

zenius



Kampus  
Merdeka  
INDONESIA JAYA

# Final Project Presentation

Nomor Kelompok: 4

Nama Mentor: Rachmadio Noval L

Nama:

- Alif Alfarisi
- Rahwini Harpa Helda

Machine Learning Class

Program Studi Independen Bersertifikat  
Zenius Bersama Kampus Merdeka



# Petunjuk

- Waktu presentasi adalah 5 menit (tentatif, tergantung dari banyaknya kelompok yang mendaftarkan diri)
- Waktu tanya jawab adalah 5 menit
- Silakan menambahkan gambar/visualisasi pada slide presentasi
- Upayakan agar tetap dalam format poin-poin (ingat, ini presentasi, bukan esai)
- Jangan masukkan *code* ke dalam slide presentasi (tidak usah memasukan screenshot jupyter notebook)

1. Latar Belakang
2. Explorasi Data dan Visualisasi
3. Modelling
4. Kesimpulan

# Latar Belakang

# Latar Belakang Project

Sumber Data: <https://www.kaggle.com/datasets/barun2104/telecom-churn?datasetId=567482>

Problem: **classification**

Tujuan:

- Melakukan prediksi dataset Customer Churn berdasarkan AccountWeeks, ContractRenewal, DataPlan, DataUsage, CustServCalls, DayMins, DayCalls, MonthlyCHarge, OverageFee, dan RoamMins.
- Menentukan Model terbaik untuk dataset Customer Churn

# Explorasi Data dan Visualisasi

# Business Understanding

Customer churn didefinisikan sebagai kecenderungan pelanggan untuk berhenti melakukan interaksi dengan sebuah perusahaan. Perusahaan telekomunikasi memiliki kebutuhan untuk mengetahui apakah pelanggan akan berhenti berlangganan atau tidak, karena biaya untuk mempertahankan pelanggan yang sudah ada jauh lebih sedikit dibandingkan memperoleh pelanggan baru.

Pelanggan yang berhenti karena sebab eksternal seperti berpindah lokasi, kematian, atau alasan lainnya yang tidak sulit untuk dilakukan karena kita dapat mempelajari karakteristik pelanggan yang dapat dilihat dari profil pelanggan. Permasalahan ini dapat dijawab dengan membuat sebuah model Machine Learning yang dapat memprediksi apakah seorang pelanggan akan churn atau tidak. Harapannya, dengan adanya model ini, pihak perusahaan telekomunikasi dapat melakukan tindak preventif bagi pelanggan yang berpeluang besar untuk churn.

# Data Cleansing

Dari dataset yang diperoleh, didapatkan informasi bahwa terdapat 3333 baris dan 11 kolom dengan 10 independent variables dan 1 dependent variable

Untuk Independent variable terdiri dari kolom AccountWeeks, ContractRenewal, DataPlan, DataUsage, CustServCalls, DayMins, DayCalls, MonthlyCharge, OverageFee, dan RoamMins. Sedangkan dependent terdiri dari kolom Churn.



# Data Cleansing

	Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins	DayCalls	MonthlyCharge	OverageFee	RoamMins
0	0	128	1	1	2.70	1	265.1	110	89.0	9.87	10.0
1	0	107	1	1	3.70	1	161.6	123	82.0	9.78	13.7
2	0	137	1	0	0.00	0	243.4	114	52.0	6.06	12.2
3	0	84	0	0	0.00	2	299.4	71	57.0	3.10	6.6
4	0	75	0	0	0.00	3	166.7	113	41.0	7.42	10.1
...	...	...	...	...	...	...	...	...	...	...	...
3328	0	192	1	1	2.67	2	156.2	77	71.7	10.78	9.9
3329	0	68	1	0	0.34	3	231.1	57	56.4	7.67	9.6
3330	0	28	1	0	0.00	2	180.8	109	56.0	14.44	14.1
3331	0	184	0	0	0.00	2	213.8	105	50.0	7.98	5.0
3332	0	74	1	1	3.70	0	234.4	113	100.0	13.30	13.7

3333 rows x 11 columns

# Data Cleansing

```
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
```

Dari informasi data bisa dilihat ada data dengan non-null. Data terdiri dari 2 tipe data, yaitu :

