Final Presentation

Data Science

Dibuat oleh: Rizky Adi Pratama

Challenge 1

BINAR

Table of Content

BINAR

01. Background & Objektif

Dataset Covid-19 di Indonesia dibuat untuk mengetahui berbagai faktor yang dapat menjadi bahan pertimbangan dalam pengambilan suatu keputusan terkait dengan tingkat keketatan setiap provinsi di Indonesia.

Objektif:

1. Berapa jumlah total kasus Covid-19 aktif yang baru paling besar menurut provinsi?
2. Dimana saja provinsi yang memiliki jumlah total kematian karena Covid-19 paling sedikit?
3. Bagaimana gambaran terkait tanggal-tanggal ketika rate kasus recovered di Indonesia paling tinggi?
4. Bagaimana gambaran total case fatality rate dan case recovered rate dari masing-masing lokasi?
5. Kapan saja tanggal-tanggal saat total kasus Covid-19 mulai menyentuh angka 30.000-an?

BINAR

02. Data Understanding

Dataset merupakan informasi terkait COVID-19 di Indonesia sejak Maret 2020 hingga September 2022. yang tersebar di **34 provinsi**. Sebagai tambahan, dataset ini merupakan kumpulan informasi dari berbagai sumber data yang dirangkum menjadi satu dataset COVID-19 Indonesia.

38 Fitur **31.8K Data**

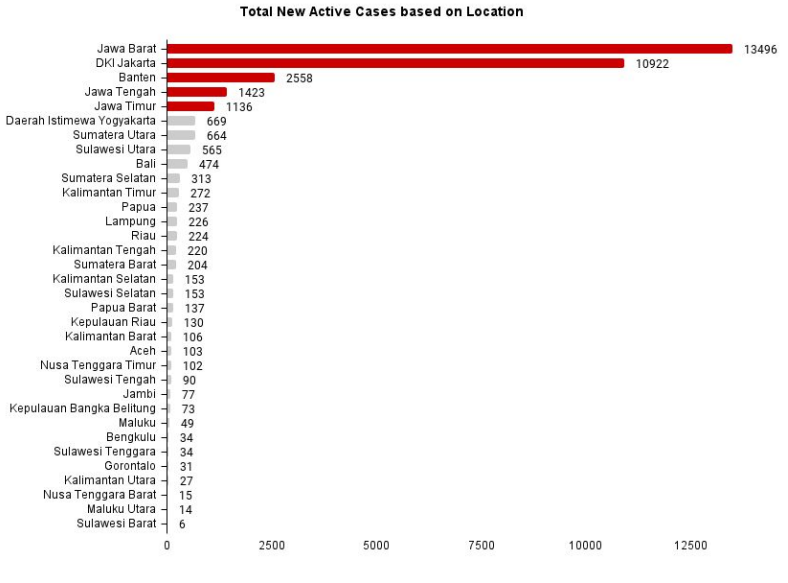
Download data Link

03. Data Analysis

03. Data Analysis

1. Berapa jumlah total kasus Covid-19 aktif yang baru paling besar menurut provinsi?

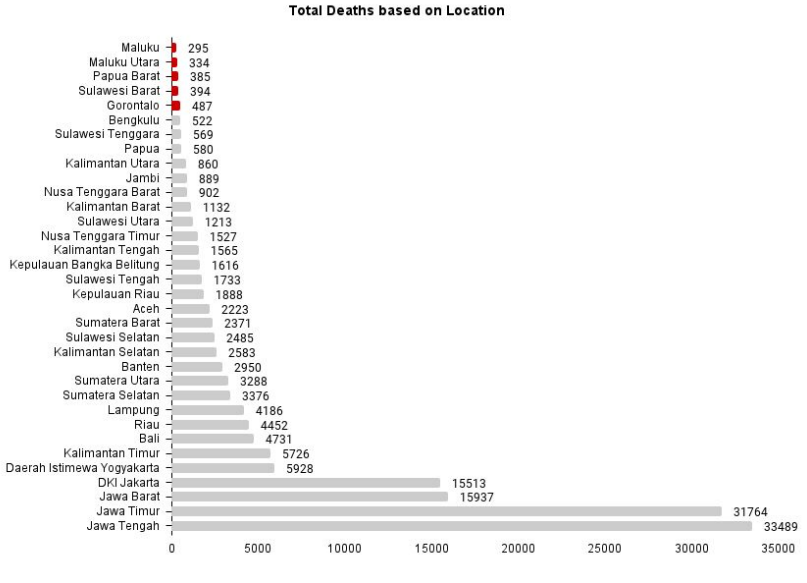
Terlihat bahwa **Jawa Barat** menjadi provinsi dengan total kasus aktif baru tertinggi selama 1 Januari 2020 - 16 September 2022 sebanyak 13.496 kasus. Diikuti **DKI Jakarta** sebanyak 10.922 kasus, dan **Banten** sebanyak 2.558 kasus. Selain itu, provinsi lain yang menyentuh angka di atas 1.000 kasus adalah **Jawa Tengah** dan **Jawa Timur**. Hal ini dapat menjadi insight bahwa pulau Jawa sangat rawan muncul kasus baru tiap harinya yang mungkin diakibatkan karena pulau Jawa merupakan pulau dengan penduduk terpadat di Indonesia.



