

Visualization & Machine Learning

Kelompok DS5-6

FGA x Binar Academy - Data Science

Meet our team!

DS5 - Team 6



Rifqi
Mufiddin



An Naffila Putri
Prasari





Table of contents



Challenge Chapter 1 - Visualization

Melakukan Query di Big Query

Merancang Dashboard

Challenge Chapter 2 - Machine Learning

Business Understanding

Exploratory Data Analysis

Data Preprocessing

Modeling

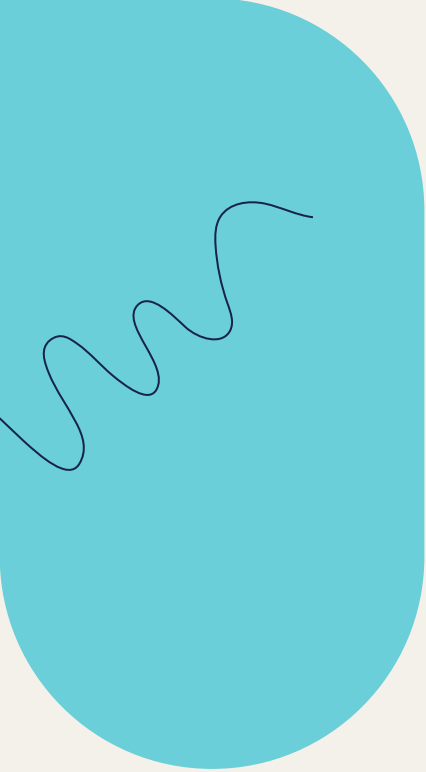
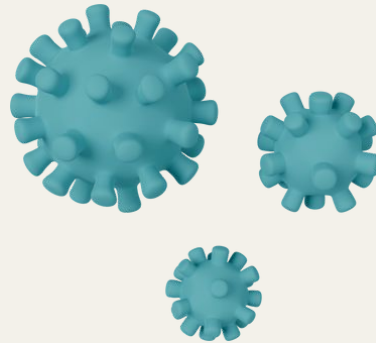
Model Evaluation

Conclusion



Challenge Chapter 1

Covid-19 Analysis



01

Melakukan Query di Big Query



Soal 1

Jumlah total kasus Covid-19 aktif yang baru di setiap provinsi lalu diurutkan berdasarkan jumlah kasus yang paling besar.

SQL Query

SELECT

Province,

SUM(New_Active_Cases) AS Total_New_Active_Cases

FROM `ds5-6-challenge-1.dataset_covid19.covid`

GROUP BY

Province

ORDER BY

Total_New_Active_Cases **DESC**

Row	Province	Total_New_Active_Cases
1	<i>null</i>	28460
2	Jawa Barat	13496
3	DKI Jakarta	10922
4	Banten	2558
5	Jawa Tengah	1423
6	Jawa Timur	1136
7	Daerah Istimewa Yogyakarta	669
8	Sumatera Utara	664
9	Sulawesi Utara	565
10	Bali	474

Results per page: 50 1 - 35 of 35

Insight

Top 3 provinsi dengan jumlah kasus aktif paling tinggi, yaitu:

1. *null* (Indonesia)
2. Jawa Barat
3. DKI Jakarta

Soal 2

Mengambil 2 (dua) location iso code yang memiliki jumlah total kematian karena Covid-19 paling sedikit.

SQL Query

```
SELECT
  Location_ISO_Code,
  SUM(Total_Deaths) AS Total_Deaths
FROM
  `ds5-6-challenge-1.dataset_covid19.covid`
GROUP BY
  Location_ISO_Code
ORDER BY
  Total_Deaths ASC
LIMIT 2
```

Row	Location_ISO_Code	Total_Deaths
1	ID-MA	147196
2	ID-MU	167511

Insight

Maluku (ID-MA) dan Maluku Utara (ID-MU) menjadi provinsi dengan total kematian karena Covid-19 paling sedikit.

Soal 3

Data tentang tanggal-tanggal ketika rate kasus recovered di Indonesia paling tinggi beserta jumlah ratenya.

Row	Date	Case_Recovered_Rate	Avg_Case_Recovered_Rate
1	2020-03-06	111.0	4.4693526829268295
2	2020-04-01	5.8571	0.37828499515972897
3	2021-03-04	1.0487	0.86797382488479258
4	2021-02-28	1.0256999999999998	0.82073826530612248
5	2022-05-08	0.98930000000000007	0.97287290322580644
6	2022-06-01	0.9892	0.97318980952380951
7	2020-07-22	0.9889	0.61586691244239633
8	2022-04-30	0.9886	0.96156314285714284
9	2022-07-11	0.9872	0.97200138248847923
10	2022-08-01	0.98620000000000008	0.97019944700460825

Results per page: 50 1 - 31 of 31

Insight

Pada tanggal 6 Maret 2020, rate kasus recovered paling tinggi mencapai 111% dengan rata-rata rate sebesar 4,47%.

