BINAR

Hardianto Tandi Seno - Challenge Chapter II

Data Science

Studi Kasus :

Banyaknya perusahaan telekomunikasi membuat pelanggan memiliki hak untuk dapat menentukan provider yang sesuai dengan preferensi sehingga memungkinkan terjadinya peralihan dari satu provider ke provider yang lain (Customer Churn).

Penyebab yang dapat ditimbulkan dari customer churn ini dapat menyebabkan berkurangnya pendapatan bagi perusahaan telekomunikasi sehingga penting untuk dapat ditangani

Hal ini membuat kami sebagai junior data scientist di suatu perusahaan telekomunikasi diminta untuk **melakukan prediksi customer churn** berdasarkan history data yang dimiliki oleh perusahaan tersebut

Dataset yang dimiliki perusahaan terdiri dari beberapa kolom, diantaranya :

- **state**: Negara bagian di mana pelanggan tinggal.
- **account_length**: Lamanya waktu (dalam hari atau bulan) pelanggan telah dikaitkan dengan penyedia layanan.
- **area_code**: Kode area telepon yang terkait dengan nomor telepon pelanggan.
- **international_plan**: Indikator biner (ya/tidak) yang menunjukkan apakah pelanggan memiliki paket panggilan internasional.
- **voice_mail_plan**: Indikator biner (ya/tidak) yang menunjukkan apakah pelanggan memiliki paket pesan suara.
- **number_vmail_messages**: Jumlah pesan suara yang diterima pelanggan.
- **total_day_minutes**: Jumlah total menit yang digunakan pelanggan selama panggilan siang hari.
- **total_day_calls**: Jumlah total panggilan yang dilakukan pelanggan pada siang hari.
- **total_day_charge**: Total biaya yang ditagihkan ke pelanggan untuk panggilan siang hari.
- **total_eve_minutes**: Jumlah total menit yang digunakan pelanggan selama panggilan malam.
- **total_eve_calls**: Jumlah total panggilan yang dilakukan pelanggan pada malam hari.
- **total_eve_charge**: Total biaya yang ditagihkan ke pelanggan untuk panggilan malam hari.
- **total_night_minutes**: Jumlah total menit yang digunakan pelanggan selama panggilan malam hari.
- **total_night_calls**: Jumlah total panggilan yang dilakukan pelanggan pada malam hari.
- **total_night_charge**: Total biaya yang ditagihkan ke pelanggan untuk panggilan malam hari.
- **total_intl_minutes**: Jumlah total menit internasional yang digunakan oleh pelanggan.
- **total_intl_calls**: Jumlah total panggilan internasional yang dilakukan oleh pelanggan.
- **total_intl_charge**: Total biaya yang ditagihkan ke pelanggan untuk panggilan internasional.
- **number_customer_service_calls**: Jumlah panggilan layanan pelanggan yang dilakukan oleh pelanggan.
- **churn**: Indikator biner (ya/tidak) yang menunjukkan apakah pelanggan telah melakukan churn (yaitu, berhenti menggunakan layanan).

HASIL INTERPRETASI PREDIKSI CUSTOMER CHURN

Bagian 1

- Melakukan proses import library yang diperlukan (Pandas, Numpy, Matplotlib, Seaborn, scikit-learn, pickle, dll)
- Membaca Train Data dengan menggunakan Pandas

	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
0	OH	107	area_code_415	no	yes	26	161.6	123	27.47	195.5
1	NJ	137	area_code_415	no	no	0	243.4	114	41.38	121.2
2	OH	84	area_code_408	yes	no	0	299.4	71	50.90	61.9
3	OK	75	area_code_415	yes	no	0	166.7	113	28.34	148.3
4	MA	121	area_code_510	no	yes	24	218.2	88	37.09	348.5

Hasil read Train data dengan Pandas (Beberapa kolom awal)

Bagian 2

Melakukan beberapa proses analisis statistika dan EDA, diantaranya :

```
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
(4250, 20)
```

Melihat adanya missing value dan shape pada data

Bagian 2

```
data[data.duplicated()]
```

state	account_length	area_code	international_plan	voice_mail_plan	number_vmail_messages	total_day_minutes	total_day_calls

