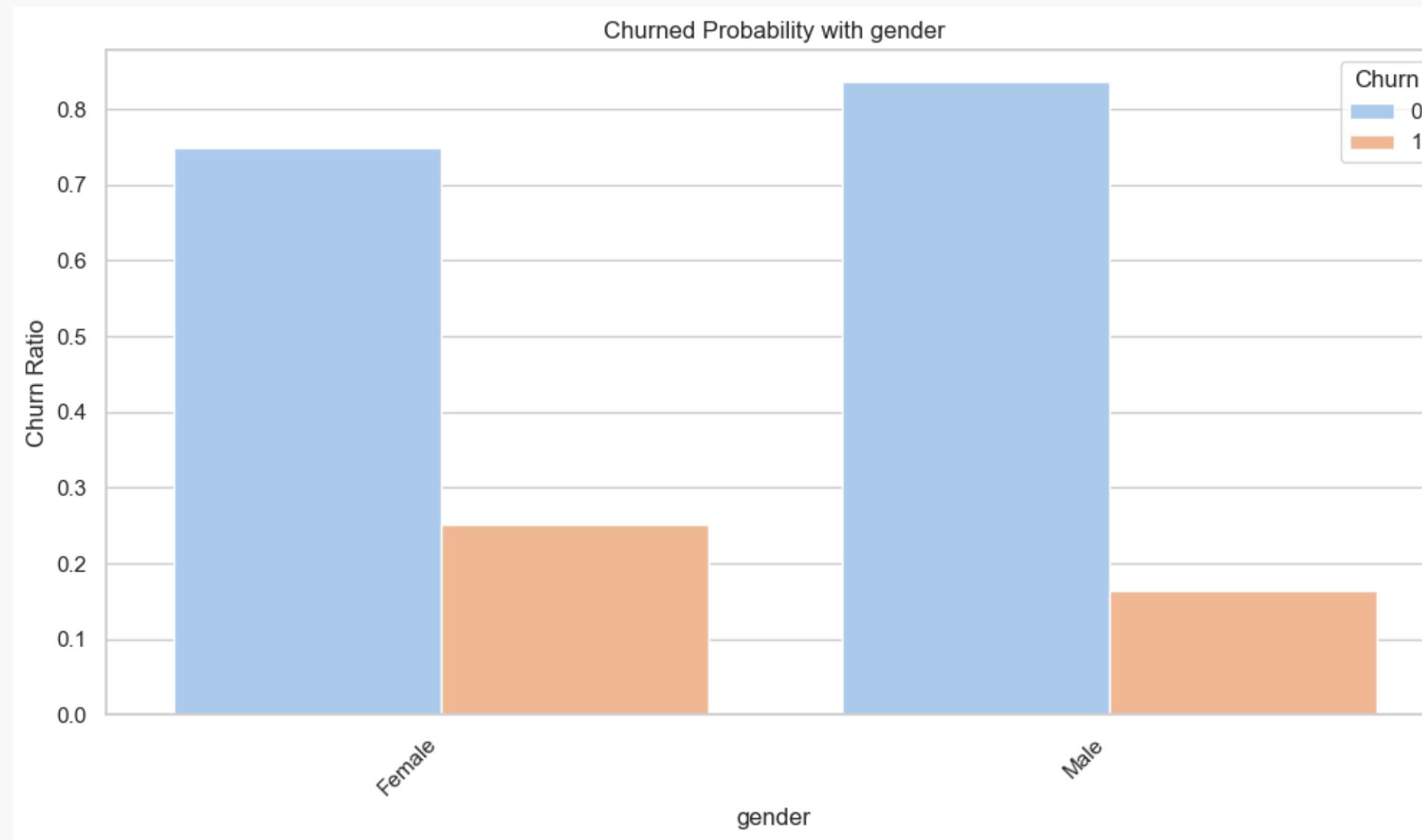
Feature gender, age, balance, active_member & country
untuk di analisa

CHURN ANALYSIS



Pie Chart menunjukkan proporsi jumlah customer yang churn dan retained. dari total 10.000 customer, sebanyak 20.4% (2.040 customer) merupakan churn, sedangkan 79,6% (7.960 customer) merupakan retained.