SQL Query

```
WITH RankedDates AS (  
    SELECT  
        Date,  
        Case_Recovered_Rate,  
        EXTRACT(YEAR FROM Date) AS Year,  
        EXTRACT(MONTH FROM Date) AS Month,  
        ROW_NUMBER() OVER(PARTITION BY EXTRACT(YEAR FROM Date),  
        EXTRACT(MONTH FROM Date) ORDER BY Case_Recovered_Rate DESC)  
        AS Rank,  
        AVG(Case_Recovered_Rate) OVER(PARTITION BY EXTRACT(YEAR  
        FROM Date), EXTRACT(MONTH FROM Date)) AS  
        Avg_Case_Recovered_Rate  
    FROM `ds5-6-challenge-1.dataset_covid19.covid`  
)  
SELECT  
    Date,  
    Case_Recovered_Rate,  
    Avg_Case_Recovered_Rate  
FROM RankedDates  
WHERE Rank = 1  
ORDER BY Case_Recovered_Rate DESC
```


Soal 4

Total case fatality rate dan case recovered rate dari masing-masing location iso code yang diurutkan dari data yang paling rendah.

SQL Query

SELECT

Location_ISO_Code,

AVG(Case_Fatality_Rate) AS Avg_Case_Fatality_Rate,

AVG(Case_Recovered_Rate) AS Avg_Case_Recovered_Rate

FROM `ds5-6-challenge-1.dataset_covid19.covid`

GROUP BY

Location_ISO_Code

ORDER BY

Avg_Case_Fatality_Rate, Avg_Case_Recovered_Rate **ASC**

Row	Location_ISO_Code	Avg_Case_Fatality_Rate	Avg_Case_Recovered_Rate
1	ID-KU	0.015837028824833708	0.81344412416851442
2	ID-NT	0.017903932584269679	0.78743910112359483
3	ID-PA	0.018607158590308367	0.66985969162995584
4	ID-JA	0.019040439560439569	0.83574637362637394
5	ID-SG	0.021379021739130418	0.80615695652173891
6	ID-KB	0.022820199778024405	0.85635271920088751
7	ID-SR	0.024146059933407321	0.81339877913429559
8	ID-SN	0.024651372118551035	0.85103995609220662
9	ID-SB	0.026560066371681439	0.834350774336284
10	ID-PB	0.026948062015503902	0.83853676633443952

Results per page: 50 1 - 35 of 35

Insight

Top 3 provinsi dengan rata-rata tingkat keparahan kasus Covid-19 paling rendah, yaitu:

1. Kalimantan Utara (ID-KU)
2. Nusa Tenggara Timur (ID-NT)
3. Papua (ID-PA)

Soal 5

Data tentang tanggal-tanggal saat total kasus Covid-19 mulai menyentuh angka 30.000-an.

SQL Query

```
SELECT
    Date,
    Total_Cases
FROM
    `ds5-6-challenge-1.dataset_covid19.covid`
WHERE
    Total_Cases >= 30000
```

Row	Date	Total_Cases
1	2020-06-06	30514
2	2020-06-07	31186
3	2020-06-08	32033
4	2020-06-09	33075
5	2020-06-10	34316
6	2020-06-11	35295
7	2020-06-12	36406
8	2020-06-13	37420
9	2020-06-14	38277
10	2020-06-15	39294

Results per page: 50 1 – 50 of 14399

Insight

Total kasus Covid-19 mulai menyentuh angka 30.000 pada tanggal 6 Juni 2020.

Soal 6

Jumlah data yang tercatat ketika kasus Covid-19 lebih dari atau sama dengan 30.000.

SQL Query

```
SELECT  
  COUNT(*) AS Total_Data  
FROM  
  `ds5-6-challenge-1.dataset_covid19.covid`  
WHERE  
  Total_Cases >= 30000
```