03. Data Analysis

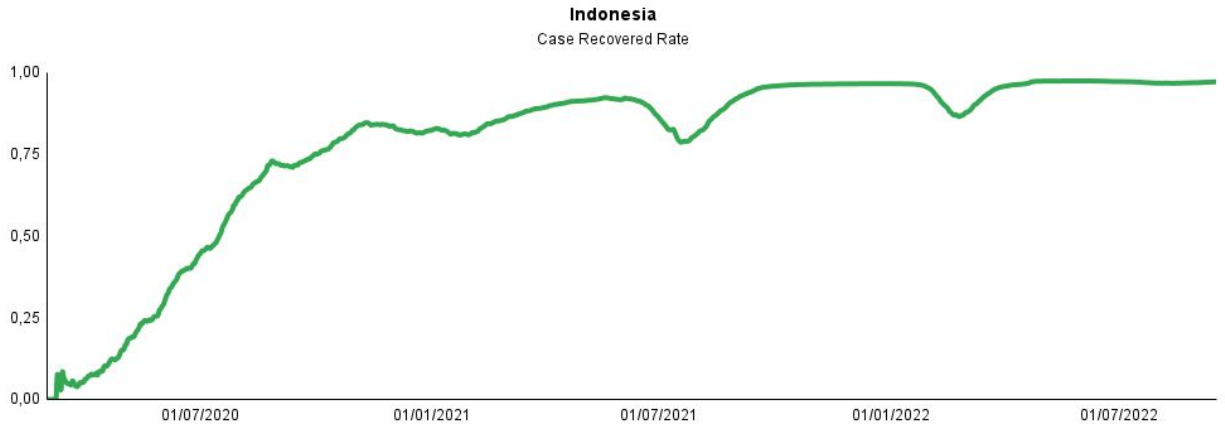
2. Dimana saja provinsi yang memiliki jumlah total kematian karena Covid-19 paling sedikit?

Maluku menjadi provinsi dengan jumlah kematian terendah selama 1 Januari 2020 - 16 September 2022 sebanyak 295 kasus. Diikuti **Maluku Utara** sebanyak 334 kasus, dan **Papua Barat** sebanyak 385 kasus. Perlu dilakukan analisis lebih lanjut untuk melihat faktor-faktor apa saja mengakibatkan provinsi-provinsi tersebut memiliki jumlah kematian yang rendah akibat Covid-19..



03. Data Analysis

3. Bagaimana gambaran terkait tanggal-tanggal ketika rate kasus recovered di Indonesia paling tinggi?

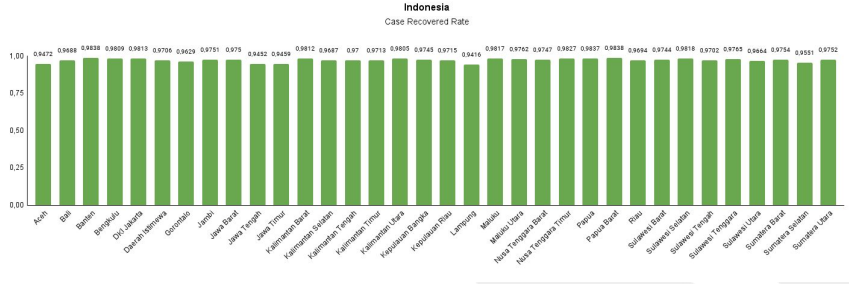


Tingkat kesembuhan akibat Covid-19 di Indonesia mulai stabil di atas 75% sejak tahun 2021. Kondisi ini merupakan dampak dari pelaksanaan vaksinasi Covid-19 yang dilaksanakan pertama kali pada Januari 2021.

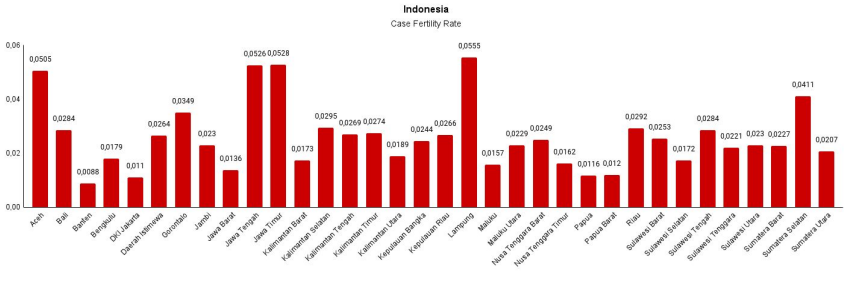
03. Data Analysis

4. Bagaimana gambaran Total case fatality rate dan case recovered rate dari masing-masing lokasi?

Case Recovered Rate (CRR) merupakan parameter yang memberikan informasi tentang seberapa efektif sistem perawatan kesehatan dalam memulihkan pasien akibat COVID-19. Semakin tinggi CRR, semakin banyak orang yang pulih. Terlihat bahwa angka CRR paling tinggi adalah **Papua Barat** sebesar 0,9838, diikuti **Banten** 0,838, dan **Papua** sebesar 0,9837.



Case Fatality Rate (CFR) merupakan parameter yang mencerminkan tingkat keparahan penyakit akibat COVID-19. Semakin tinggi CFR, semakin berat dampaknya. Terlihat bahwa angka CFR paling tinggi adalah **Lampung** sebesar 0,0555, diikuti **Jawa Timur** 0,0528, dan **Jawa Tengah** sebesar 0,0526.



03. Data Analysis

5. Kapan saja tanggal-tanggal saat total kasus Covid-19 mulai menyentuh angka 30.000-an?

Terlihat bahwa total kasus Covid-19 yang mulai menyentuh angka 30,000 tersebar di berbagai provinsi di Indonesia dimulai sejak Agustus 2020 di DKI Jakarta dan Jawa Timur. Disusul Jawa Tengah dan Jawa Barat pada Oktober 2020 dan Sulawesi Selatan pada Desember 2020. Selanjutnya total kasus Covid-19 yang mulai menyentuh angka 30,000 terus menyebar di provinsi lain.