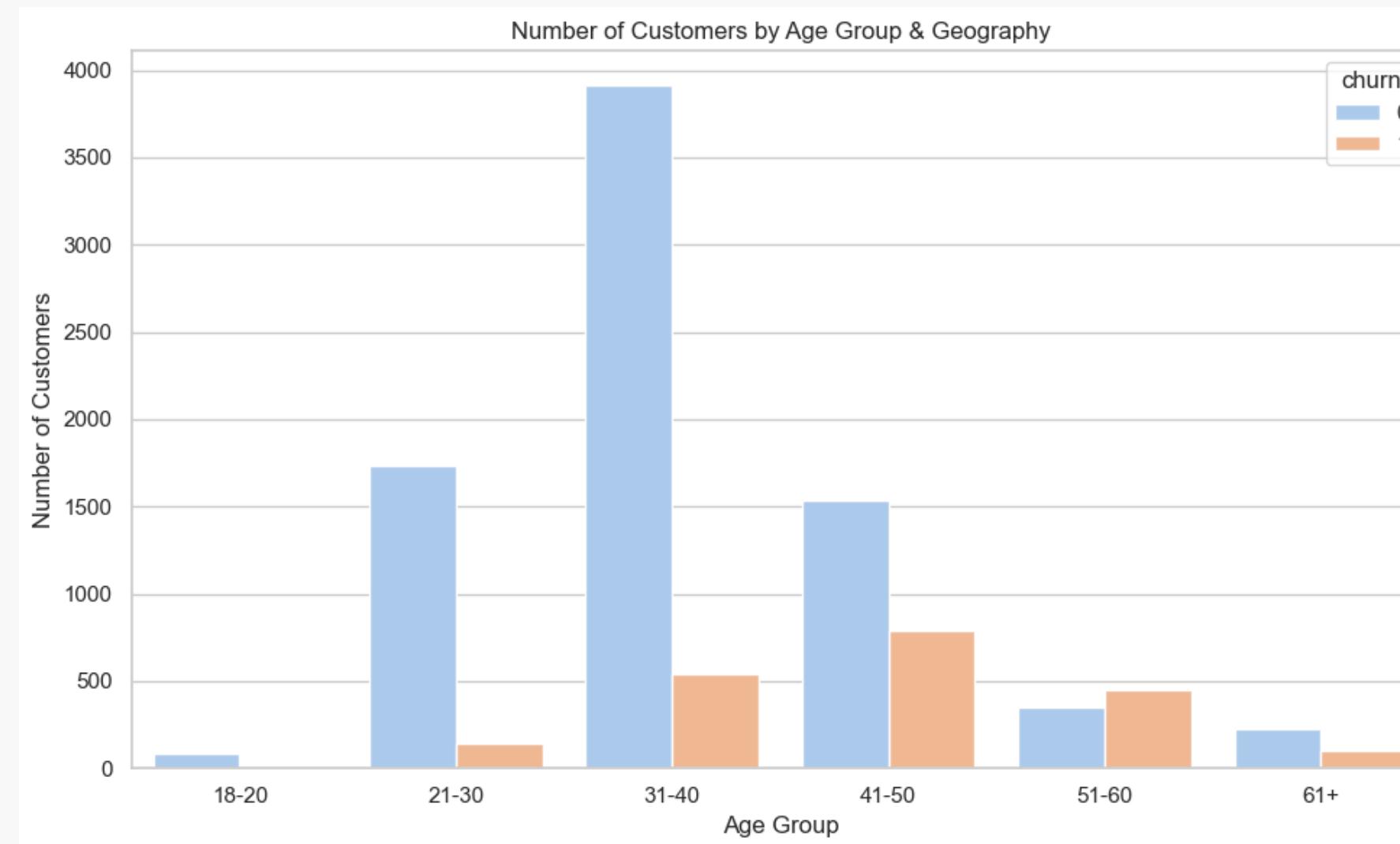
GENDER VS CHURN ANALYSIS



gender	churn	ratio
Female	0	0.749285
Female	1	0.250715
Male	0	0.835441
Male	1	0.164559

Berdasarkan gender, ratio churn terbesar terjadi pada Female sebesar 0.250715, Perubahan dalam prioritas atau kebutuhan pelanggan Female dapat memengaruhi keputusan mereka untuk berhenti menggunakan layanan atau produk tertentu.

