Final Project Presentation

Nomor Kelompok: 5
Nama Mentor: Aditya Bariq Ikhsan
Nama:

- Noeril Agian Septa Dinata
- Stefanus Adyan Mardhikaputra

Machine Learning Class

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

Latar Belakang

Latar Belakang Project

Sumber Data: https://www.kaggle.com/datasets/barun2104/telecom-churn?datasetId=567482
Problem: **Classification**

Tujuan:

- Prediksi churn pada data pelanggan telekomunikasi dan membuat recomendasi ke perusahaan apa yang dilakukan agar custumer tidak churn

Explorasi Data dan Visualisasi

Business Understanding

Telecom_churn, dalam industri telekomunikasi, pelanggan dapat memilih dari beberapa penyedia layanan dan secara aktif beralih dari satu operator ke operator lainnya. Faktor-faktor yang mempengaruhi churn pada data telekomunikasi yaitu pelayanan, kontrak, dan biaya bulanan.

Churn merupakan pelanggan yang keluar atau tidak berlangganan kembali dari suatu bisnis, churn itu harus diantisipasi karena berdampak kepada kerugian bisnis.

Column Definition

Churn = 1 if customer cancelled service, 0 if not
AccountWeeksnumber = of weeks customer has had active account
ContractRenewal = 1 if customer recently renewed contract, 0 if not
DataPlan = 1 if customer has data plan, 0 if not
DataUsagegigabytes = of monthly data usage
CustServCallsnumber = of calls into customer service
DayMinsaverage = daytime minutes per month
DayCallsaverage = number of daytime calls
MonthlyChargeaverage = monthly bill
OverageFeelargest = overage fee in last 12 months

Data Cleansing

Data telecom churn ini tidak memiliki missing value, sudah dicek pada type data, jumlah baris, dan mencari nilai kosong. Semua tindakan yang dilakukan tidak menemukan data yang perlu diubah atau dihapus. Dimensi pada data telecom churn adalah baris, colom (3333,11).

Pada data memiliki outlier namun tidak diubah, karena nanti akan berdampak kepada model yang akan dibuat.

Data Cleansing

```
#check data missing
df.isna().sum()
```

```
Churn              0
AccountWeeks       0
ContractRenewal    0
DataPlan           0
DataUsage          0
CustServCalls      0
DayMins            0
DayCalls           0
MonthlyCharge      0
OverageFee         0
RoamMins           0
dtype: int64
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3333 entries, 0 to 3332
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Churn            3333 non-null   int64  
 1   AccountWeeks     3333 non-null   int64  
 2   ContractRenewal  3333 non-null   int64  
 3   DataPlan         3333 non-null   int64  
 4   DataUsage        3333 non-null   float64
 5   CustServCalls    3333 non-null   int64  
 6   DayMins          3333 non-null   float64
 7   DayCalls         3333 non-null   int64  
 8   MonthlyCharge    3333 non-null   float64
 9   OverageFee       3333 non-null   float64
 10  RoamMins         3333 non-null   float64
dtypes: float64(5), int64(6)
memory usage: 286.6 KB
```

```
#check duplicate
df.duplicated().any()
```

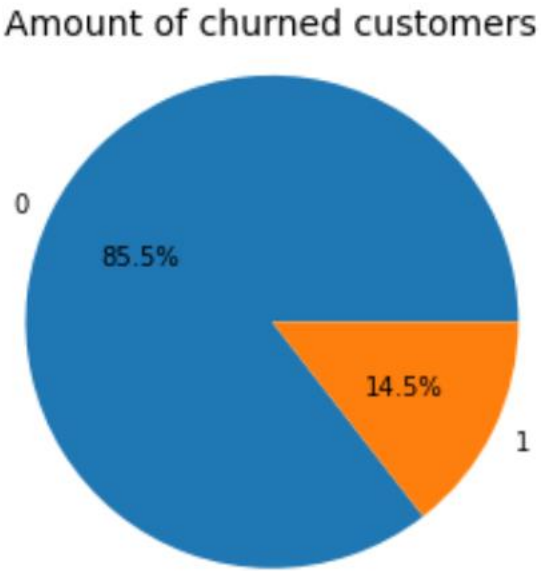
```
False
```

Exploratory Data Analysis

Jumlah no churn (0) dan churn (1)