1. Tipe data int64 terdiri dari 6 kolom
2. Tipe data float64 terdiri dari 5 kolom

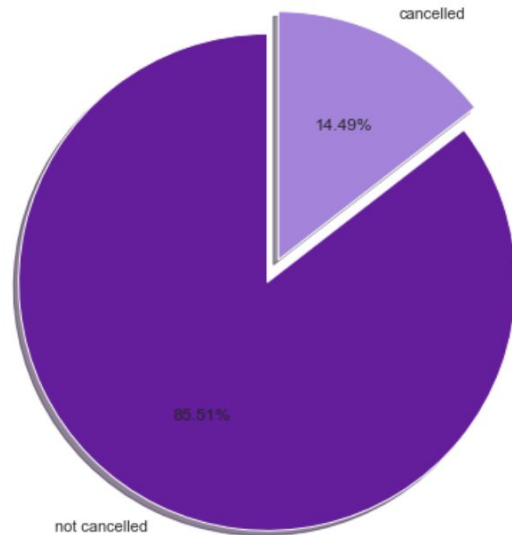
# Exploratory Data Analysis



Heatmap dari dataset Customer Churn bisa dilihat hubungan antar kolom pada gambar tersebut.

# Exploratory Data Analysis

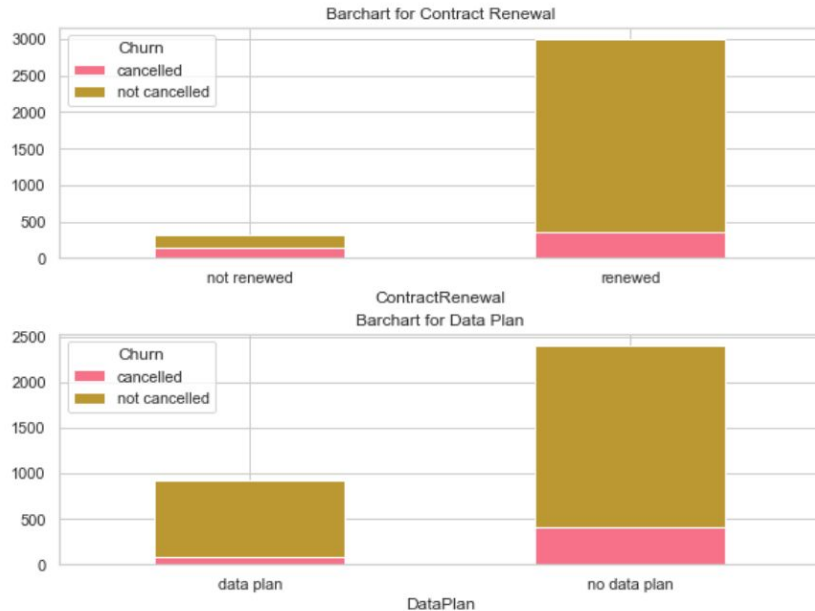
Pie Chart of Column Churn



Dapat dilihat pada Pie Chart of Column Churn terdapat 14.49% yang melakukan cancelled dan 85.51% yang tidak melakukan cancelled

Churn	
not cancelled	2850
cancelled	483

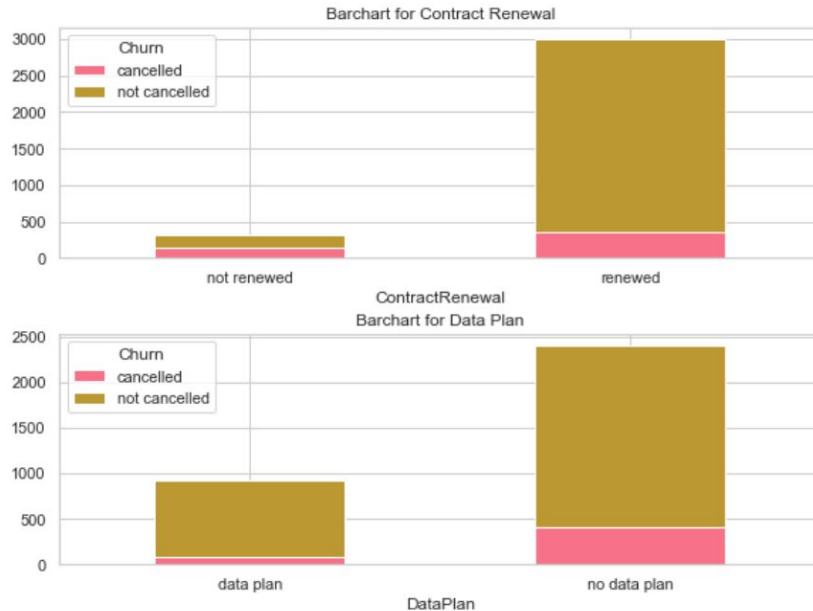
# Exploratory Data Analysis



1. Binary Features; 'ContractRenewal' terhadap 'Churn'

Bisa dilihat lebih banyak pelanggan memperpanjang kontrak daripada yang tidak memperpanjang kontrak, dan yang melakukan cancelled juga lebih banyak daripada pelanggan yang tidak memperpanjang kontrak