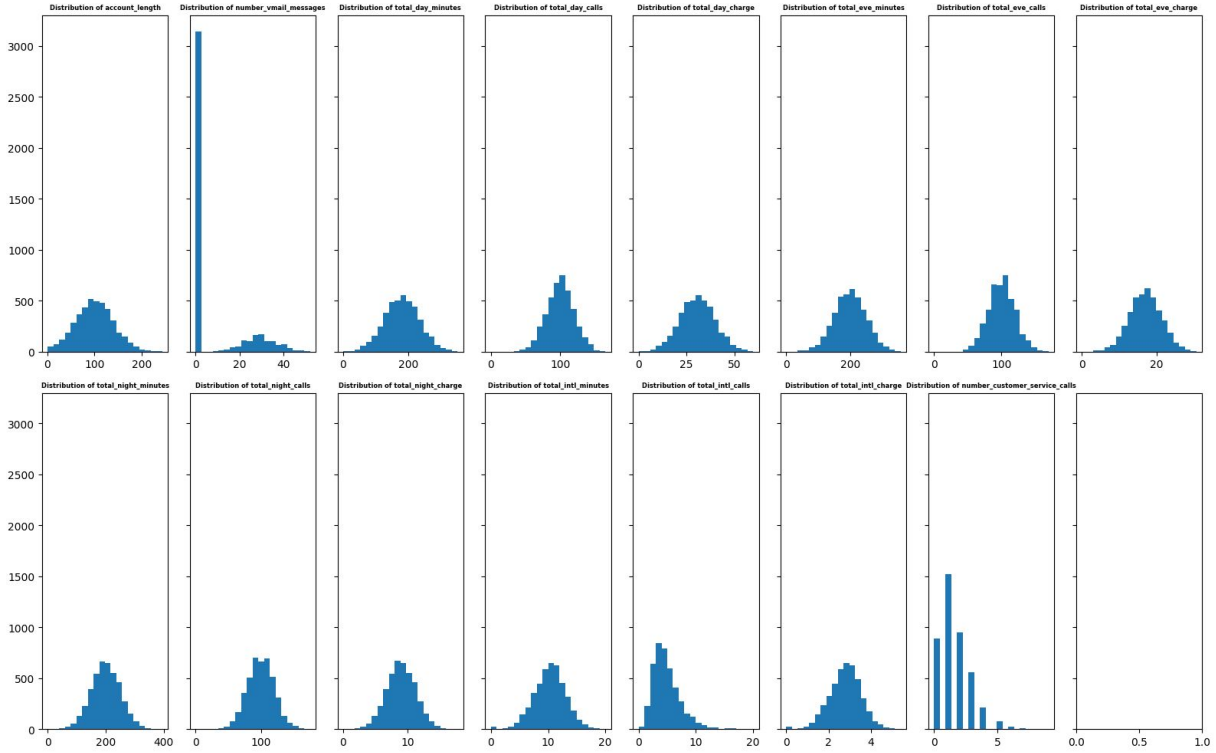
Melihat kemungkinan adanya data yang terduplikat

```
data.describe()
```

	account_length	number_vmail_messages	total_day_minutes	total_day_calls	total_day_charge	total_eve_minutes	total_eve_calls	total_eve_charge	total_night_minutes
count	4250.000000	4250.000000	4250.000000	4250.000000	4250.000000	4250.000000	4250.000000	4250.000000	4250.000000
mean	100.236235	7.631765	180.259600	99.907294	30.644682	200.173906	100.176471	17.015012	200.527882
std	39.698401	13.439882	54.012373	19.850817	9.182096	50.249518	19.908591	4.271212	50.353548
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	73.000000	0.000000	143.325000	87.000000	24.365000	165.925000	87.000000	14.102500	167.225000
50%	100.000000	0.000000	180.450000	100.000000	30.680000	200.700000	100.000000	17.060000	200.450000
75%	127.000000	16.000000	216.200000	113.000000	36.750000	233.775000	114.000000	19.867500	234.700000
max	243.000000	52.000000	351.500000	165.000000	59.760000	359.300000	170.000000	30.540000	395.000000