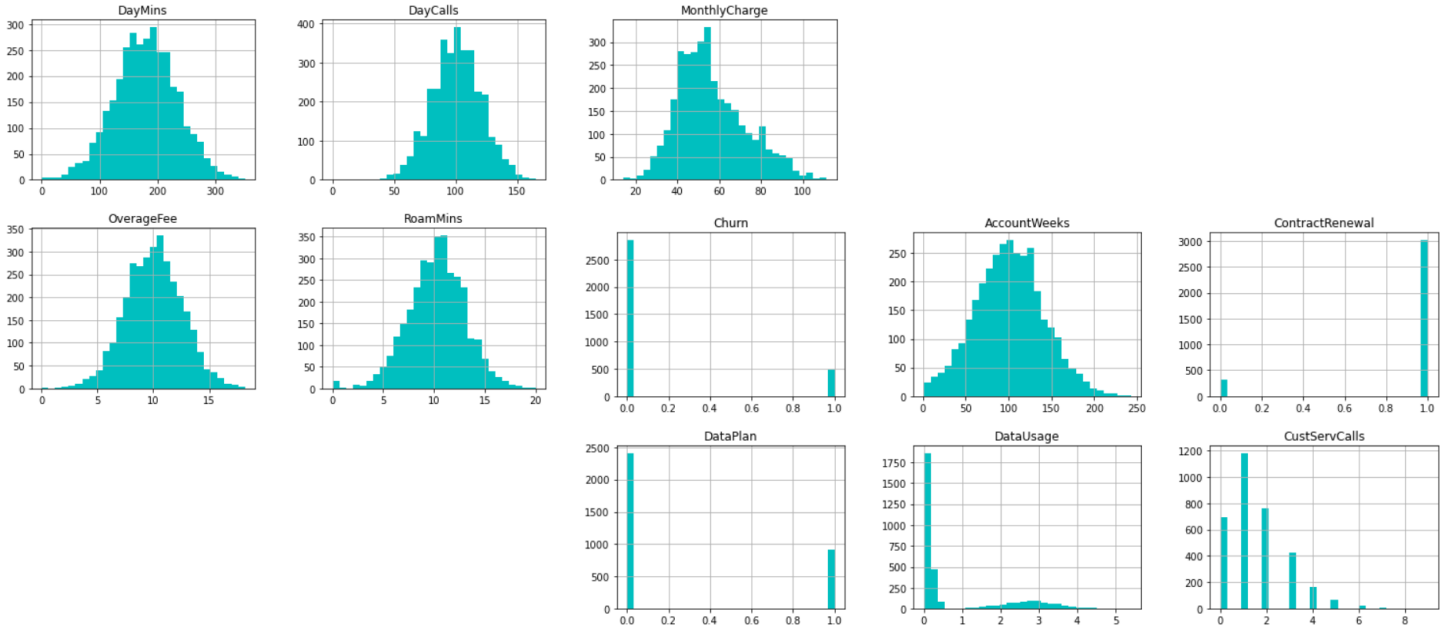
```
#melihat jumlah churn and no churn
df["Churn"].value_counts()
```

```
0    2850
1     483
Name: Churn, dtype: int64
```