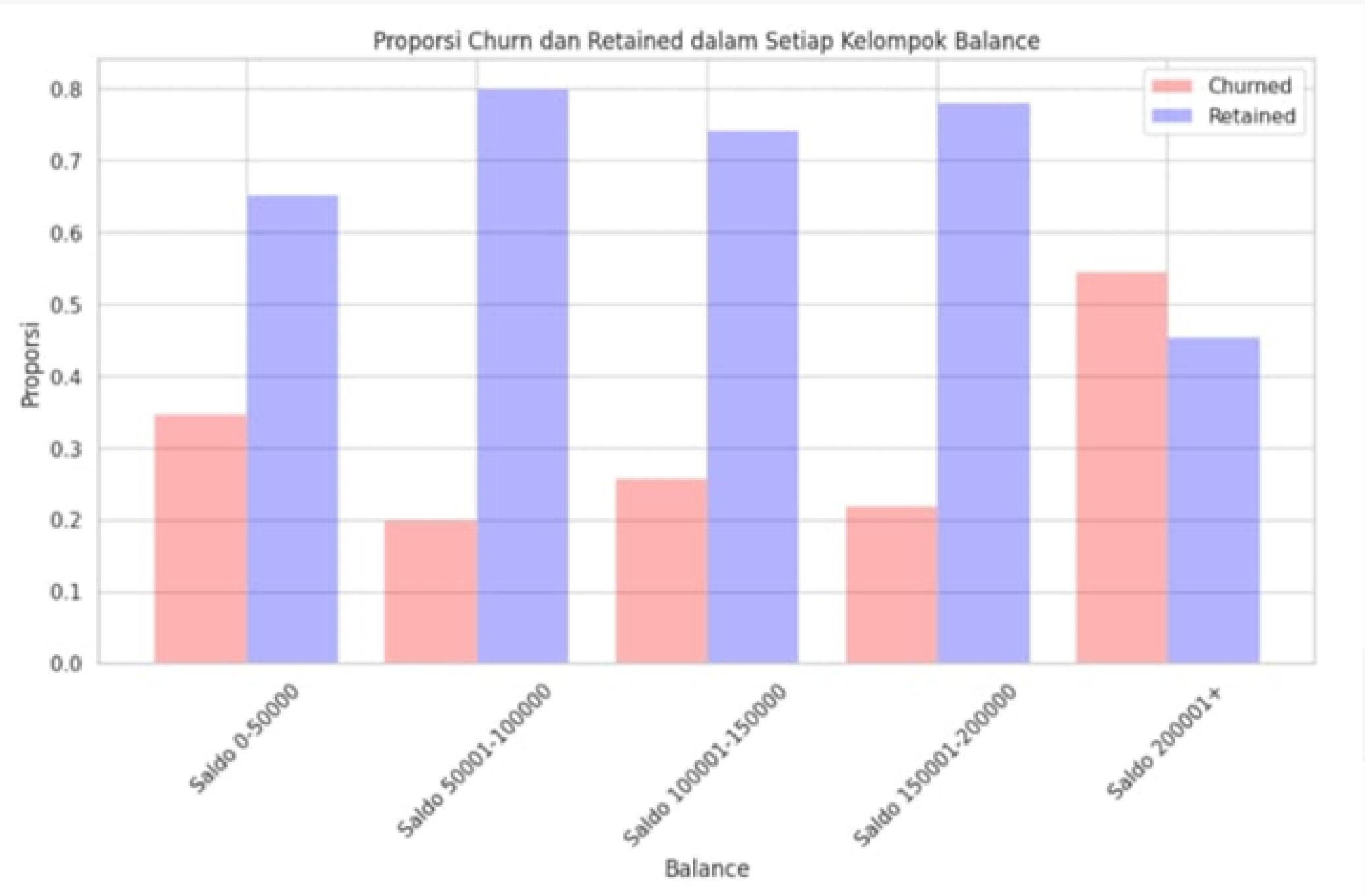
CUSTOMER AGE DISTRIBUTION VS CHURN



Age Group	Number of Customers
41-50	788
31-40	538
51-60	448
21-30	143
61+	104
18-20	5

menunjukkan bahwa orang-orang di rentang usia 49 hingga 57 tahun memiliki kemungkinan lebih tinggi untuk berhenti menggunakan layanan. Pada usia ini, banyak orang menghadapi beban keuangan yang lebih besar, seperti membeli rumah, pendidikan anak, dan persiapan pensiun. Ini dapat mengubah prioritas keuangan mereka dan membuat mereka mengurangi pengeluaran yang tidak penting, termasuk langganan layanan.