Query Link

date	Location	total_cases
2020-08-18	DKI Jakarta	30463
2020-08-23	Jawa Timur	30078
2020-10-13	Jawa Tengah	30131
2020-10-14	Jawa Barat	30392
2020-12-29	Sulawesi Selatan	30029
2021-01-09	Kalimantan Timur	30119
2021-02-12	Riau	30027
2021-02-13	Bali	30134
2021-03-05	Banten	30189
2021-03-12	Sumatera Barat	30042
2021-03-13	Daerah Istimewa Yogyakarta	30027
2021-04-10	Kalimantan Selatan	30156
2021-05-09	Sumatera Utara	30058

date	Location	total_cases
2021-07-06	Sumatera Selatan	30152
2021-07-09	Kepulauan Riau	30637
2021-07-17	Kalimantan Tengah	30042
2021-07-21	Nusa Tenggara Timur	31090
2021-07-23	Lampung	30387
2021-07-25	Kepulauan Bangka Belitung	30009
2021-08-10	Sulawesi Tengah	30933
2021-08-12	Kalimantan Barat	30316
2021-08-15	Papua	30108
2021-08-18	Sulawesi Utara	30225
2021-08-22	Aceh	30077
2021-08-22	Kalimantan Utara	30074
2022-02-08	Jambi	30029
2022-02-11	Nusa Tenggara Barat	30399
2022-03-05	Papua Barat	30000

Dashboard

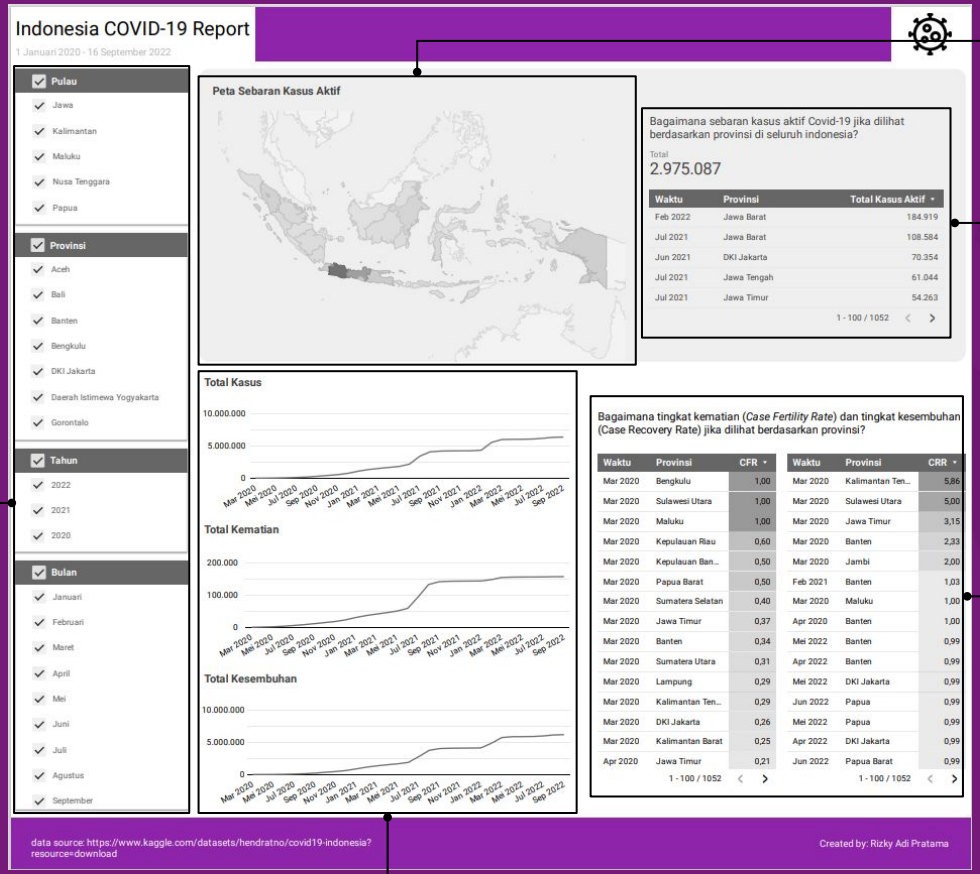
BINAR

Tujuan dari pembuatan dashboard ini adalah untuk mengetahui gambaran perkembangan kasus Covid-19 di Indonesia sejak 1 Januari 2020 hingga 16 September 2022, yang dilihat berdasarkan sebaran pulau, provinsi, tahun, dan bulan.

Menu drop-down untuk memfilter tampilan laporan Covid-19 berdasarkan pulau, provinsi, tahun, dan bulan.

Line chart untuk menggambarkan trend Covid-19 di Indonesia berdasarkan total kasus, total kematian, dan total kesembuhan.

Dashboard Link



Bagian ini menggambarkan sebaran kasus aktif Covid-19 di berbagai provinsi. Semakin pekat warnanya, maka semakin besar jumlah kasus aktif di daerah tersebut.

Bagian ini menggambarkan total kasus aktif Covid-19 dan tabel dibawah menunjukan top 5 kasus aktif terbesar berdasarkan waktu (*m-y*) dan sebaran provinsi.

Bagian ini menggambarkan kondisi Covid-19 yang dilihat berdasarkan waktu (*m-y*) dan provinsi menggunakan parameter *Case Fertility Rate* dan *Case Recovered Rate*.

BINAR

Tools

Challenge 2

BINAR

Table of Content

BINAR

01. Background & Objektif

Perkembangan industri telekomunikasi sekarang sangat 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. Namun, 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.