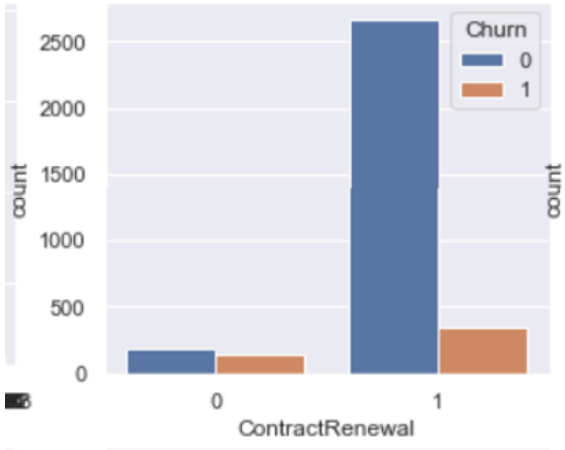


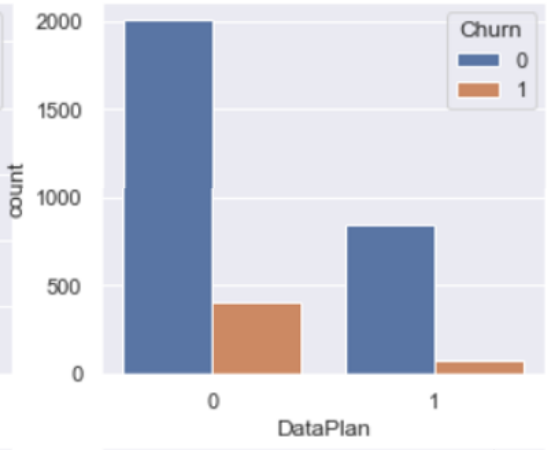
Exploratory Data Analysis

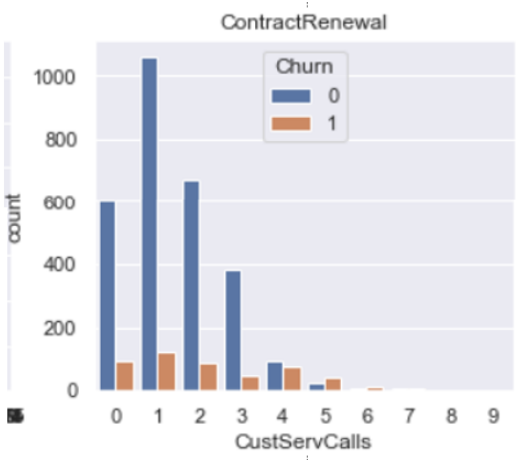
Histogram :