BALANCE VS CHURN

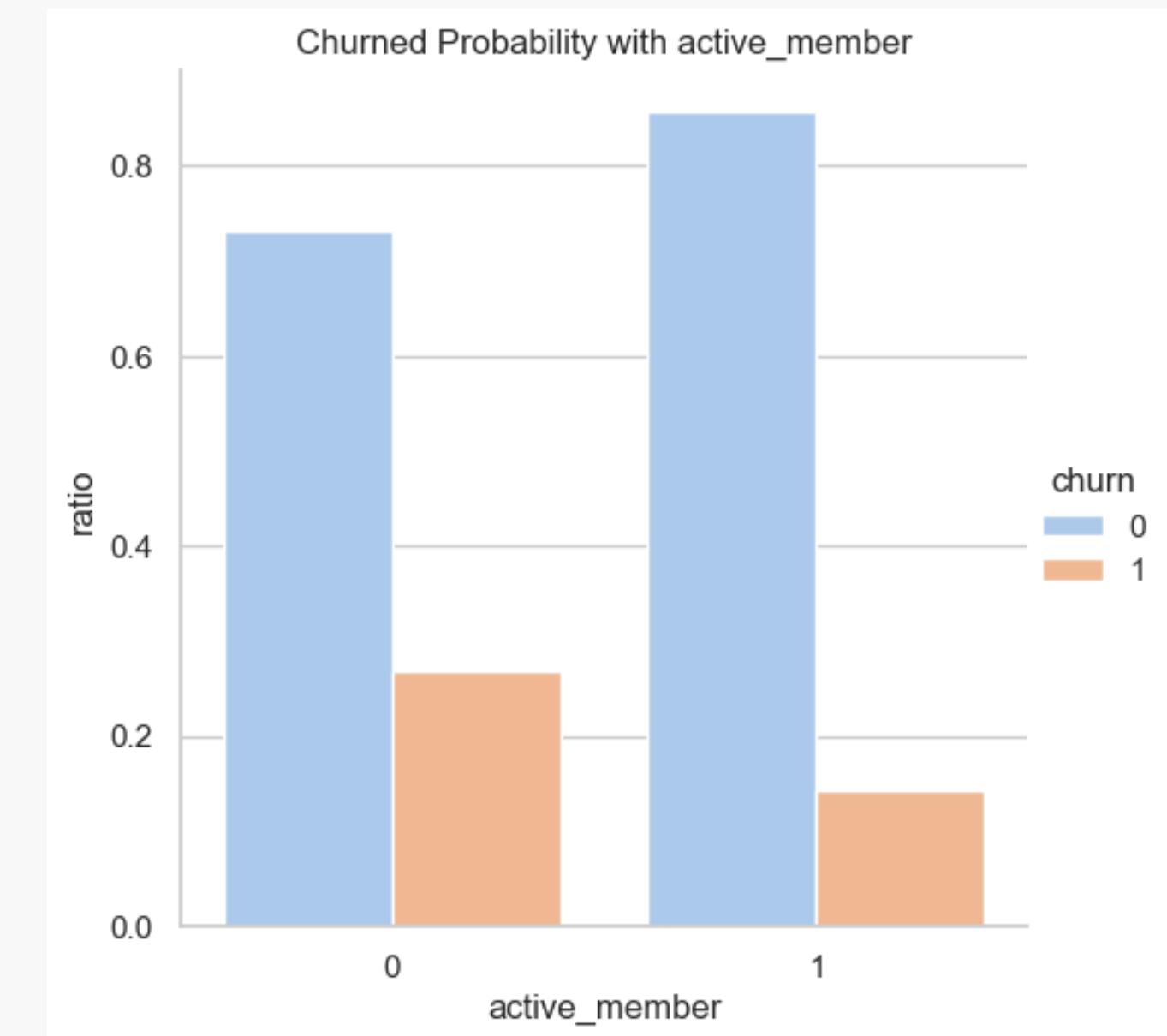


BALANCE_RANGE	CHURN	RATIO
0 Saldo 0-50000	0	0.653333
1 Saldo 0-50000	1	0.346667
2 Saldo 50001-100000	0	0.801193
3 Saldo 50001-100000	1	0.198807
4 Saldo 100001-150000	0	0.742298
5 Saldo 100001-150000	1	0.257702
6 Saldo 150001-200000	0	0.780749
7 Saldo 150001-200000	1	0.219251
8 Saldo 200001+	1	0.545455
9 Saldo 200001+	0	0.454545

Berdasarkan Saldo uang didalam akun customer (balance) terlihat bahwa churn terbesar terjadi pada kelompok saldo > 200,001 dengan proporsi 55% dan churn terkecil terjadi pada kelompok saldo antara 50,001 - 100,000 dengan proporsi 20%.

ACTIVE MEMBER VS CHURN

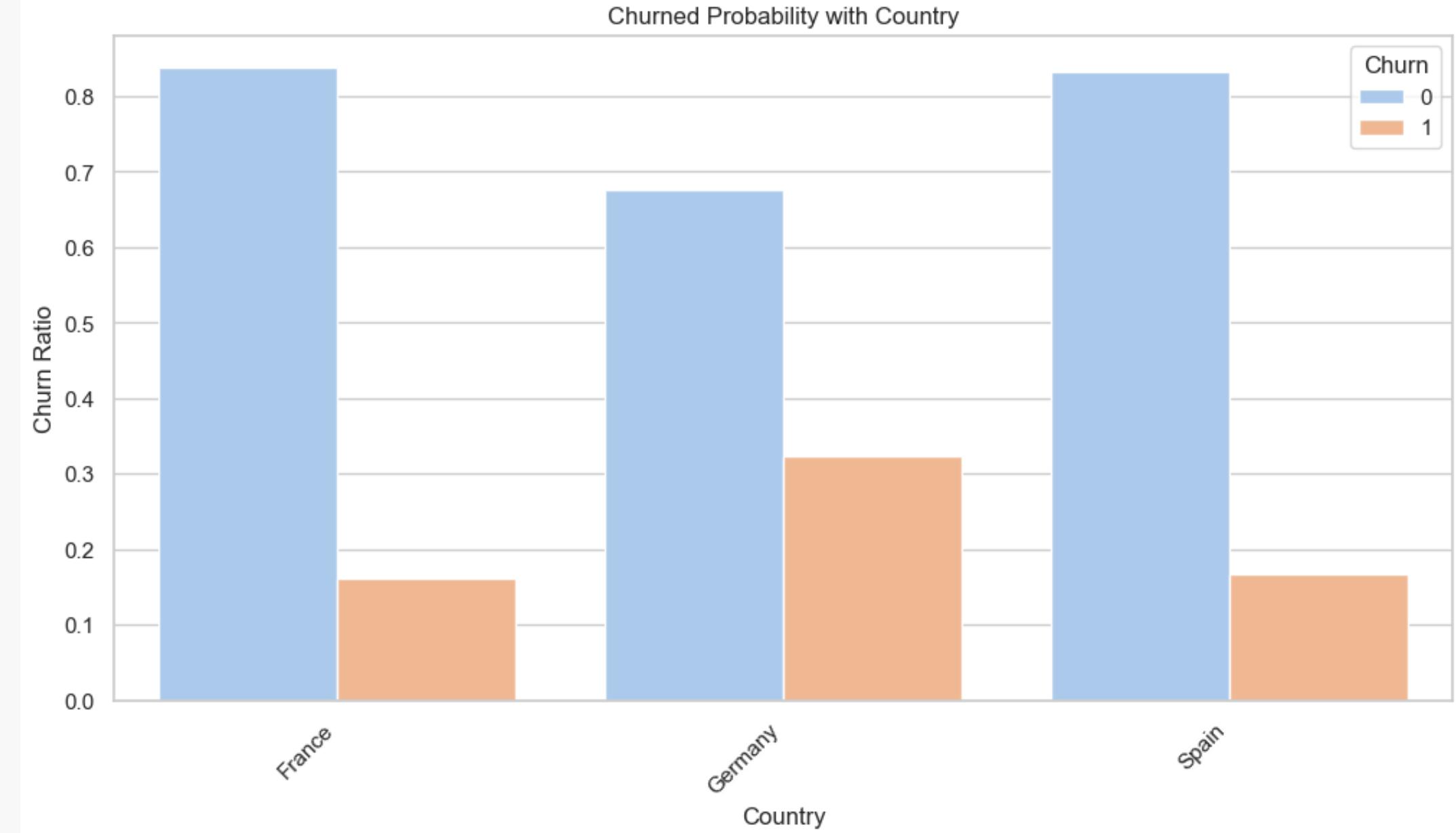
	active_member	churn	ratio
0	0	0	0.731491
1	0	1	0.268509
2	1	0	0.857309
3	1	1	0.142691



Temuan ini menunjukkan bahwa orang yang bukan pengguna aktif memiliki kemungkinan lebih besar untuk berhenti menggunakan layanan ("churn"). Artinya, mereka yang tidak aktif dalam menggunakan layanan cenderung meninggalkannya.