Row	Total_Data
1	14399

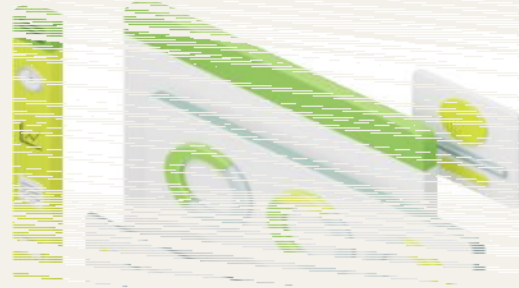
Insight

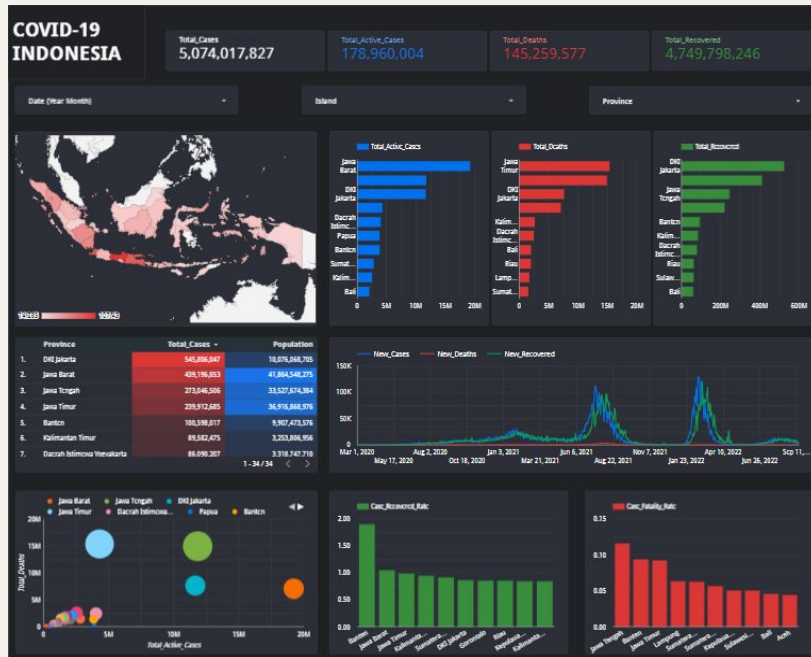
Terdapat 14.399 data yang tercatat untuk total kasus Covid-19 lebih dari atau sama dengan 30.000.



02

Merancang dashboard





Link: [Dashboard Covid-19 Indonesia](#)



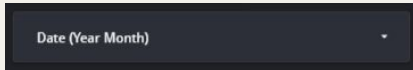
Scorecards

Scorecards ini memberikan informasi cepat tentang COVID-19. *Total_Cases* menunjukkan seberapa banyak orang yang terkena, *Total_Active_Cases* menunjukkan yang masih sakit, *Total_Deaths* menunjukkan yang meninggal, dan *Total_Recovered* menunjukkan yang sembuh. Dengan ini, *scorecards* memberikan gambaran singkat tentang situasi COVID-19.

Total_Cases 5,074,017,827	Total_Active_Cases 178,960,004	Total_Deaths 145,259,577	Total_Recovered 4,749,798,246
------------------------------	-----------------------------------	-----------------------------	----------------------------------

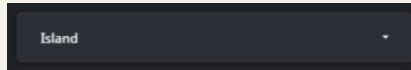


Filter Dropdown List



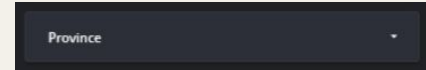
Date (Year Month)

Memungkinkan *user* untuk merinci analisis berdasarkan rentang waktu tertentu, membantu mengidentifikasi tren seiring waktu.



Island

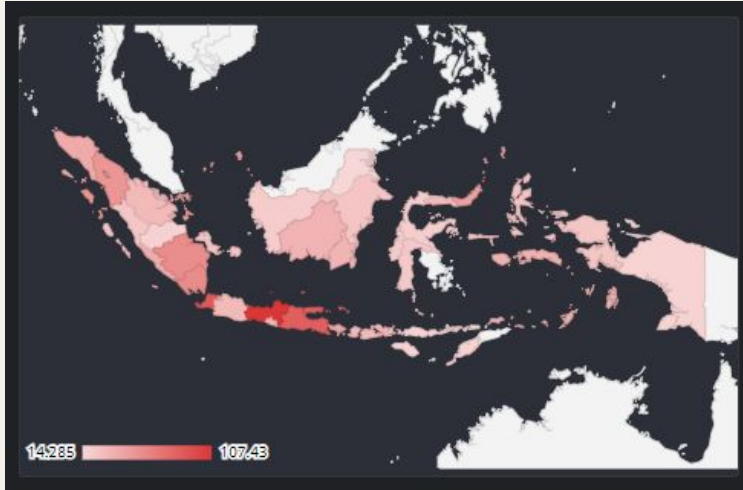
Memfasilitasi pemfilteran data berdasarkan lokasi pulau, berguna untuk menganalisis perbedaan antar pulau dalam hal penyebaran dan dampak COVID-19.



Province

Memberikan fleksibilitas kepada *user* untuk mempersempit analisis ke tingkat provinsi, membantu memahami situasi setempat dengan lebih detail.