Oleh sebab itu, perlunya dilakukan **prediksi customer churn** agar perusahaan bisa **memetakan strategi bisnis untuk mempertahankan pelanggan**.

Objektif:
1. Mengetahui karakteristik dan distribusi data
2. Mengetahui hubungan antar fitur
3. Berhasil memprediksi customer churn dengan akurasi terbaik.

BINAR

02. Data Understanding

Dataset merupakan informasi terkait customer churn yang tersebar di **51 state** dan **3 area code** yang berbeda. Sebagai tambahan, dataset merupakan kumpulan informasi terkait **paket layanan yang digunakan**, **total panggilan**, **total durasi panggilan**, **total biaya selama melakukan panggilan**, dan **frekuensi penggunaan layanan customer service**.

20 Fitur

- 5 kategorik
- 15 numerik

4,250 Data

598 Churn

Challenge 2

BINAR

Gambaran dataset

	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
...	...	...	...	...	...	...	...	...	...	...
4245	MT	83	area_code_415	no	no	0	188.3	70	32.01	243.8
4246	WV	73	area_code_408	no	no	0	177.9	89	30.24	131.2
4247	NC	75	area_code_408	no	no	0	170.7	101	29.02	193.1
4248	HI	50	area_code_408	no	yes	40	235.7	127	40.07	223.0
4249	VT	86	area_code_415	no	yes	34	129.4	102	22.00	267.1

4250 rows × 20 columns

Dataset Link

BINAR

03. EDA

BINAR

03. EDA

A. Univariate Analysis

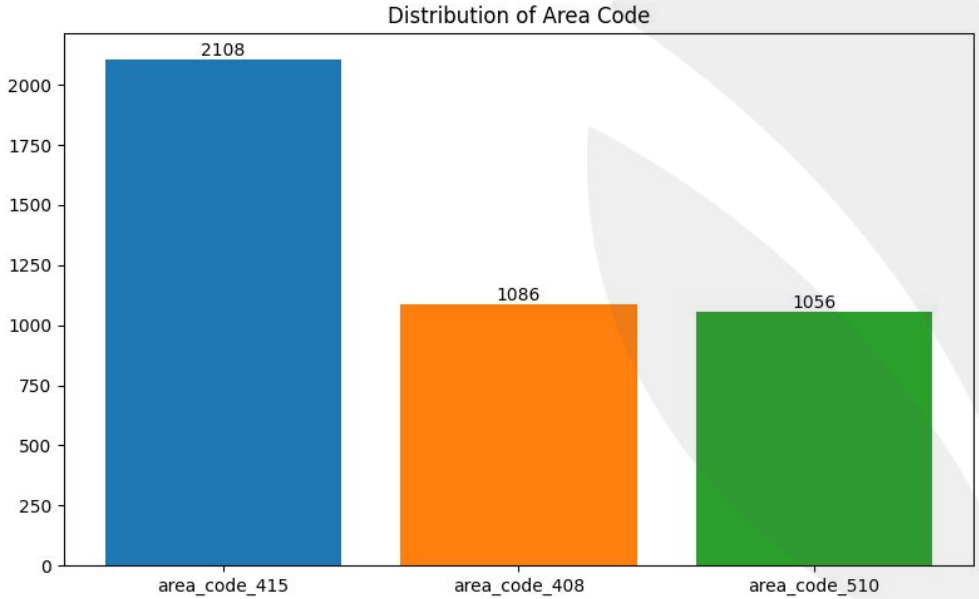
Tabel disamping merupakan statistik deskriptif untuk fitur numerik. Terlihat bahwa, **number_vmail_messages** memiliki nilai standar deviasi yang lebih besar dari pada nilai mean. Hal ini menandakan terlalu banyak varians pada fitur ini. Sementara itu, **number_customer_service_calls** memiliki nilai std yang mendekati nilai mean. Hal ini juga menandakan terlalu banyak variasi pada fitur ini.

	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

03. EDA

A. Univariate Analysis

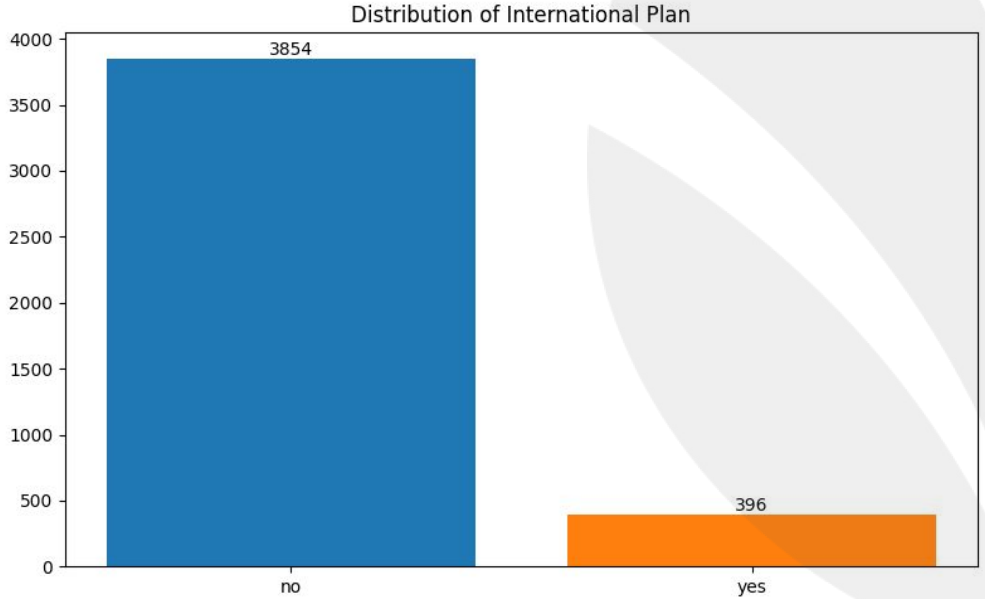
Lebih dari setengah sebaran customer berasal dari **area code 415**.