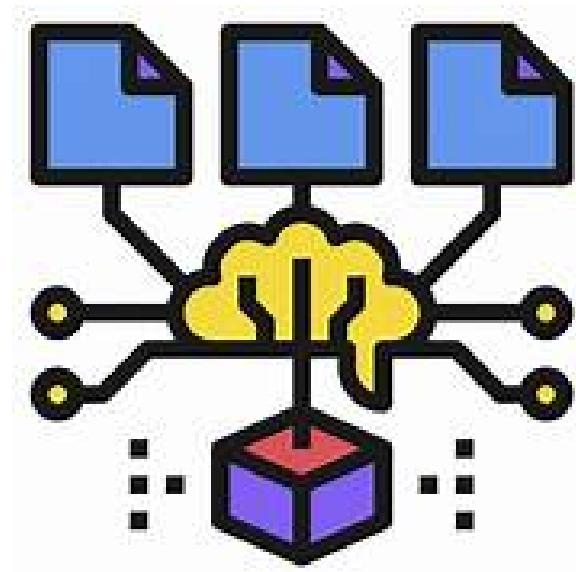
COUNTRY VS CHURN ANALYSIS

country	churn	ratio
France	0	0.838452
France	1	0.161548
Germany	0	0.675568
Germany	1	0.324432
Spain	0	0.833266
Spain	1	0.166734



Berdasarkan country, ratio churn terbesar terjadi untuk negara Germany sebesar 0.324432, diikuti oleh Spain dan France sebesar 0.166734 dan 0.161548.

DATA PREPROCESSING



DATA PREPROCESSING

	customer_id	credit_score	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	churn
count	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

[] data.dtypes

	customer_id	credit_score	country	gender	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	churn	dtype: object
customer_id													int64
credit_score													int64
country													object
gender													object
age													int64
tenure													int64
balance													float64
products_number													int64
credit_card													int64
active_member													int64
estimated_salary													float64
churn													int64
dtype: object													

Dari hasil analisa data, terdapat beberapa kolom yang masih berupa object, data-data pada kolom ini perlu dimanipulasi agar dapat dikenali oleh algoritma Machine Learning.

```
[ ] # List kolom-kolom yang ingin diubah menjadi tipe data kategori.  
columns_to_change = ["gender", "country"]  
  
# Menggunakan metode astype untuk mengubah tipe data kolom-kolom tersebut.  
data[columns_to_change] = data[columns_to_change].astype("category")
```

DATA PREPROCESSING

```
[ ] frames = [data,dummytf.fit_transform(data.country)]
dfNew=pd.concat(frames,axis=1,join='inner')
dfNew["gender"] = LabelEncoder().fit_transform(dfNew["gender"])
# move the column to end of list using index, pop and insert
columns = list(dfNew)
columns.insert(900, columns.pop(columns.index('churn')))
dfNew = dfNew.loc[:, columns]

dfNum=dfNew.drop('country',axis=1)
dfNum
```

	customer_id	credit_score	gender	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	Spain	Germany	France	churn	
0	15634602	619	0	42	2	0.00		1	1	1	101348.88	0	0	1	1
1	15647311	608	0	41	1	83807.86		1	0	1	112542.58	1	0	0	0
2	15619304	502	0	42	8	159660.80		3	1	0	113931.57	0	0	1	1
3	15701354	699	0	39	1	0.00		2	0	0	93826.63	0	0	1	0
4	15737888	850	0	43	2	125510.82		1	1	1	79084.10	1	0	0	0
...	
9995	15606229	771	1	39	5	0.00		2	1	0	96270.64	0	0	1	0
9996	15569892	516	1	35	10	57369.61		1	1	1	101699.77	0	0	1	0
9997	15584532	709	0	36	7	0.00		1	0	1	42085.58	0	0	1	1
9998	15682355	772	1	42	3	75075.31		2	1	0	92888.52	0	1	0	1
9999	15628319	792	0	28	4	130142.79		1	1	0	38190.78	0	0	1	0

10000 rows × 14 columns

Proprocessing kedua yang dilakukan adalah Encoding. Kita menggunakan metode One-Hot Encoding contohnya terlihat pada Kolom Country yang dibagi kedalam 3 Kolom dengan isi indeks 1 atau 0 yang merepresentasikan masing-masing negara pada data awal.

DATA PREPROCESSING

Setelah dilakukan Data Preprocessing, didapatkan Dataset “dfNum” yang bisa diproses/dikenali dengan lebih baik oleh Model Machine Learning. Untuk selanjutnya digunakan Dataset “dfNum” untuk diproses dengan Machine Learning.

▶ dfNum

👤

	customer_id	credit_score	gender	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	Spain	Germany	France	churn	
0	15634602	619	0	42	2	0.00		1	1	1	101348.88	0	0	1	1
1	15647311	608	0	41	1	83807.86		1	0	1	112542.58	1	0	0	0
2	15619304	502	0	42	8	159660.80		3	1	0	113931.57	0	0	1	1
3	15701354	699	0	39	1	0.00		2	0	0	93826.63	0	0	1	0
4	15737888	850	0	43	2	125510.82		1	1	1	79084.10	1	0	0	0
...	
9995	15606229	771	1	39	5	0.00		2	1	0	96270.64	0	0	1	0
9996	15569892	516	1	35	10	57369.61		1	1	1	101699.77	0	0	1	0
9997	15584532	709	0	36	7	0.00		1	0	1	42085.58	0	0	1	1
9998	15682355	772	1	42	3	75075.31		2	1	0	92888.52	0	1	0	1
9999	15628319	792	0	28	4	130142.79		1	1	0	38190.78	0	0	1	0