Geo Chart



Membantu *user* mengidentifikasi provinsi dengan tingkat kematian tertinggi, memberikan informasi penting tentang sebaran tingkat keparahan di berbagai wilayah. Dengan visualisasi ini, *Geo Chart* menyediakan gambaran cepat untuk mendukung pengambilan keputusan terkait COVID-19.



Bar Charts

Bar Charts menggunakan data provinsi dan menampilkan perbandingan visual antara total kasus aktif, kematian, dan pulih. Ini membantu dengan mudah melihat sebaran efek pandemi COVID-19 di berbagai provinsi. Dengan *Bar Charts*, user dapat mengidentifikasi pola dan fokus pada wilayah-wilayah yang membutuhkan perhatian khusus.





	Province	Total_Cases ▾	Population
1.	DKI Jakarta	545,806,047	10,076,068,705
2.	Jawa Barat	439,196,053	41,864,548,275
3.	Jawa Tengah	273,046,506	33,527,674,384
4.	Jawa Timur	239,912,685	36,916,868,976
5.	Banten	100,598,017	9,907,473,576
6.	Kalimantan Timur	89,582,475	3,253,806,956
7.	Daerah Istimewa Yoevakarta	86.090.207	3.318.747.710

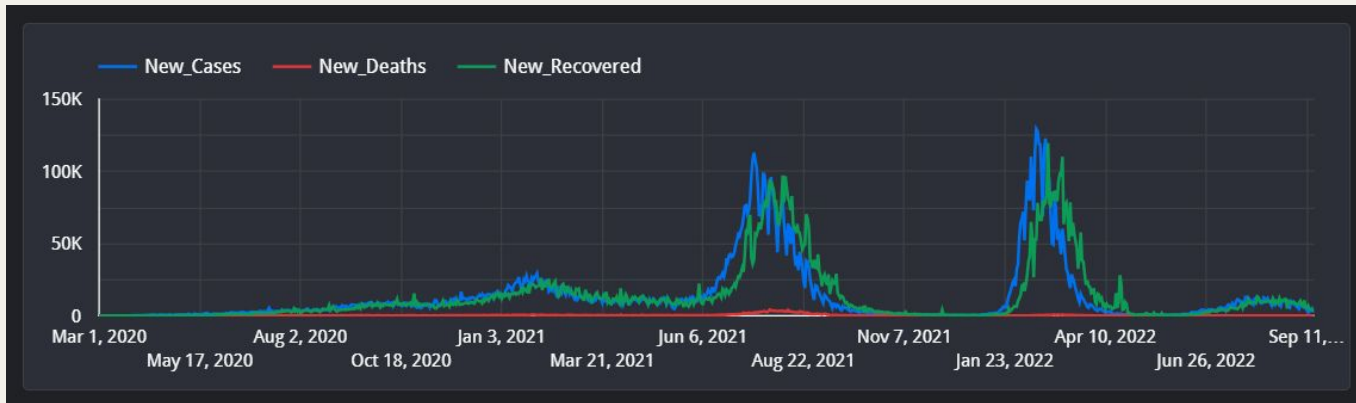
1 - 34 / 34 < >

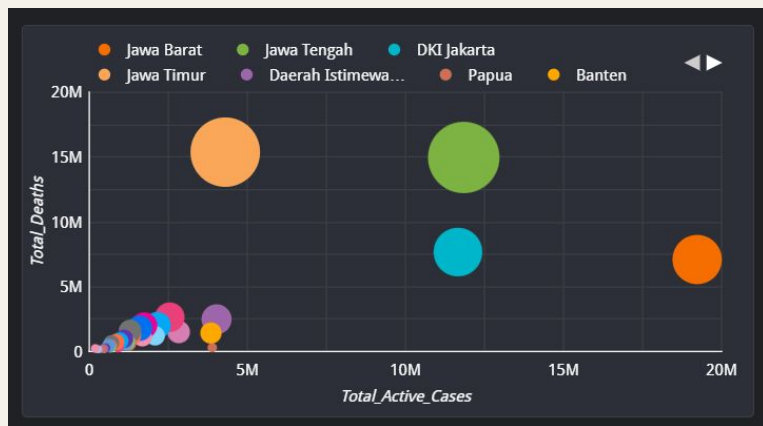
Table with Heatmap

Membantu *user* membandingkan total kasus Covid-19 dengan populasi di setiap wilayah untuk memberikan informasi terkait konteks demografis.

Time Series Charts

Time series chart digunakan untuk melihat tren harian kasus baru, kematian baru, dan pulih baru. Ini memudahkan user untuk mengetahui perkembangan kasus Covid-19 dari waktu ke waktu.