03. EDA

A. Univariate Analysis

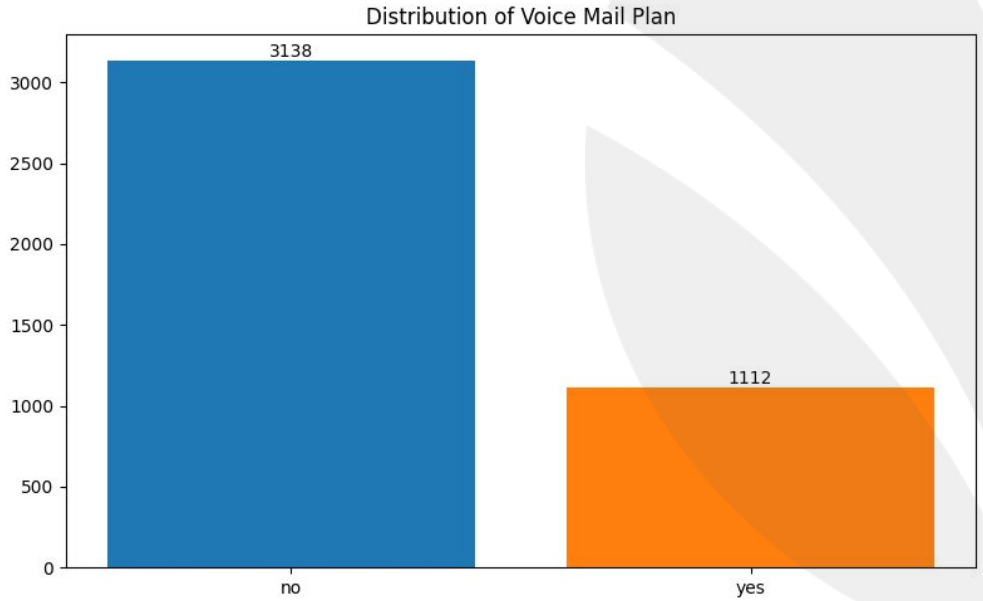
Terlihat bahwa tidak banyak customer yang menggunakan layanan paket internasional hanya sekitar **9%**.



03. EDA

A. Univariate Analysis

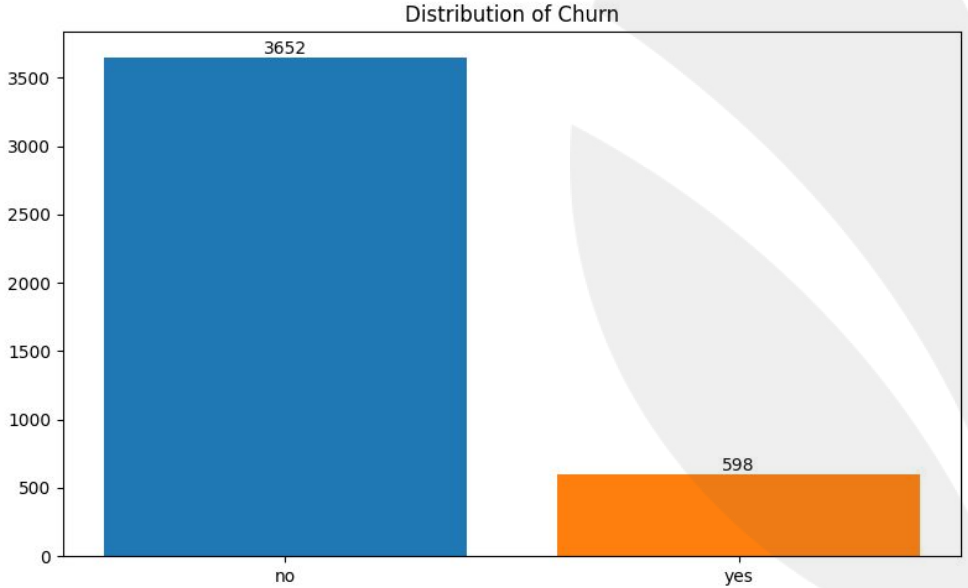
Terlihat bahwa tidak banyak customer yang menggunakan layanan paket voice mail hanya sekitar **26%**.



03. EDA

A. Univariate Analysis

Tidak banyak customer yang melakukan churn hanya sekitar **14%**. Fitur ini nantinya akan digunakan sebagai target dalam pemodelan. Disamping itu, terlihat bahwa jumlah kelas pada fitur ini sangat tidak seimbang, sehingga nantinya perlu dilakukan *balancing*.