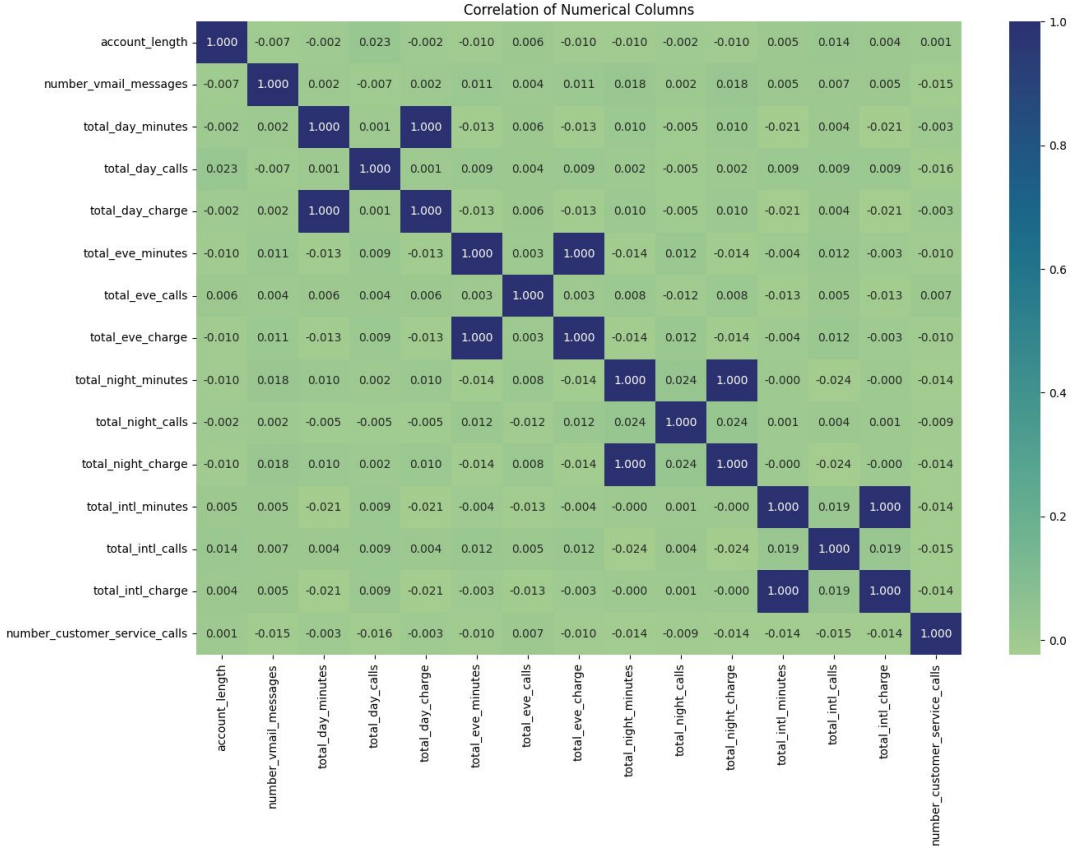
Melihat statistik deskriptif pada data numerik untuk mengetahui adanya indikasi outlier atau tidak pada data

Bagian 2



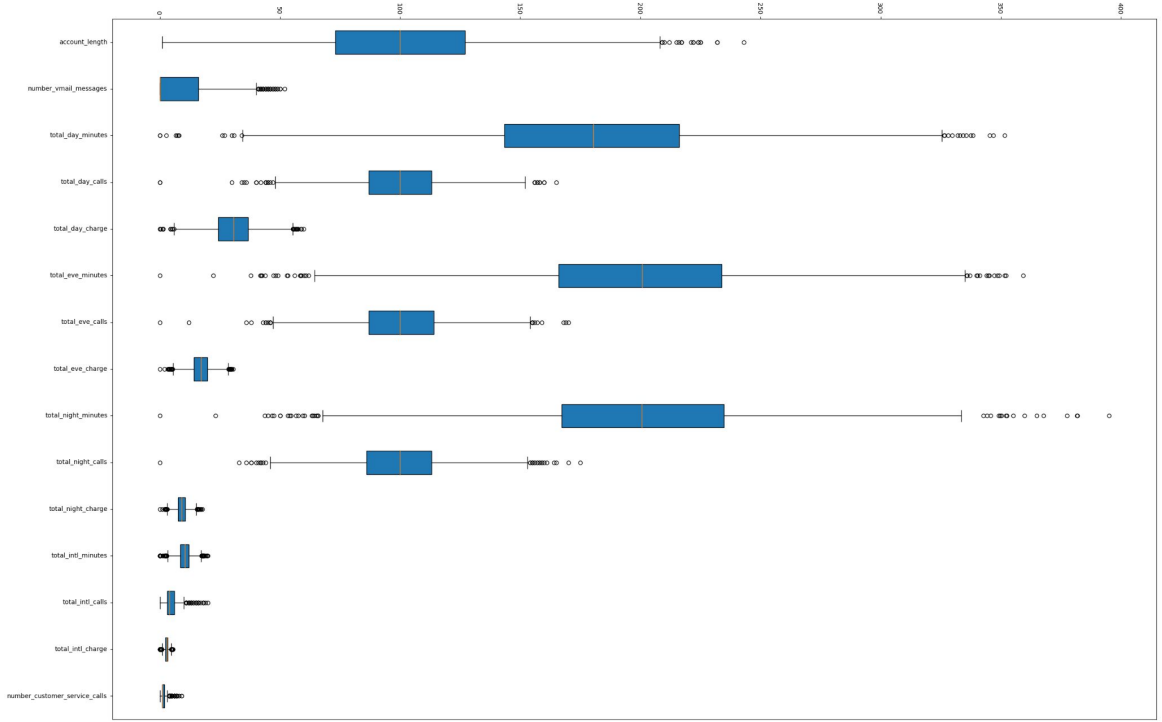
Melihat distribusi penyebaran data numerik. Bisa terlihat bahwa rata-rata kolomnya itu sudah hampir terdistribusi normal. Hanya saja terdapat beberapa kolom yang distribusi datanya tidak seimbang (skewed), seperti number_vmail_messages, total_intl_calls, number_customer_service_calls