10000 rows × 14 columns

MACHINE LEARNING MODELING



MACHINE LEARNING MODELING

Sebelum Data dapat diproses kedalam algoritma machine learning, ada beberapa manipulasi yang perlu dilakukan terhadap data yaitu :

1. Pisahkan dependent dan independent variabel
2. Memisahkan data menjadi dataset train dan test

```
[ ] #Memisahkan dataset dependent dan independent variables

#deklarasi response variable:
response = dfNum["churn"]

dfNum = dfNum.drop(columns="churn")
```

```
[ ] #Membuat data set training dan test dari dependent dan independent variables

X_train, X_test, y_train, y_test = train_test_split(dfNum, response,
stratify=response,
test_size = 0.2,
random_state = 0)
```

```
print("Number transactions X_train dataset: ", X_train.shape)
print("Number transactions y_train dataset: ", y_train.shape)
print("Number transactions X_test dataset: ", X_test.shape)
print("Number transactions y_test dataset: ", y_test.shape)
```

```
Number transactions X_train dataset: (8000, 13)
Number transactions y_train dataset: (8000,)
Number transactions X_test dataset: (2000, 13)
Number transactions y_test dataset: (2000,)
```

MACHINE LEARNING MODELING

Scalling atau standarisasi dilakukan untuk membuat data numerical atau angka memiliki rentang yang sama pada dataset yang sudah ada.

```
[ ] # Melakukan scalling (Standarisasi) pada feature

sc_X = StandardScaler()
X_train2 = pd.DataFrame(sc_X.fit_transform(X_train))
X_train2.columns = X_train.columns.values
X_train2.index = X_train.index.values
X_train = X_train2

X_test2 = pd.DataFrame(sc_X.transform(X_test))
X_test2.columns = X_test.columns.values
X_test2.index = X_test.index.values
X_test = X_test2
```

MACHINE LEARNING MODELING

```
#Bandingkan Algoritma Klasifikasi - Iterasi Pertama  
#Bandingkan Accuracy and ROC AUC Mean Metrics  
  
models = []  
  
models.append(('Logistic Regression', LogisticRegression(solver='liblinear',  
                                         random_state = 0,  
                                         class_weight='balanced')))  
  
models.append(('Decision Tree Classifier',  
             DecisionTreeClassifier(criterion = 'entropy', random_state = 0)))  
  
models.append(('Random Forest', RandomForestClassifier(  
                                         n_estimators=100, criterion = 'entropy', random_state = 0)))
```

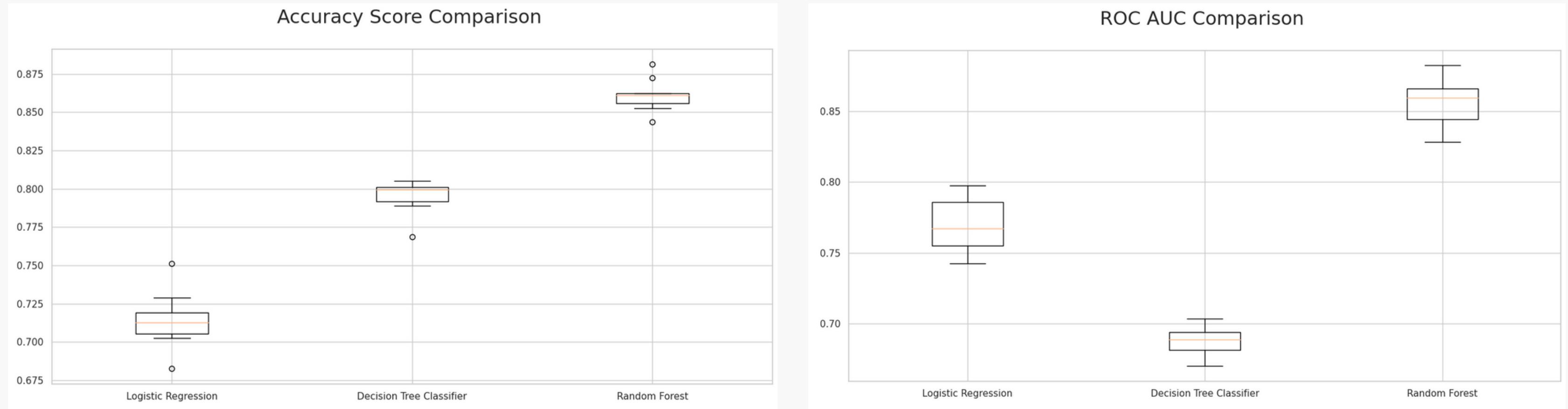
Berikut adalah salah satu indikasi dalam memilih suatu model learning yang terbaik untuk digunakan.

- Membandingkan Algoritma Klasifikasi - Iterasi Pertama
- Membandingkan Akurasi dan ROC AUC Mean Metrics

	Algorithm	ROC AUC Mean	ROC AUC STD	Accuracy Mean	Accuracy STD
2	Random Forest	85.57	1.74	86.11	0.98
0	Logistic Regression	76.97	1.85	71.38	1.72
1	Decision Tree Classifier	68.70	0.99	79.54	1.03

Random Forest memiliki nilai Akurasi dan nilai ROC AUC yang paling besar. Dipilihlah Random Forest untuk algoritma machine learning.