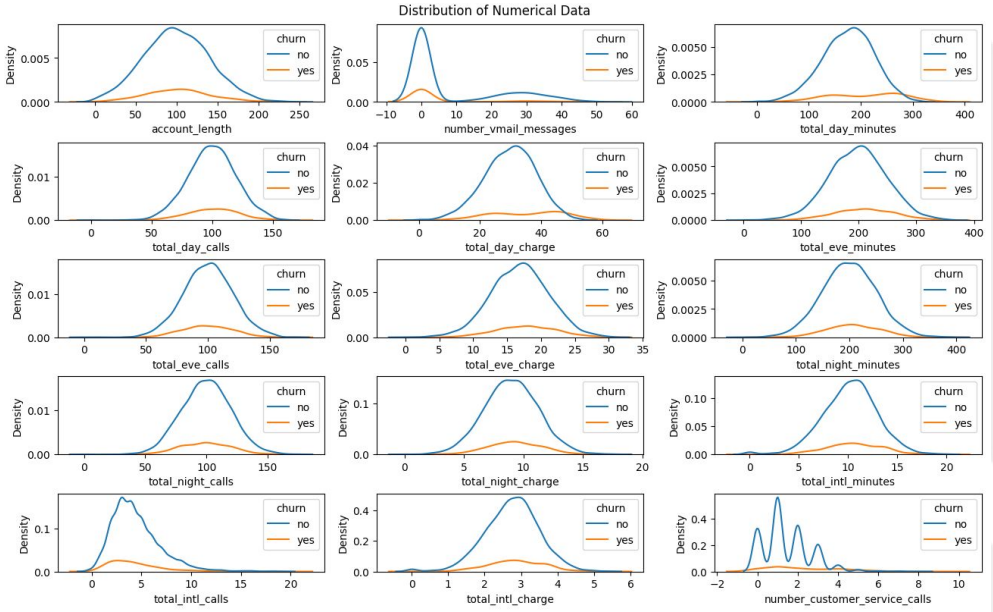
BINAR

03. EDA

A. Univariate Analysis

Fitur **number_vmail_messages**, **total_intl_calls**, dan **number_customer_service_calls** mengalami skewed positif sementara fitur lainnya cenderung simetris. Ketiga fitur memiliki banyak nilai ekstrim di dalamnya, khususnya pada fitur **number_vmail_messages.**

Kondisi tersebut dapat mempengaruhi hasil analisis dan model terbaik yang terbentuk nantinya. Dimana terkadang, dapat memberikan hasil yang salah (menyesatkan).



03. EDA

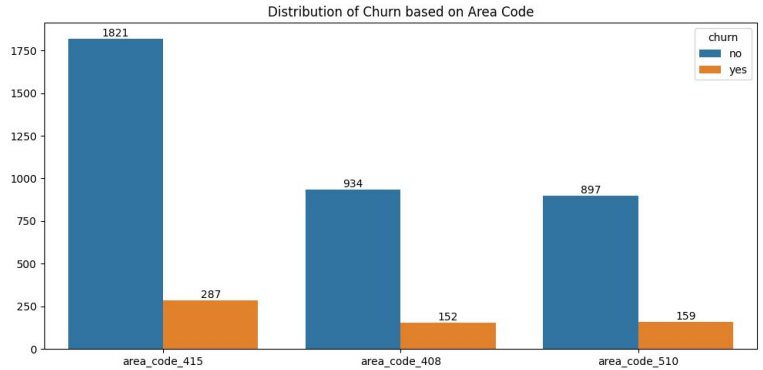
B. Multivariate Analysis

	churn	account_length		
		max	mean	min
0	no	243	99.924973	1
1	yes	225	102.137124	2

Customer yang masih setia menggunakan layanan provider, rata-rata telah memakai provider selama 100 hari. Sementara customer yang memutuskan untuk pindah layanan provider (churn), rata-rata telah memakai provider lebih lama 2 hari dari customer yang tidak churn.

03. EDA

B. Multivariate Analysis



Pada area code 408, 415, 510 secara berturut-turut persentase customer churn sebesar 14%, 13.6%, dan 15.1%. Berdasarkan informasi ini, persentase terbesar untuk customer berhenti menggunakan layanan provider adalah pada **area code 510**. Perusahaan penyedia provider perlu melakukan analisis lebih lanjut terkait faktor-faktor yang menjadi alasan customer berhenti menggunakan provider pada area tersebut.

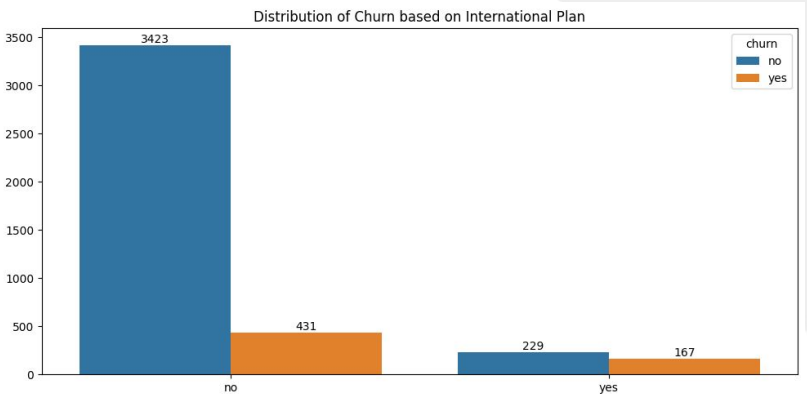
03. EDA

B. Multivariate Analysis

Diperoleh informasi bahwa:

- Untuk customer yang tidak menggunakan paket layanan internasional, tidak banyak customer yang melakukan churn. Bahkan tidak sampai setengah dari seluruh customer pada kategori ini.
- Untuk customer yang menggunakan paket layanan internasional, cukup banyak customer churn.

Perusahaan penyedia provider perlu melakukan evaluasi terhadap paket layanan internasional yang mungkin dari segi harga yang terlalu mahal jika dibandingkan dengan provider lain, sehingga banyak customer yang berhenti menggunakan provider dengan alasan tersebut.