Bagian 2

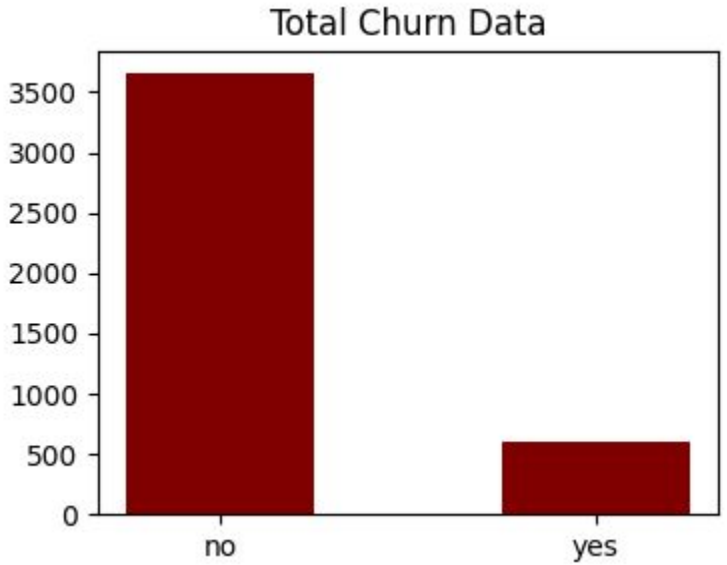


Melihat korelasi antar kolom pada data numerik. Hasil dari pengecekan korelasi pada tiap kolom dengan menggunakan heatmap menunjukkan bahwa terdapat beberapa kolom yang memiliki hubungan korelasi (munculnya pola yang sama pada waktu yang bersamaan) yang positif sempurna (bernilai 1).

Bagian 2



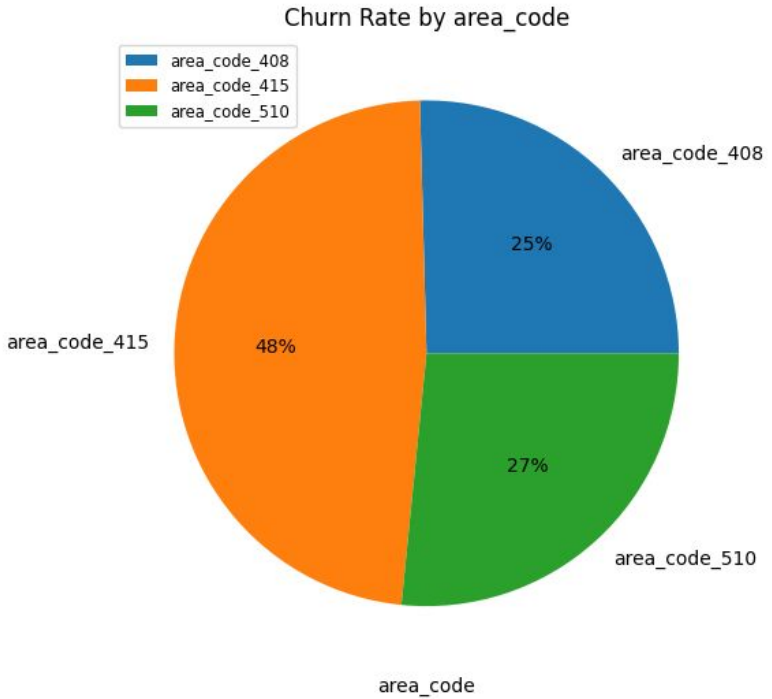
Melihat potensi outlier pada kolom data numerik dengan menggunakan boxplot. Hasilnya menunjukkan bahwa terdapat beberapa kolom yang memiliki outlier, sehingga akan dihapus dengan menggunakan metode z-score



Melihat perbandingan data terkait klasifikasi churn atau tidak pada kolom target. Datanya bersifat imbalanced (tidak seimbang).