Exploratory Data Analysis

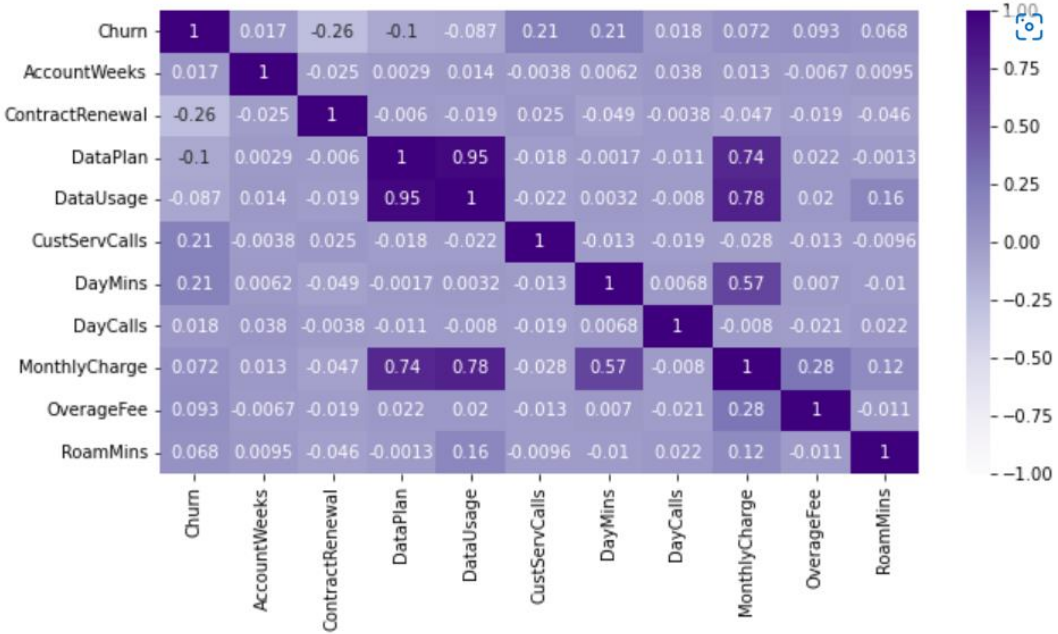






Exploratory Data Analysis

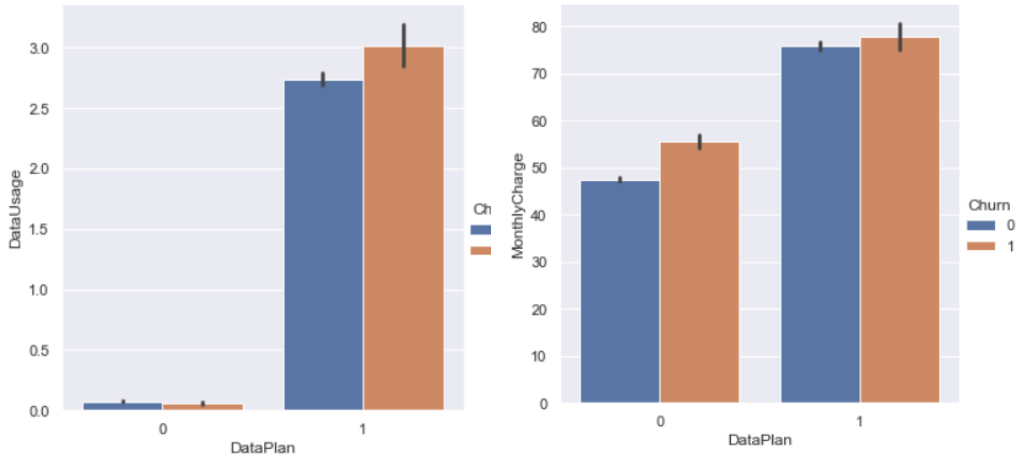
Correlation : Pada gambar ini dapat diketahui nilai tertinggi



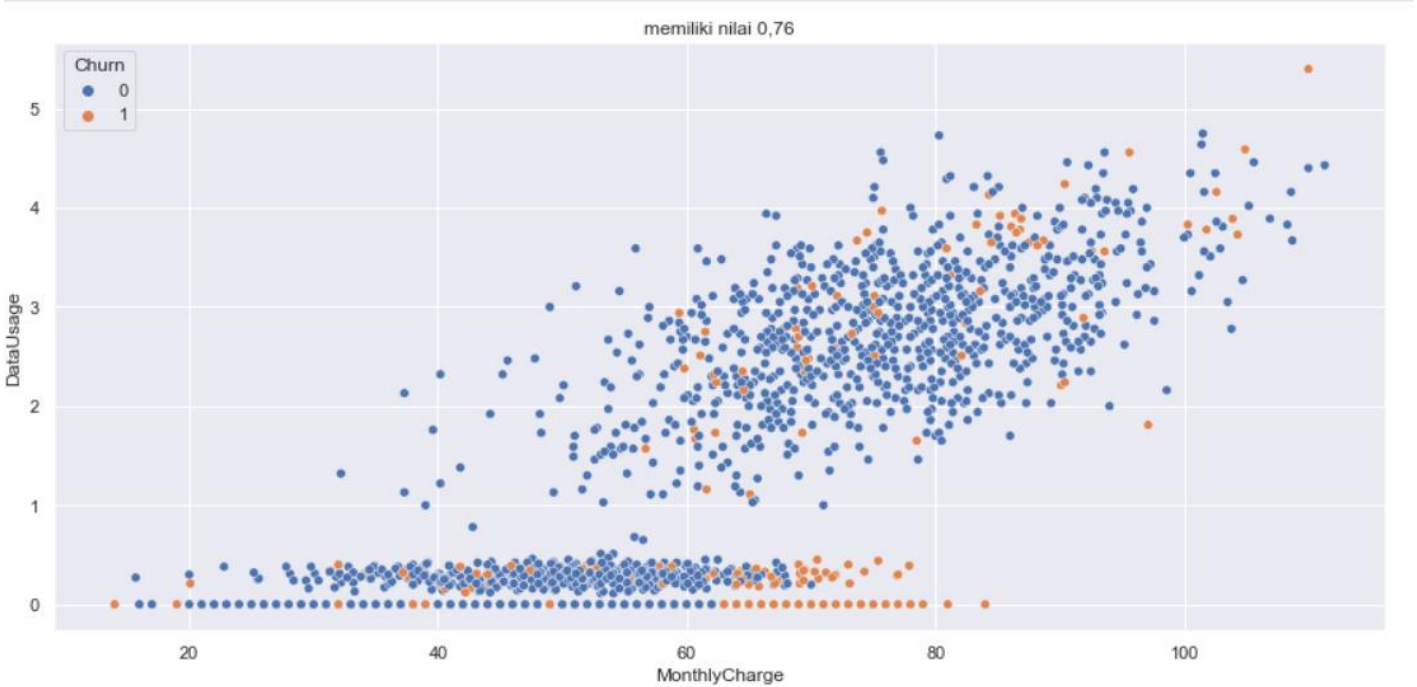
Exploratory Data Analysis

Visualisasi HeatMap :

```
# nilai 9,5
sns.catplot(x="DataPlan", y="DataUsage", hue="Churn", kind="bar", data=df)
# memiliki nilai 0,74
sns.catplot(x="DataPlan", y="MonthlyCharge", hue="Churn", kind="bar", data=df)
```