# Exploratory Data Analysis



## 2. Binary Features; 'DataPlan' terhadap 'Churn'

Bisa dilihat lebih banyak pelanggan tidak memiliki dataplan daripada yang memiliki dataplan, dan di no plan data lebih banyak daripada yang lainnya untuk churn.

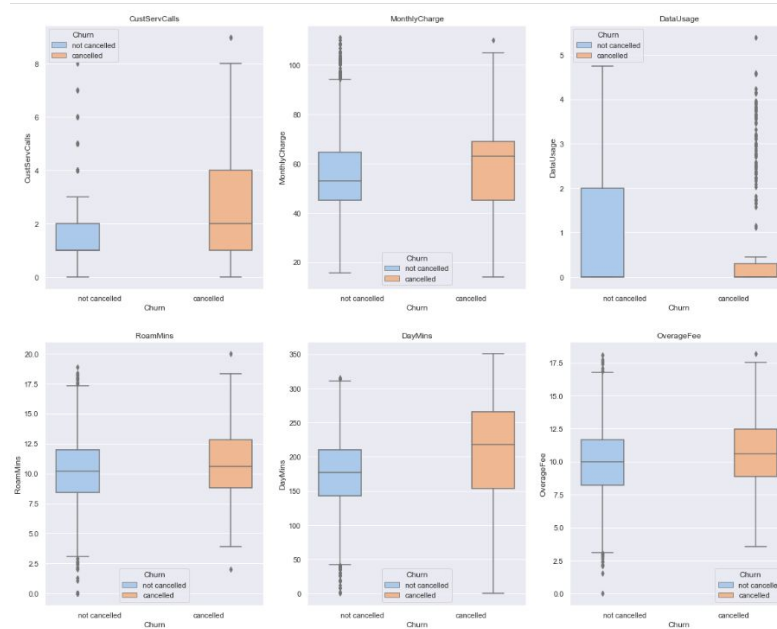
# Exploratory Data Analysis

Feature	Mean
Churn	0.144914
AccountWeeks	101.064806
ContractRenewal	0.903090
DataPlan	0.276628
DataUsage	0.816475
CustServCalls	1.562856
DayMins	179.775098
DayCalls	100.435644
MonthlyCharge	56.305161
OverageFee	10.051488
RoamMins	10.237294

Feature	Mean
Churn	0.000000
AccountWeeks	100.793684
ContractRenewal	0.934737
DataPlan	0.295439
DataUsage	0.862151
CustServCalls	1.449825
DayMins	175.175754
DayCalls	100.283158
MonthlyCharge	55.816246
OverageFee	9.954618
RoamMins	10.158877

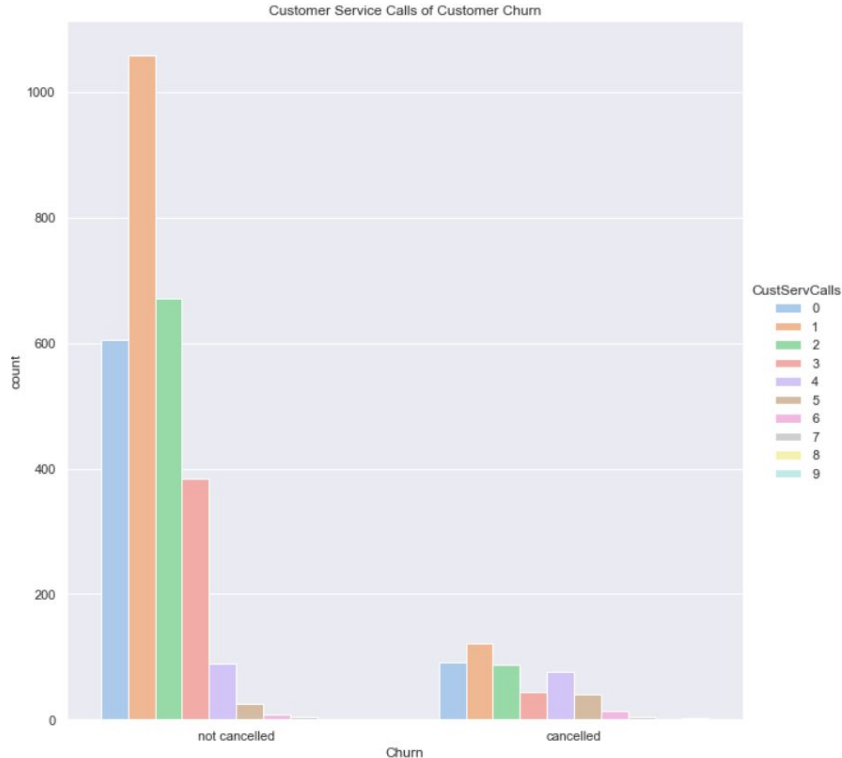
Disamping merupakan data rata-rata panggilan yang dilakukan oleh pelanggan.