Bagian 2

Kolom target ini akan dilakukan proses label encoding untuk dapat memudahkan analisis lebih lanjut. Hasil analisis terkait dengan tingkat churn pada customer adalah sebagai berikut :



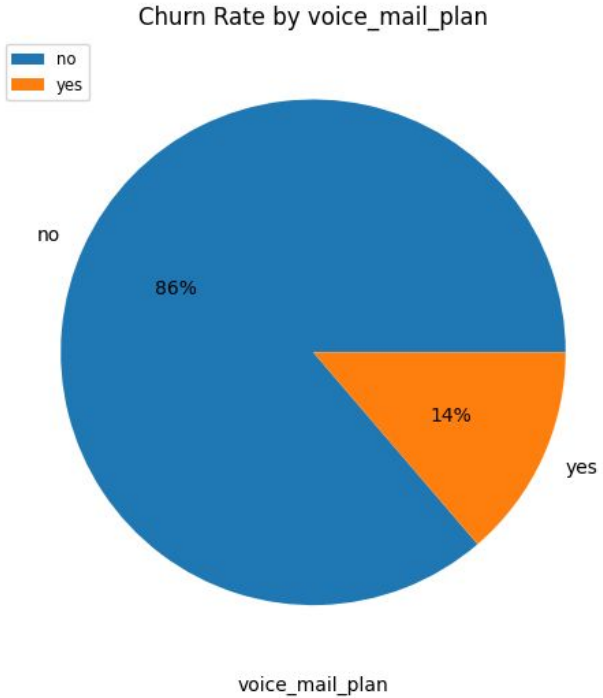
Pada grafik disamping, terlihat bahwa area_code_415 memiliki tingkat churn yang tinggi dibandingkan dengan 2 area_code lainnya.

Hal ini membuat perusahaan dapat mengidentifikasi faktor-faktor unik yang mempengaruhi perilaku pelanggan di daerah tersebut.

Selain itu, bisa dilakukan analisis lanjutan untuk faktor-faktor tertentu seperti kualitas layanan, harga, atau preferensi pelanggan yang berperan

Bagian 2

Kolom target ini akan dilakukan proses label encoding untuk dapat memudahkan analisis lebih lanjut. Hasil analisis terkait dengan tingkat churn pada customer adalah sebagai berikut :

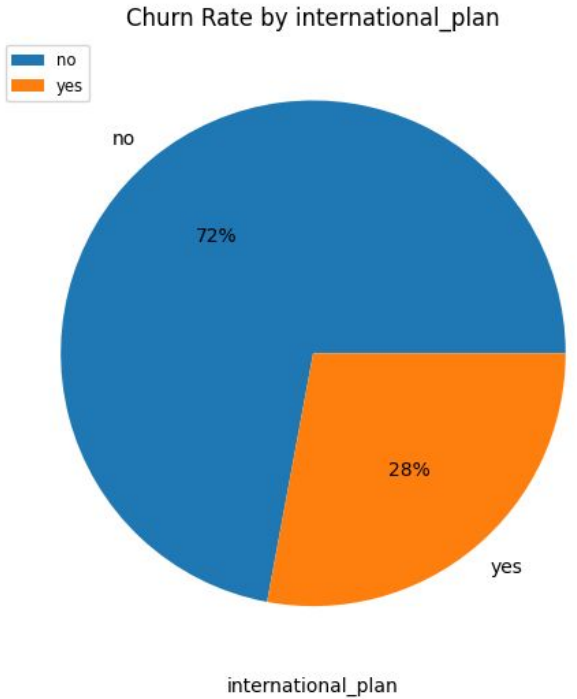


Pada grafik disamping, terlihat bahwa customer yang tidak mengaktifkan voice_main_plan memiliki tingkat churn yang sangat tinggi dibandingkan dengan yang mengaktifkan voice_mail_plan-nya

Penyebab hal ini terjadi bisa karena customer merasa kurang puas dengan layanan tersebut atau mungkin mereka tidak menggunakan fitur-fitur lain yang disediakan oleh layanan tersebut. Ini bisa menjadi indikator bahwa perusahaan perlu memperbaiki atau meningkatkan layanan voice mail plan agar customer bisa bertahan.

Bagian 2

Kolom target ini akan dilakukan proses label encoding untuk dapat memudahkan analisis lebih lanjut. Hasil analisis terkait dengan tingkat churn pada customer adalah sebagai berikut :