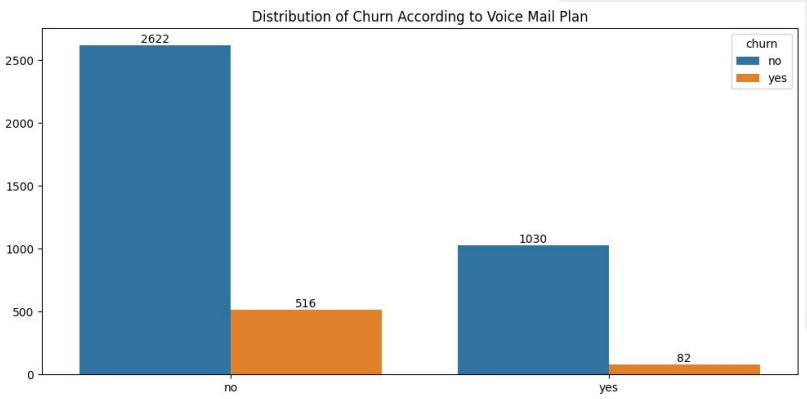
03. EDA

B. Multivariate Analysis

Diperoleh informasi bahwa:

- Untuk customer yang tidak menggunakan paket layanan pesan suara, terlihat tidak banyak customer churn pada kelompok ini.
- Untuk customer yang menggunakan paket layanan pesan suara, terlihat sangat sedikit customer churn pada kelompok ini.

Paket layanan pesan suara yang disediakan oleh provider ini dapat dikatakan memuaskan bagi customer yang mungkin dari segi harga atau kualitas layanan yang lebih baik dari pada provider lain. Perusahaan penyedia provider dapat menjadikan paket layanan ini menjadi produk unggulan untuk menarik customer baru.



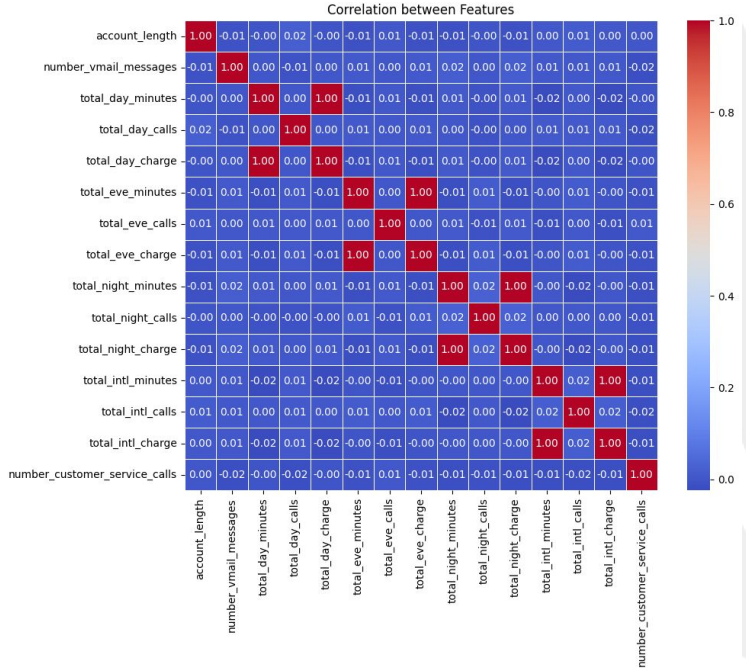
03. EDA

B. Multivariate Analysis

Diperoleh informasi bahwa:

- **total_day_minutes** vs **total_day_charge** berkorelasi positif sempurna.
- **total_eve_minutes** vs **total_eve_charge** berkorelasi positif sempurna.
- **total_night_minutes** vs **total_night_charge** berkorelasi positif sempurna.
- **total_intl_minutes** vs **total_intl_charge** berkorelasi positif sempurna.

Dapat disimpulkan bahwa semakin lama customer melakukan panggilan maka biaya yang dikeluarkan akan semakin besar. Selain itu, korelasi yang tinggi dari keempat kombinasi tersebut dapat diartikan jika terdapat gejala multikolinieritas. Sehingga nantinya salah satu fitur dari keempat kombinasi tersebut dapat dihapus.



03. EDA

B. Multivariate Analysis

```
alpha: 0.05

Churn vs Area Code
chi2: 1.2166542631365147 p-value: 0.5442605842955197

Churn vs International Plan
chi2: 282.65349013787664 p-value: 1.9831895448817517e-63

Churn vs Voice Mail Plan
chi2: 55.10981373962457 p-value: 1.139803854851859e-13
```

- **Churn vs Area Code:** Terdapat bukti bahwa tidak ada hubungan yang signifikan antara keputusan pelanggan untuk menghentikan langganan provider dan kode wilayah customer.
- **Churn vs International Plan**: Terdapat bukti bahwa ada hubungan yang signifikan antara keputusan pelanggan untuk menghentikan langganan provider dan apakah mereka memiliki paket panggilan internasional.
- **Churn vs Voice Mail Plan**: Terdapat bukti bahwa ada hubungan yang signifikan antara keputusan pelanggan untuk menghentikan langganan provider dan apakah mereka memiliki paket voicemail.

Karena **Area Code** tidak berkorelasi dengan **churn**, maka fitur ini tidak akan digunakan.