Exploratory Data Analysis

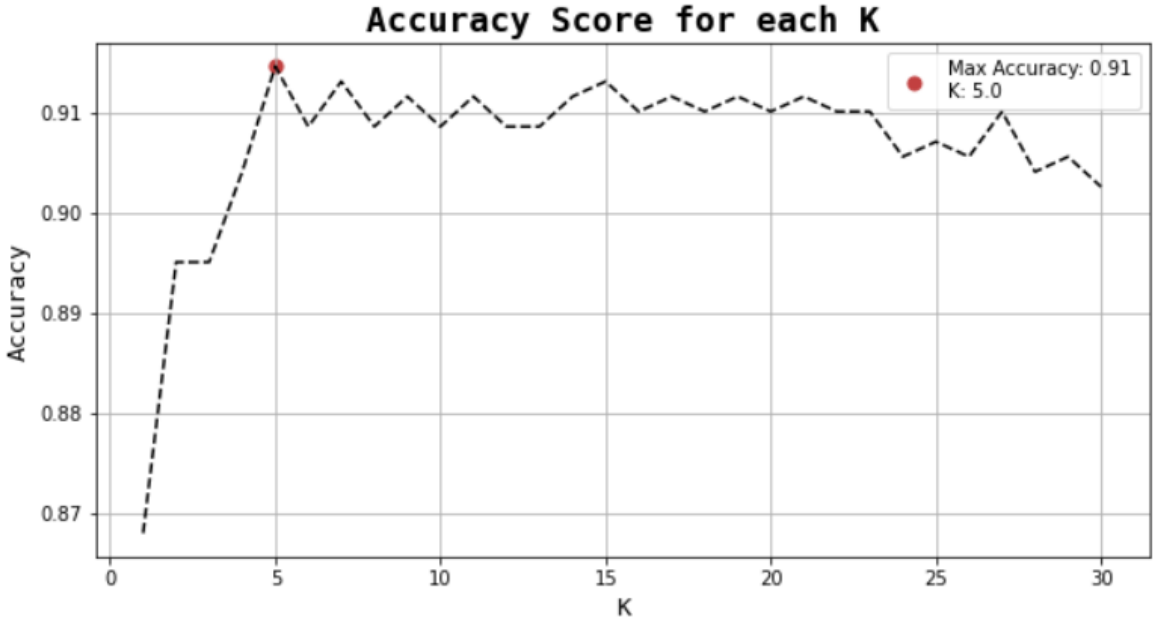


Exploratory Data Analysis

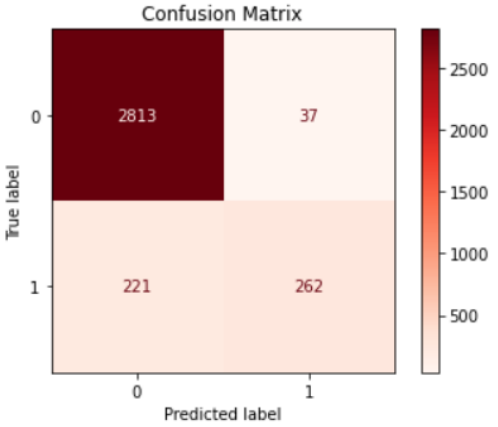
Modelling

Modelling

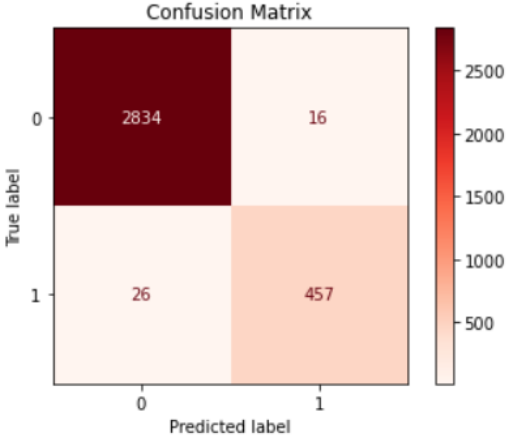
- Metode train test split / cross validation yang digunakan yaitu model list **K-Fold Cross Validation**
- Metrik untuk melakukan evaluasi (Confusion Matrik)
- Jenis model awal yang dicoba (KNN)
- Jenis model lain yang turut dicoba, serta tindakan-tindakan apa saja yang dilakukan untuk mencoba menambah akurasi model (random forest, DecisionTree)
- Model final (random forest) karena memiliki akurasi yang tinggi
- Kolom-kolom apa saja yang menjadi prediktor dan target variable untuk model final (DayMins, MonthlyCharge, CustServCalls dan OverageFee)
- Menggunakan pustaka pembelajaran mesin scikit-learn untuk melakukan prosedur pemisahan train-test.
- Cara mengevaluasi algoritma pembelajaran mesin untuk klasifikasi
- Menggunakan pustaka pembelajaran mesin scikit-learn untuk melakukan prosedur pemisahan train-test.