# Exploratory Data Analysis



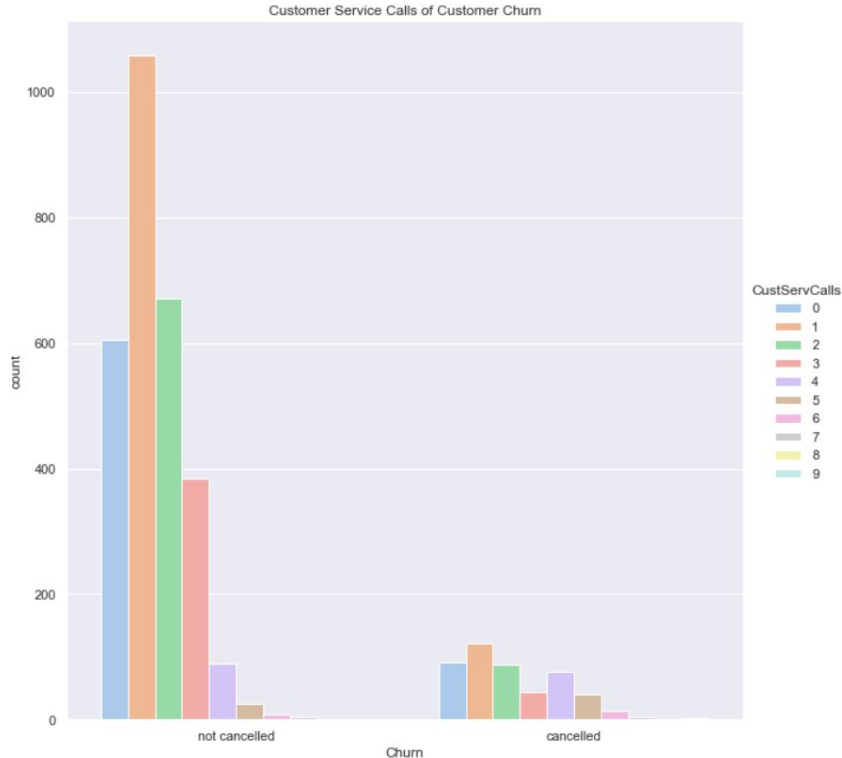


# Exploratory Data Analysis



Berdasarkan grafik Customer Service Calls of Customer Churn dapat ditarik kesimpulan pelanggan yang not cancelled service melakukan panggilan 1 sampai 3 kali dan pelanggan yang melakukan cancelled service melakukan panggilan 4 kali jika dibandingkan dengan 3 kali panggilan, dan pelanggan yang melakukan panggilan 1 kali termasuk cancelled dan not cancelled service.

# Exploratory Data Analysis



Berdasarkan grafik Customer Service Calls of Customer Churn dapat ditarik kesimpulan pelanggan yang not cancelled service melakukan panggilan 1 sampai 3 kali dan pelanggan yang melakukan cancelled service melakukan panggilan 4 kali jika dibandingkan dengan 3 kali panggilan, dan pelanggan yang melakukan panggilan 1 kali termasuk cancelled dan not cancelled service.