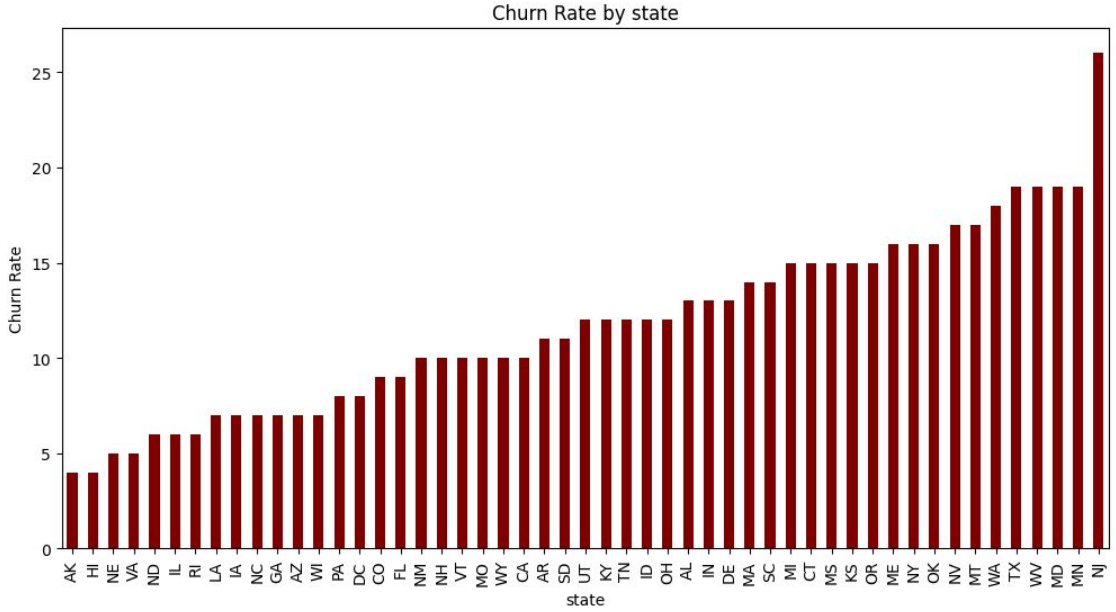


Pada grafik disamping, terlihat bahwa customer yang tidak mengaktifkan international_plan memiliki tingkat churn yang sangat tinggi dibandingkan dengan yang mengaktifkan international_plan-nya

Penyebab hal ini terjadi bisa jadi sama seperti kasus voice_mail_plan. Sehingga, perusahaan perlu memperbaiki atau meningkatkan international plan agar customer bisa bertahan.

Bagian 2

Kolom target ini akan dilakukan proses label encoding untuk dapat memudahkan analisis lebih lanjut. Hasil analisis terkait dengan tingkat churn pada customer adalah sebagai berikut :



State AK (Alaska), menjadi state dengan tingkat churn terendah. Sedangkan NJ (New Jersey) menjadi state dengan tingkat churn tertinggi

Hal ini bisa disebabkan oleh kualitas provider di tiap daerah itu berbeda, persaingan harga yang cukup ketat sehingga memungkinkan customer untuk berganti provider sesuai dengan layanan yang sesuai dengan keinginan, dan lain-lain

Bagian 3

Melakukan penambahan kolom fitur pada data berupa total menit customer per hari, total panggilan yang dilakukan, dan total biaya yang dikeluarkan customer

total_minutes	total_calls	total_charges
611.5	329	55.54
527.2	328	59.00
558.2	248	65.02
501.9	356	49.36
779.3	314	76.28

Bagian 3

Melakukan tes hipotesis pada beberapa kolom fitur terhadap kolom target pada data, dan berikut hasilnya :

```
Test statistic: -13.439870415300485
p-value: 2.332712336238055e-40
Reject the null hypothesis. There is a significant difference in average total_minutes between churned and non-churned customers.
```

Ini dapat diinterpretasikan sebagai indikasi bahwa variabel total menit (total_minutes) mungkin memiliki pengaruh atau peran yang signifikan dalam memprediksi perilaku churn pelanggan.

```
Test statistic: 0.3027328330651385
p-value: 0.7621082965640166
Fail to reject the null hypothesis. There is no significant difference in average total_calls between churned and non-churned customers.
```

Ini dapat diinterpretasikan sebagai indikasi bahwa variabel total panggilan (total_call) tidak memiliki pengaruh atau peran yang signifikan dalam memprediksi perilaku churn pelanggan.

```
Test statistic: -15.69270961944851
p-value: 5.424598212128414e-54
Reject the null hypothesis. There is a significant difference in average total_charges between churned and non-churned customers.
```

Ini dapat diinterpretasikan sebagai indikasi bahwa variabel total biaya (total_charges) memiliki pengaruh atau peran yang signifikan dalam memprediksi perilaku churn pelanggan.