Bubble Chart

Bubble chart menggunakan data total kasus aktif dan total kematian setiap provinsi untuk membandingkan hubungan antara total kematian dengan total kasus aktif dengan lebih cepat dan mudah.



Column Charts

Column charts menggunakan data provinsi dan menampilkan informasi tingkat kematian dan tingkat kesembuhan. Hal ini dapat menjadi indikator keberhasilan penanganan Covid-19 di setiap provinsi.



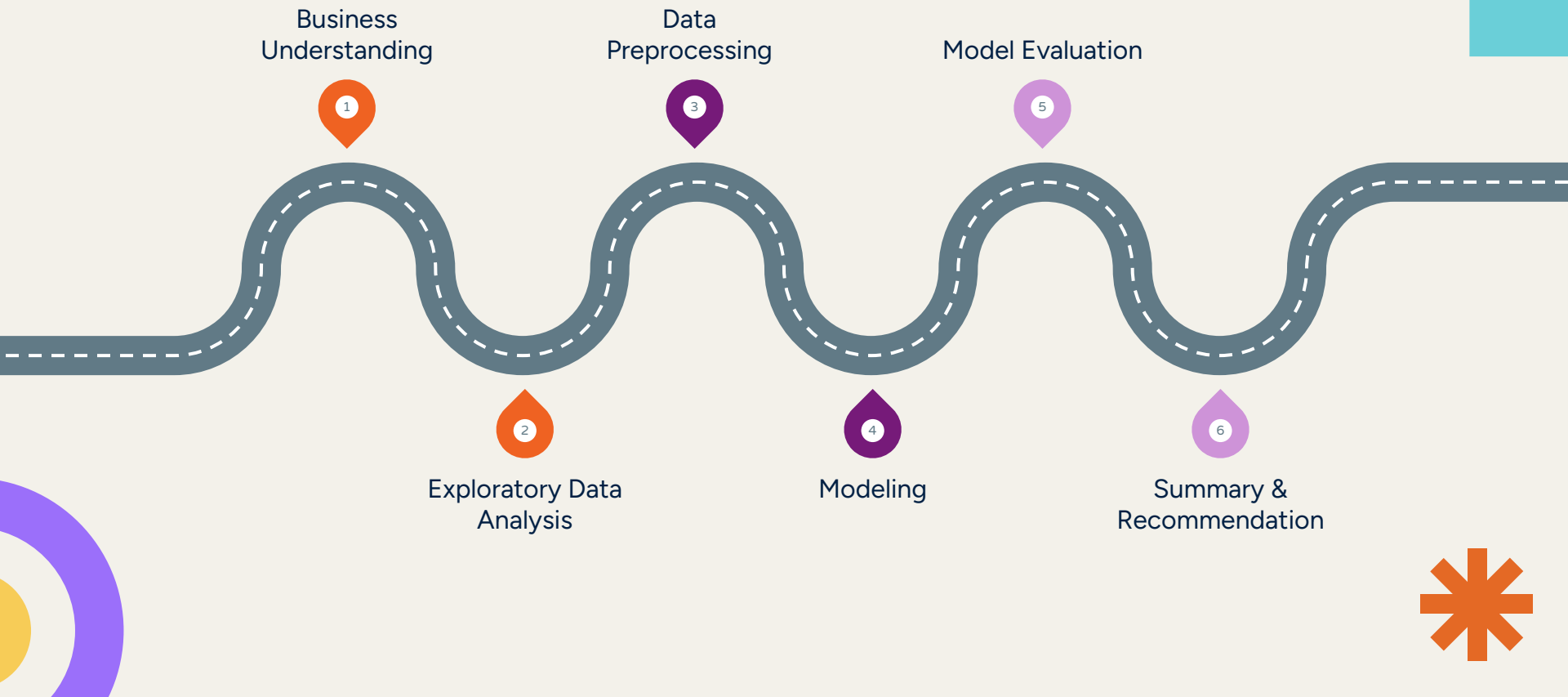


Challenge Chapter 2

Telco Customer Churn Prediction



Roadmap



01

Business Understanding



Problem Statement

Perkembangan industri telekomunikasi telah **memperketat persaingan** antar provider. Di samping itu, pelanggan memiliki **hak** untuk memilih provider yang sesuai dengan kebutuhan mereka dan dapat **beralih (churn)** dari provider sebelumnya. Hal ini dapat menyebabkan berkurangnya pendapatan bagi perusahaan telekomunikasi sehingga penting untuk ditangani.

Objective

Membangun **model klasifikasi** yang dapat mengenali pelanggan yang **berpotensi beralih (churn)** dari layanan telekomunikasi. Dengan demikian, perusahaan dapat mengambil tindakan untuk mempertahankan pelanggan.