# Modelling

# Normalisasi Data

Split data for train and test dengan menggunakan data testing sebanyak 20% dan Data Training Sebanyak 80%

# Normalisasi Data

- Melakukan Resampling Data

```
Downsampled Dataset Class Counts:
```

```
0      386
```

```
1      386
```

```
Name: Churn, dtype: int64
```

```
Upsampled Dataset Class Counts:
```

```
1     2280
```

```
0     2280
```

```
Name: Churn, dtype: int64
```

# MODELLING

## - Hasil Tuning Parameter Random Forest

### IMBALANCED CLASSES

Best roc\_auc-score during GS: 0.9030325668483563

Best params from GS:

```
{'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 444}
```

Accuracy Score: 0.9565217391304348

AUROC Score: 0.9301862904684391

### DOWNSAMPLED MAJORITY CLASS

Best roc\_auc-score during GS: 0.8853929620163387

Best params from GS:

```
{'max_depth': 11, 'min_samples_leaf': 4, 'min_samples_split': 6, 'n_estimators': 370}
```

Accuracy Score: 0.8875562218890555

AUROC Score: 0.9195876288659793

### UPSAMPLED MINORITY CLASS

# MODELLING

## - Hasil Tuning Parameter Decision Tree

### IMBALANCED CLASSES

Best roc\_auc-score during GS: 0.8657849825612984

Best params from GS:

```
{'max_depth': 6, 'min_samples_leaf': 4, 'min_samples_split': 5}
```

Accuracy Score: 0.9415292353823088

AUROC Score: 0.921287755471152

### DOWNSAMPLED MAJORITY CLASS

Best roc\_auc-score during GS: 0.8449924534340119

Best params from GS:

```
{'max_depth': 6, 'min_samples_leaf': 4, 'min_samples_split': 5}
```

Accuracy Score: 0.8860569715142429

AUROC Score: 0.8886507505878097

### UPSAMPLED MINORITY CLASS

Best roc\_auc-score during GS: 0.9640322022160663

Best params from GS:

```
{'max_depth': 16, 'min_samples_leaf': 2, 'min_samples_split': 7}
```

Accuracy Score: 0.8725637181409296

AUROC Score: 0.820898896726352

# MODELLING

## - Hasil Tuning Parameter Logistic Regression

### IMBALANCED CLASSES

Best roc\_auc-score during GS: 0.8096941800889169

Best params from GS:

```
{'max_iter': 630, 'penalty': 'l2', 'tol': 0.00018661761457749353}
```

Accuracy Score: 0.8650674662668666

AUROC Score: 0.8317236389943932

### DOWNSAMPLED MAJORITY CLASS

Best roc\_auc-score during GS: 0.8140980231889323

Best params from GS:

```
{'max_iter': 630, 'penalty': 'l2', 'tol': 0.00018661761457749353}
```

Accuracy Score: 0.7586206896551724

AUROC Score: 0.8280340025321035

### UPSAMPLED MINORITY CLASS

Best roc\_auc-score during GS: 0.8142255309325946

Best params from GS:

```
{'max_iter': 630, 'penalty': 'l2', 'tol': 0.00018661761457749353}
```

Accuracy Score: 0.7646176911544228

AUROC Score: 0.8348164224995478



# MODELLING

RandomForestClassifier Confusion Matrix:

```
[[559  11]  
 [ 23  74]]
```

Precision: 0.980701754385965

Recall: 0.9604810996563574

Accuracy: 0.9490254872563718

DecisionTreeClassifier Confusion Matrix:

```
[[527  43]  
 [ 30  67]]
```

Precision: 0.9245614035087719

Recall: 0.9461400359066428

Accuracy: 0.8905547226386806

LogisticRegression Confusion Matrix:

```
[[436 134]  
 [ 27  70]]
```

Precision: 0.7649122807017544

Recall: 0.9416846652267818

Accuracy: 0.7586206896551724

Dapat dilihat hasil dari Precision, Recall, dan Accuracy yang paling besar adalah **Random Forest**.

# Conclusion

# Conclusion

Dari modelling yang telah dilakukan terhadap dataset Customer Churn, dapat ditarik bahwa Random Forest yang paling baik untuk Model dataset tersebut. Dengan adanya model untuk memprediksi customer churn, pihak perusahaan telekomunikasi dapat dengan mudah mengetahui pelanggan mana yang memiliki kecenderungan untuk churn. Dari sini, pihak marketing dapat melakukan promosi produk dengan sifat kontrak yang jangkanya lebih panjang sehingga kedua pelanggan ini dapat bertahan lebih lama.

# Terima kasih!

Ada pertanyaan?

zenius



Kampus  
Merdeka  
INDONESIA JAYA