Bagian 4

Melakukan beberapa tahapan data preprocessing sebelum masuk ke pemodelan :

```
Length rows before remove outlier: 4250
Length rows after remove outlier: 4031
```

1. Menghilangkan outlier pada data
2. Melakukan encoding pada categorical data (Label Encoding & One-Hot Encoding)
3. Melakukan Feature Scaling menggunakan Normalization (MinMaxScaler)

Detail penerapannya dapat dilihat pada file .ipynb

Bagian 5

Melakukan proses pemodelan dengan 3 model, yaitu Logistic Regression, KNN, dan Decision Tree. Dalam proses pemodelan ini dibagi menjadi beberapa pengujian, diantaranya :

Penerapan langsung tanpa adanya parameter dan metode tambahan :

```
Training Logistic Regression...
Evaluation metrics for Logistic Regression:
Accuracy: 0.8682
Precision: 0.6735
Recall: 0.2558
F1 Score: 0.3708
              precision    recall  f1-score   support

           0       0.88      0.98      0.93       721
           1       0.67      0.26      0.37       129

    accuracy                           0.87       850
   macro avg       0.78      0.62      0.65       850
weighted avg       0.85      0.87      0.84       850
```

```
Training K-NN...
Evaluation metrics for K-NN:
Accuracy: 0.8541
Precision: 0.5641
Recall: 0.1705
F1 Score: 0.2619
              precision    recall  f1-score   support

           0       0.87      0.98      0.92       721
           1       0.56      0.17      0.26       129

    accuracy                           0.85       850
   macro avg       0.72      0.57      0.59       850
weighted avg       0.82      0.85      0.82       850
```

```
Training DecisionTreeClassifier...
Evaluation metrics for DecisionTreeClassifier:
Accuracy: 0.9482
Precision: 0.8058
Recall: 0.8682
F1 Score: 0.8358
              precision    recall  f1-score   support

           0       0.98      0.96      0.97       721
           1       0.81      0.87      0.84       129

    accuracy                           0.95       850
   macro avg       0.89      0.92      0.90       850
weighted avg       0.95      0.95      0.95       850
```

Terlihat bahwa model Decision Tree mendapatkan nilai Accuracy, Precision, Recall, dan F-1 Score yang lebih baik dan stabil dari 2 model lainnya.

Sedangkan untuk model Logistic Regression dan KNN, nilai Precision, Recall, dan F-1 Scorenya terbilang rendah

Bagian 5

Melakukan proses pemodelan dengan 3 model, yaitu Logistic Regression, KNN, dan Decision Tree. Dalam proses pemodelan ini dibagi menjadi beberapa pengujian, diantaranya :

Penerapan metode SMOTE untuk imbalancing data :

Logistic Regression

	precision	recall	f1-score	support
0	0.94	0.78	0.85	721
1	0.37	0.71	0.49	129
accuracy			0.77	850
macro avg	0.65	0.75	0.67	850
weighted avg	0.85	0.77	0.80	850

KNN

	precision	recall	f1-score	support
0	0.92	0.75	0.83	721
1	0.31	0.61	0.41	129
accuracy			0.73	850
macro avg	0.61	0.68	0.62	850
weighted avg	0.82	0.73	0.76	850

Metode ini hanya diaplikasikan untuk 2 model saja terkait dengan pengujian yang pertama.

Hasilnya menunjukkan bahwa nilai Precision & Accuracy pada kedua model mengalami penurunan, namun nilai Recall & F-1 Score nya mengalami peningkatan meskipun nilainya masih terbilang cukup rendah.

Bagian 5

Melakukan proses pemodelan dengan 3 model, yaitu Logistic Regression, KNN, dan Decision Tree. Dalam proses pemodelan ini dibagi menjadi beberapa pengujian, diantaranya :

Penerapan metode GridSearchCV :

```
Best parameters for Logistic Regression: {'C': 10}
Best score for Logistic Regression: 0.8697058823529412
              precision    recall  f1-score   support

           0       0.88      0.98      0.93       721
           1       0.67      0.26      0.38       129

    accuracy                           0.87       850
   macro avg       0.77      0.62      0.65       850
weighted avg       0.85      0.87      0.84       850
```

```
Best parameters for K-NN: {'n_neighbors': 9}
Best score for K-NN: 0.8791176470588236
              precision    recall  f1-score   support

           0       0.87      0.99      0.92       721
           1       0.67      0.16      0.25       129

    accuracy                           0.86       850
   macro avg       0.77      0.57      0.59       850
weighted avg       0.84      0.86      0.82       850
```

```
Best parameters for DecisionTreeClassifier: {'max_depth': 5}
Best score for DecisionTreeClassifier: 0.9708823529411765
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       721
           1       0.96      0.84      0.90       129

    accuracy                           0.97       850
   macro avg       0.97      0.92      0.94       850
weighted avg       0.97      0.97      0.97       850
```

Penerapan GridSearchCV ini dilakukan untuk mencari model terbaik berdasarkan hyperparameter yang digunakan dalam model tersebut.