MACHINE LEARNING MODELING



Berikut merupakan perbandingan dari ketiga machine learning pada akurasi dan juga ROC AUC. Random Forest memiliki nilai terbesar pada kedua perbandingan tersebut. Keuntungan dari Random Forest yaitu fleksibel dan mudah digunakan, serta mampu memprediksi dengan akurat dan stabil.

MACHINE LEARNING MODELING

Hasil Train dengan data Test :

```
[ ] #Train pada Data Test dan Evaluasi Model Pilihan

# Fit Random Forest terhadap Training dataset:

classifier = RandomForestClassifier(random_state = 0)
classifier.fit(X_train, y_train)

# Prediksi Test set results

y_pred = classifier.predict(X_test)

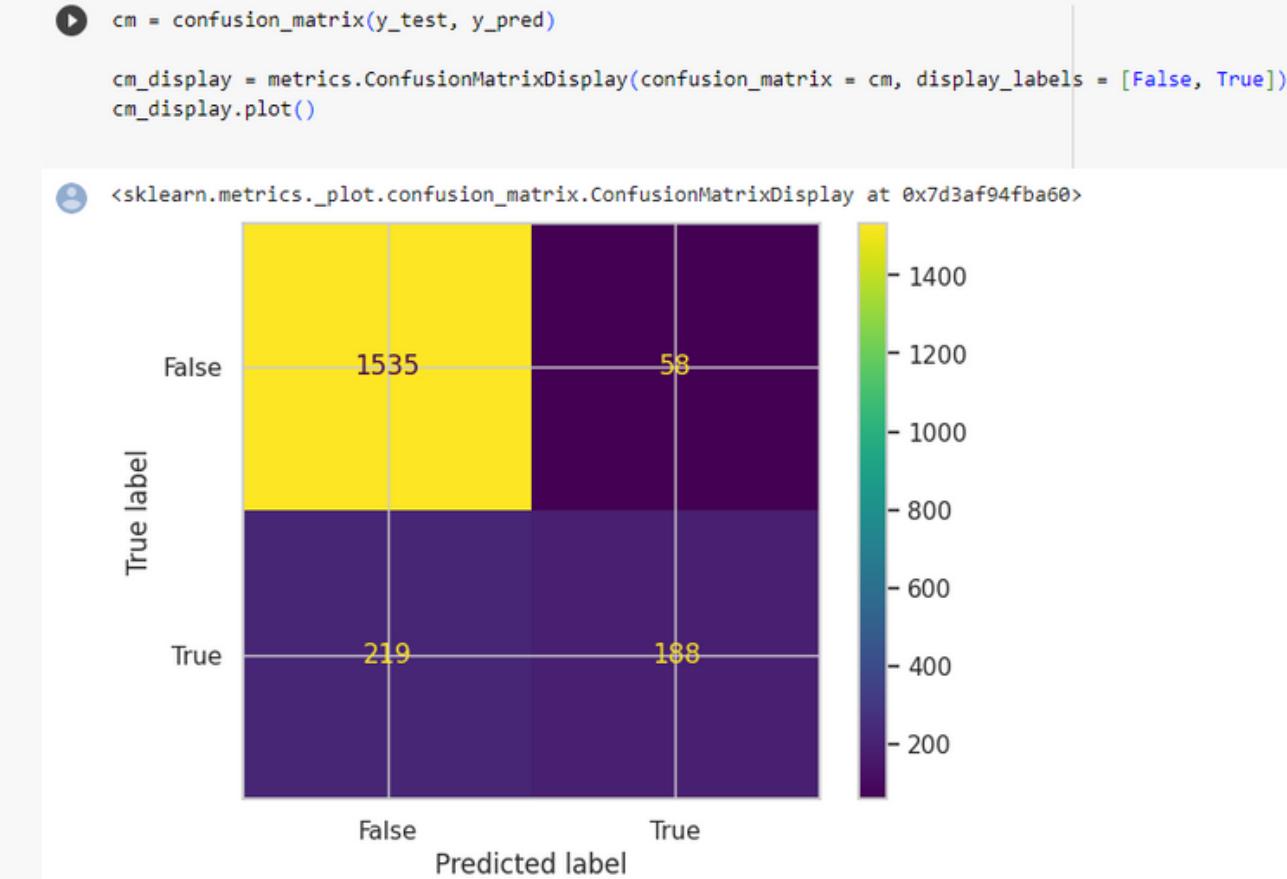
pd.DataFrame(y_pred)[0].value_counts()

0    1754
1     246
Name: 0, dtype: int64
```

Hasil Train dengan data Test :