02

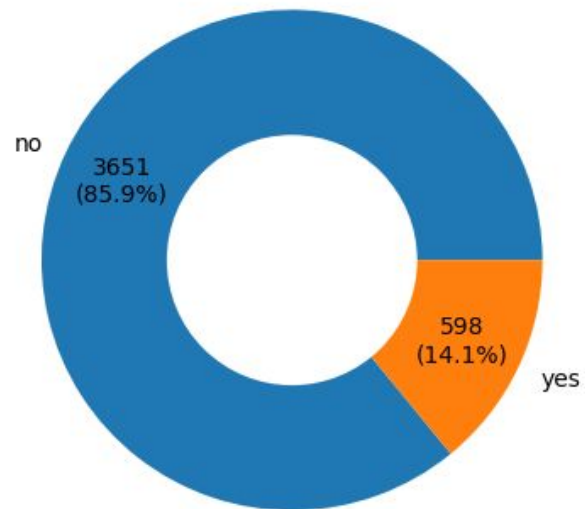
Exploratory Data Analysis

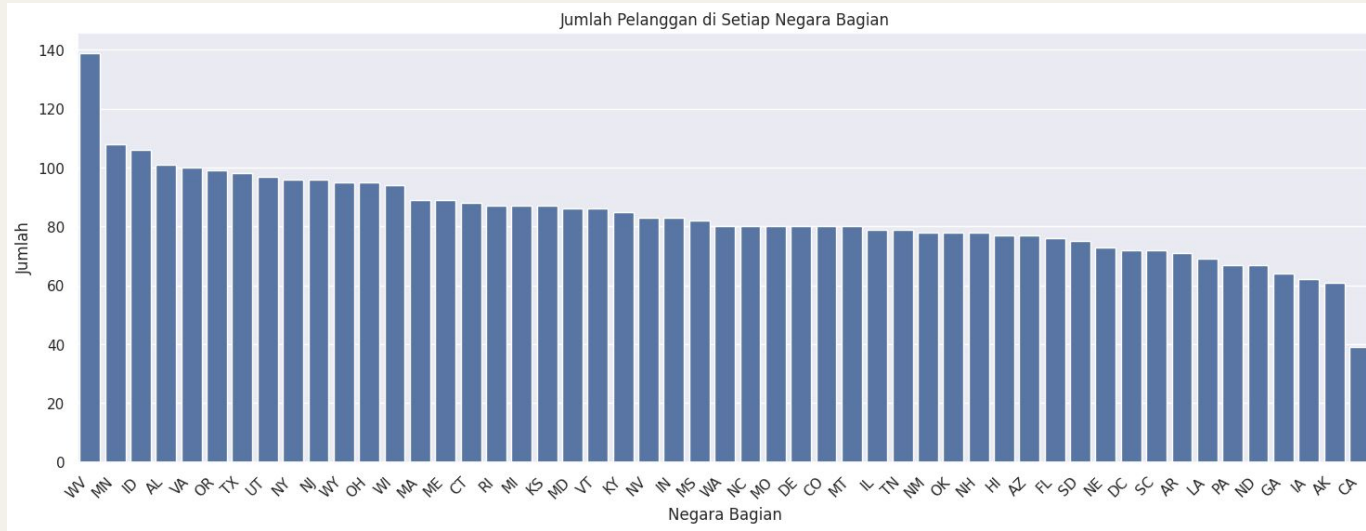




Terdapat sebanyak **14.1%** pengguna layanan telekomunikasi yang melakukan ***churn***.

Persentase Pelanggan yang Churn



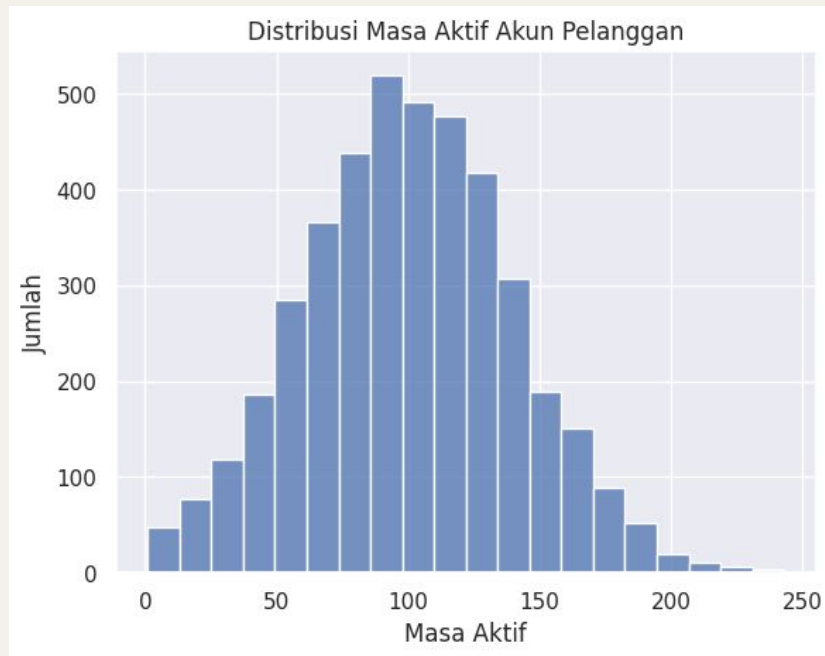


Top 5 **negara bagian** dengan jumlah pengguna layanan telekomunikasi **terbanyak**, yaitu:

1. West Virginia (WV)
2. Minnesota (MN)
3. Idaho (ID)
4. Alabama (AL)
5. Virginia (VA)

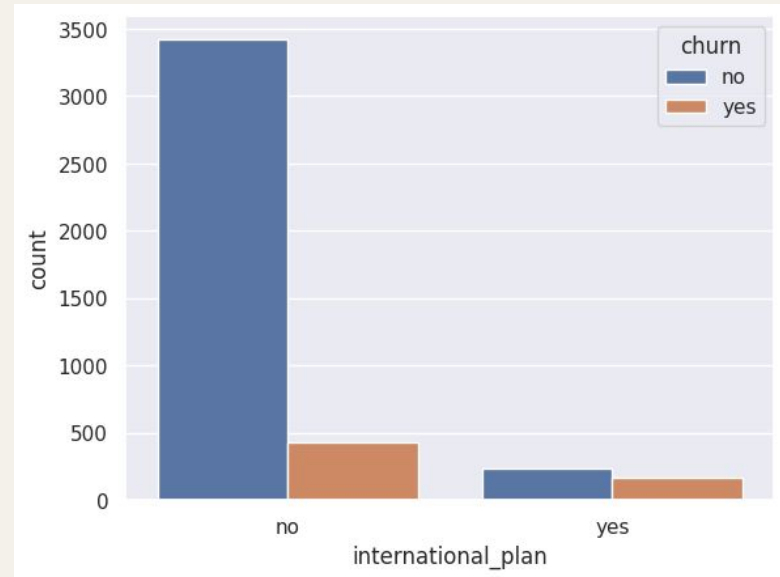


Masa aktif akun pelanggan **paling banyak** berada pada selang **87-100 hari** dengan jumlah mencapai lebih dari **500 pelanggan**.



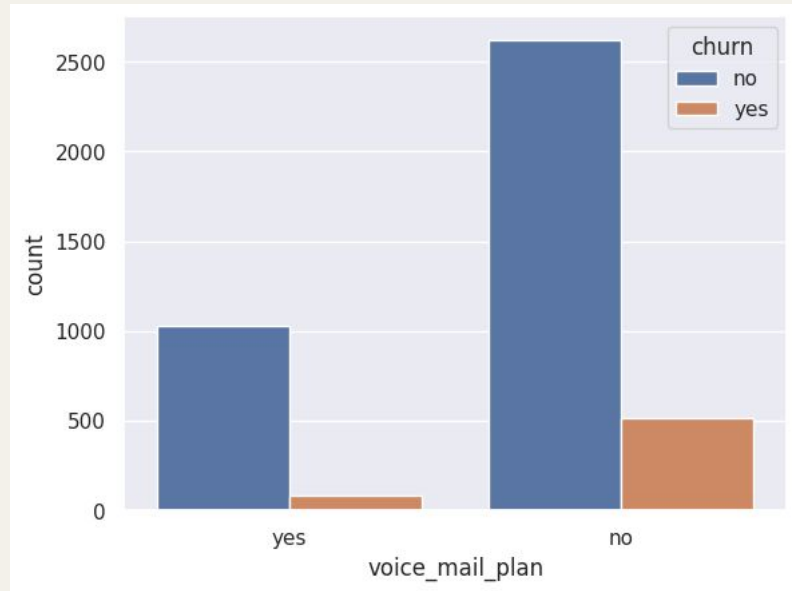
Pengguna layanan telekomunikasi yang **berlangganan Layanan Internasional** lebih mungkin untuk melakukan **churn** yaitu sebesar **42.17%**.

	international_plan	Churn_NO	Churn_YES	Total	Churn Rate
0	no	3423	431	3854	11.18
1	yes	229	167	396	42.17
2	All	3652	598	4250	14.07