Hasilnya menunjukkan bahwa model Logistic Regression & KNN tidak mengalami peningkatan yang berarti.

Sedangkan model Decision Tree peningkatannya cukup terasa jika nilainya dibandingkan dengan nilai hasil model tanpa parameter dan metode tertentu.

Bagian 5

Melakukan proses pemodelan dengan 3 model, yaitu Logistic Regression, KNN, dan Decision Tree. Dalam proses pemodelan ini dibagi menjadi beberapa pengujian, diantaranya :

Penerapan metode PCA :

```
Training Logistic Regression...
Evaluation metrics for Logistic Regression:
Accuracy: 0.8674
Precision: 0.6471
Recall: 0.2115
F1 Score: 0.3188
              precision    recall  f1-score   support

           0       0.88      0.98      0.93       907
           1       0.65      0.21      0.32       156

    accuracy                           0.87      1063
   macro avg       0.76      0.60      0.62      1063
weighted avg       0.84      0.87      0.84      1063
```

```
Training K-NN...
Evaluation metrics for K-NN:
Accuracy: 0.8598
Precision: 0.5686
Recall: 0.1859
F1 Score: 0.2802
              precision    recall  f1-score   support

           0       0.87      0.98      0.92       907
           1       0.57      0.19      0.28       156

    accuracy                           0.86      1063
   macro avg       0.72      0.58      0.60      1063
weighted avg       0.83      0.86      0.83      1063
```

```
Training DecisionTreeClassifier...
Evaluation metrics for DecisionTreeClassifier:
Accuracy: 0.9078
Precision: 0.6747
Recall: 0.7179
F1 Score: 0.6957
              precision    recall  f1-score   support

           0       0.95      0.94      0.95       907
           1       0.67      0.72      0.70       156

    accuracy                           0.91      1063
   macro avg       0.81      0.83      0.82      1063
weighted avg       0.91      0.91      0.91      1063
```

Penerapan PCA ini dilakukan untuk mereduksi dimensi dari data yang mengalami peningkatan akibat proses one-hot encoder yang telah dilakukan sebelumnya.

Hasilnya menunjukkan bahwa model Logistic Regression & KNN terdapat beberapa penurunan metric.

Sedangkan model Decision Tree penurunannya cukup terasa jika nilainya dibandingkan dengan nilai hasil model tanpa parameter dan metode tertentu.

Bagian 5

```
Best parameters for DecisionTreeClassifier: {'max_depth': 5}
Best score for DecisionTreeClassifier: 0.9708823529411765
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       721
           1       0.96      0.84      0.90       129

    accuracy                           0.97       850
   macro avg       0.97      0.92      0.94       850
weighted avg       0.97      0.97      0.97       850
```

Hasil model akhir yang memiliki metrik evaluasi yang baik yaitu menggunakan Decision Tree dengan tambahan metode hyperparameter tuning dengan GridSearchCV

Bagian 6

	Probablilites Prediction	Count
0	0.000000	86
1	0.005263	46
2	0.011561	42
3	0.022467	510
4	0.024793	22
5	0.090909	37
6	0.125000	4
7	0.272727	3
8	0.727273	3
9	0.875000	4
10	0.909091	37
11	0.975207	22
12	0.977533	510
13	0.988439	42
14	0.994737	46
15	1.000000	86

- Melakukan save pada hasil pemodelan (dalam hal ini Decision Tree + GridSearchCV)
- Mengakses test data lalu lakukan flow yang sama seperti pada training data
- Melakukan load model yang telah di save sebelumnya
- Menguji prediksi pada test data

Bagian 6

	Probablilites Prediction	Count
0	0.000000	86
1	0.005263	46
2	0.011561	42
3	0.022467	510
4	0.024793	22
5	0.090909	37
6	0.125000	4
7	0.272727	3
8	0.727273	3
9	0.875000	4
10	0.909091	37
11	0.975207	22
12	0.977533	510
13	0.988439	42
14	0.994737	46
15	1.000000	86

Hasil pengujian pada test data dengan melihat probabilitas hasil prediksi menghasilkan beberapa kesimpulan:

- Model cenderung sangat yakin pada beberapa prediksi (seperti pada probabilitas 0 dan 1)

- Jumlah sampel terbesar terdapat pada probabilitas 0.022467 dan 0.977533, yang mungkin mengindikasikan bahwa model mungkin cukup baik dalam membedakan dua kelas dan memberikan prediksi yang tegas

Alur Selengkapnya dapat dilihat pada file .ipynb yang dikumpulan bersama Presentasi ini

BINAR

Thank You