1754 bernilai 0 / False

246 bernilai 1/ True



Diplot kedalam Confusion Matrix

Dari 1754 data, 219 merupakan False Negatif

Dari 246 data. 58 merupakan False Positif

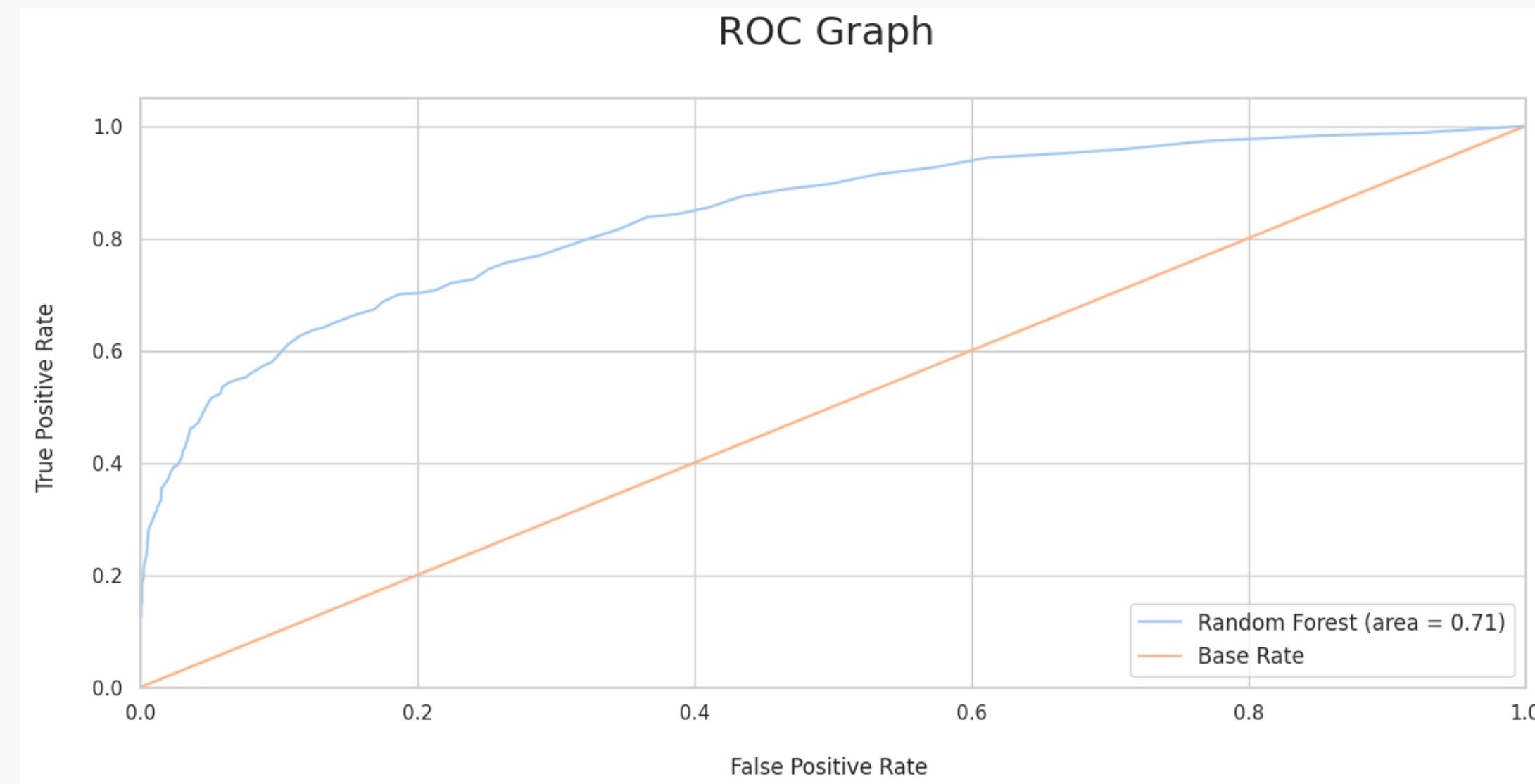
Akurasi = True Positif + True Negatif / Total Data Test

$$= ((1535 + 188)/2000)*100\%$$

$$= 86.15\%$$

MACHINE LEARNING MODELING

Evaluasi Model dengan ROC Graph : Menghitung luas area dibawah kurva ROC



Dari kurva diatas, didapatkan nilai Area Under Curve untuk Kurva ROC sebesar 0.71 yang mengindikasikan bahwa model memiliki kualitas yang lebih baik dibandingkan memprediksi secara acak (Nilai AUC-ROC >0.5).

CONCLUSION



CONCLUSION

- Dari ketiga machine learning yang sudah dilakukan test, dapat disimpulkan menggunakan machine learning Random Forest dengan pertimbangan nilai akurasi dan AUC ROC terbesar.
- Rentang umur yang rentan untuk berhenti berlangganan yaitu 49 hingga 57 tahun, karena mengurangi beban pengeluaran yang tidak prioritas. Kita bisa budgeting secara personal, agar pelanggan memiliki opsi untuk melanjutkan.
- Saldo >200.000 menjadi penyumbang churn terbesar dan hal ini bisa kita kurangi angka churn dengan memberikan reward kepada customer yang sudah memiliki saldo >200.000 agar tetap bertahan.
- Customer yang tidak aktif cenderung berhenti berlangganan, cara mengatasinya yaitu kita bisa memunculkan notifikasi dan memberikan penawaran yang menarik dengan promo yang berkelanjutan, agar customer bisa kembali aktif untuk rentang waktu yang sudah ditentukan.

MODEL DEPLOYMENT

← → ⌂ bankcustomer-churn.streamlit.app

Customer ID
72322212 - +

Credit Score
645 - +

Gender
 Male
 Female

Age
44 - +

Tenure
8 - +

Balance
113755,78 - +

Number of Products
2 - +

Has Credit Card

Active Member

Estimated Salary
149756,71 - +

Country
France

Predict



Welcome to Customer Churn Prediction

Prediction Result

Customer: Churn

Prediction Probabilities

Probability of Churn: 59.00%

Probability of Retained: 41.00%

PEMBAGIAN TUGAS:

EDA & VISUALIZATION :

- 1. SUSI**
- 2. SONI AGUNG WAHYUDIYANTA**

MACHINE LEARNING MODEL & DEPLOYMENT :

- 1. AL-AZHARY PUTERA SATRIA**
- 2. KATON CAHYO ANDARU**
- 3. DINDA THALIA FAHIRA**

THANK YOU