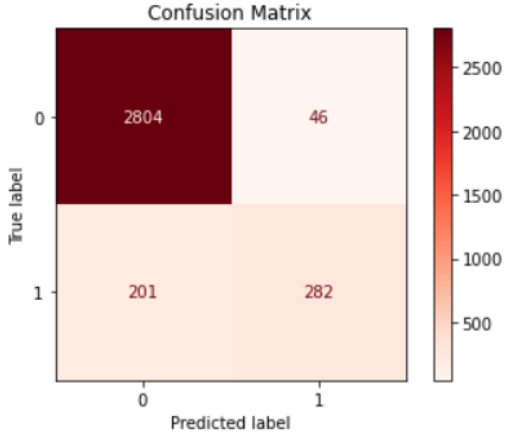
KNeighborsClassifier
Accuracy Score : 0.9145427286356822



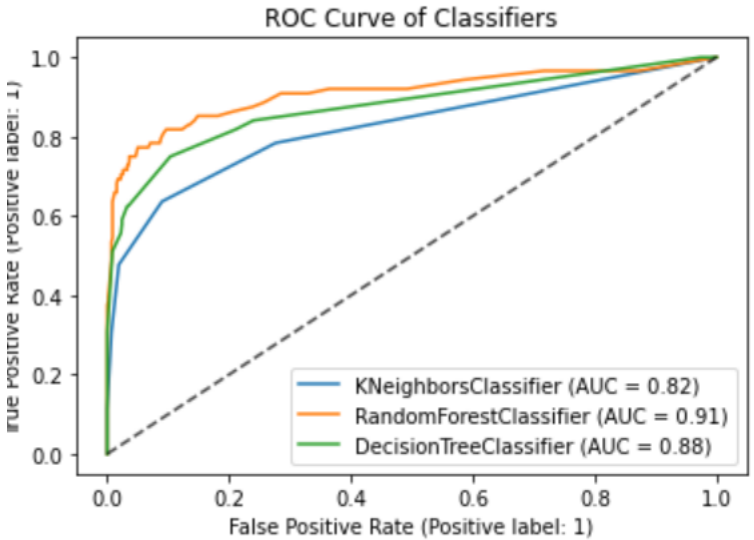
RandomForestClassifier
Accuracy Score : 0.9370314842578711



DecisionTreeClassifier
Accuracy Score : 0.9220389805097451



- Perbandingan nilai dari 3 model



Cross Validation Score (CVS)

- Perbandingan CVS dari 3 model machine learning

	Model Name	CVS
0	KNeighborsClassifier	0.903088
1	RandomForestClassifier	0.936397
2	DecisionTreeClassifier	0.921394

- Ada 4 kolom dengan nilai tertinggi

	feature importance
DayMins	0.210207
MonthlyCharge	0.147182
CustServCalls	0.140063
OverageFee	0.106054
RoamMins	0.086793
DataUsage	0.077845
ContractRenewal	0.070375
AccountWeeks	0.064755
DayCalls	0.059946
DataPlan	0.036780

Conclusion

Saran dan Kesimpulan

- Mengoptimalkan harga waktu bicara (DayMins)
- Mempertahankan biaya bulanan atau memberi diskon kepada pelanggan, agar pelanggan tetap no churn (MonthlyCharge)
- Tingkatkan atau pertahankan pelayanan service, karena cutumer yang bertanya atau memanfaatkan pelayanan, sangat rentan terhadap churn (CustServCalls)

Terima kasih!

Ada pertanyaan?