BINAR

04. Preprocessing

04. Preprocessing

A. *Missing Value* & Duplikasi

Dari 20 fitur, terlihat **data bersih dari *missing value***. Selain itu, dataset juga **tidak mengalami duplikasi** pada tiap barisnya.

```
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
```

```
Jumlah data duplikasi: 0
```

04. Preprocessing

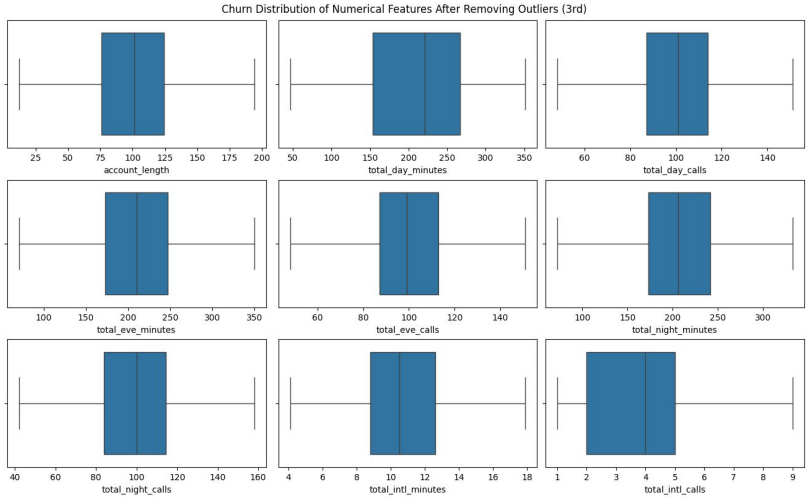
B. *Drop Unnecessary Features*

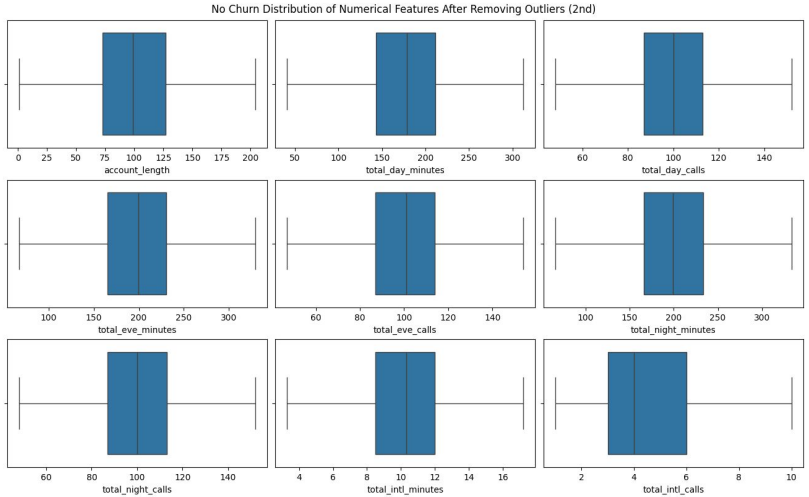
Beberapa fitur yang akan di drop:

- **state**: karena terlalu banyak label akan sulit untuk menemukan pola pada fitur tersebut.
- **area_code**: karena tidak berkorelasi dengan churn.
- **total_day_charge**, **total_eve_charge**, **total_night_charge**, dan **total_intl_charge**: untuk menghindari multikolinearitas
- **number_vmail_messages**: varians data terlalu besar, sehingga akan sulit menemukan polanya.
- **number_customer_service_calls**: nilai std mendekati mean yang berarti varians data cukup besar, sehingga akan sulit menemukan polanya.

04. Preprocessing

C. Outliers





Outliers dihapus berdasarkan masing-masing churn. Setelah dilakukan proses ini, jumlah data menjadi **3,896 baris**.

BINAR

04. Preprocessing

D. Feature Encoding

international_plan_yes	voice_mail_plan_yes
0	0
0	0
0	0
0	0
0	1

Pada tahap ini, fitur **international_plan** dan **voice_mail_plan** diubah menjadi fitur *dummy* dimana banyaknya fitur yang terbentuk adalah *k*(banyak label) - 1.

04. Preprocessing

E. Standardization

Sebelum dilakukan tahap ini, dataset perlu dibagi menjadi data train dan data validation dengan rasio 70:30.

Train

- Jumlah baris = 2,727
- churn(yes) = 380
- churn(no) = 2,347

Validation

- Jumlah baris = 1,169
- churn(yes) = 163
- churn(no) = 1006

Selanjutnya data train dan data validation dilakukan standarisasi berdasarkan pada **data train**. Hal ini dilakukan untuk memastikan bahwa informasi pada data validation tidak tercampur dengan data yang akan digunakan untuk pemodelan nantinya (data train).

04. Preprocessing

F. Balancing

Tahap balancing pada data train dilakukan menggunakan metode SMOTE untuk menyeimbangkan kelas minoritas terhadap kelas mayoritas.

Sebelum

- Jumlah baris = **2,727**
- churn(yes) = 380
- churn(no) = 2,347

Setelah

- Jumlah baris = **4,694**
- churn(yes) = 2,347
- churn(no) = 2,347

05. Pemodelan & Evaluasi

BINAR

05. Pemodelan dan Evaluasi

A. Pemodelan

Model	Presisi	Recall	Akurasi	F1-Score (weighted)
Logistic Regression	Yes: 0.27 No: 0.92	Yes: 0.60 No: 0.74	0.72	0.76
Naive Bayes	Yes: 0.24 No: 0.92	Yes: 0.63 No: 0.68	0.67	0.72
KNN	Yes: 0.33 No: 0.92	Yes: 0.57 No: 0.81	0.78	0.80
SVM	Yes: 0.45 No: 0.92	Yes: 0.54 No: 0.89	0.84	0.85
Decision Tree	Yes: 0.42 No: 0.93	Yes: 0.56 No: 0.87	0.83	0.84
Gradient Boosting	Yes: 0.82 No: 0.93	Yes: 0.56 No: 0.98	0.92	0.92

BINAR

05. Pemodelan dan Evaluasi

A. Pemodelan

Model	Hasil Prediksi (F1-Score)	
	Train	Validation
SVM	0.8891	0.8493
Gradient Boosting	0.9228	0.9159

Model terbaik yang dipilih untuk memprediksi data test adalah Gradient Boosting.

Google Colaboratory Link

BINAR

Tools

BINAR

Thank You