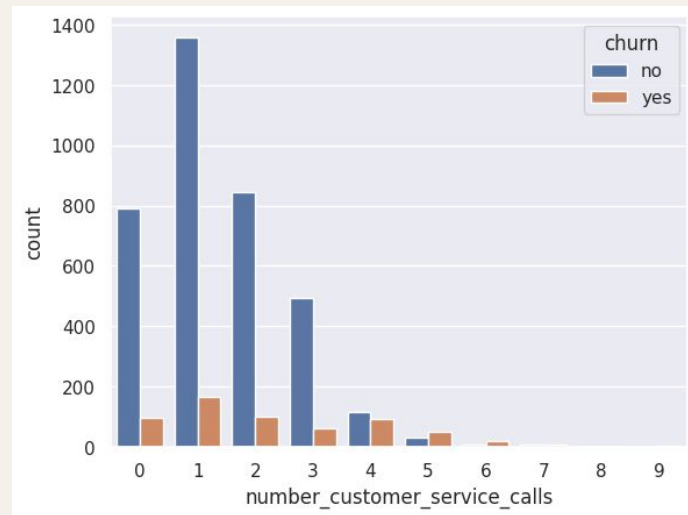
Pengguna layanan telekomunikasi yang **berlangganan Voice Mail** lebih cenderung untuk **tidak churn** yaitu hanya sebesar **7.37%**.

	voice_mail_plan	Churn_NO	Churn_YES	Total	Churn Rate
0	no	2622	516	3138	16.44
1	yes	1030	82	1112	7.37
2	All	3652	598	4250	14.07

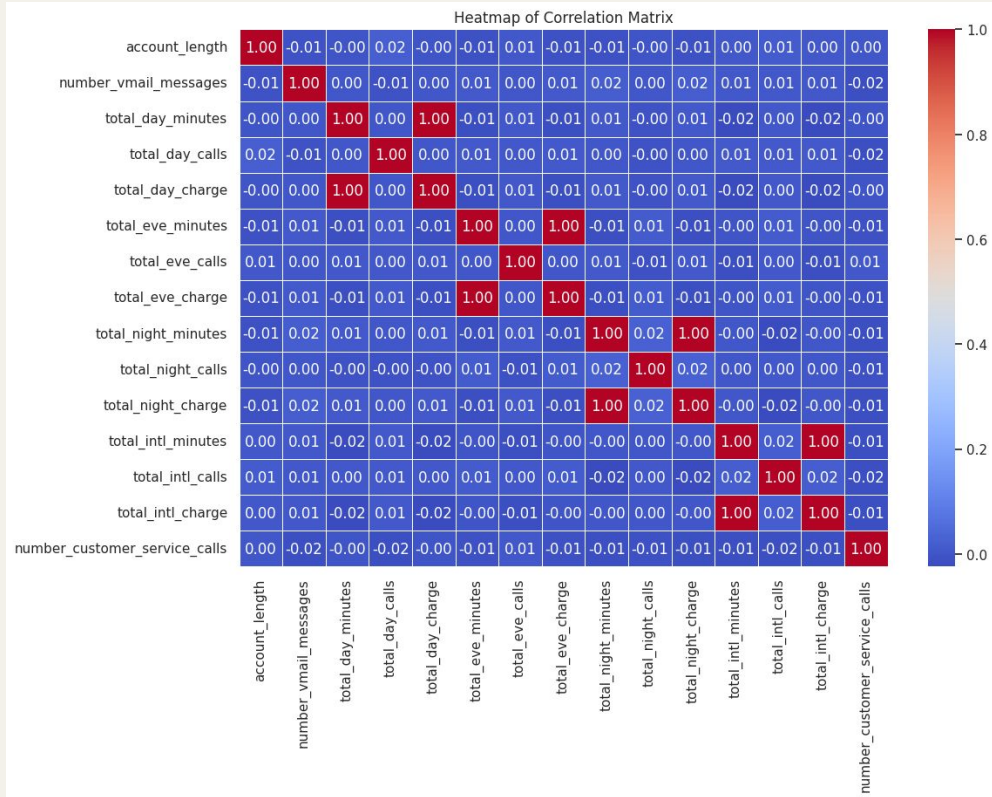


Tingkat *churn* meningkat ketika jumlah panggilan ke *Customer Service* terjadi **lebih dari 3 kali**.

	number_customer_service_calls	Churn_NO	Churn_YES	Total	Churn Rate
0	0	789	97	886	10.95
1	1	1358	166	1524	10.89
2	2	845	102	947	10.77
3	3	495	63	558	11.29
4	4	117	92	209	44.02
5	5	32	49	81	60.49
6	6	9	19	28	67.86
7	7	6	7	13	53.85
8	8	1	1	2	50.00
9	9	0	2	2	100.00
10	All	3652	598	4250	14.07



Semakin **lama durasi telepon**
maka semakin **besar charge**
atau biaya yang dikenakan.



03

Data Preprocessing





Dataset

Terdiri dari **4.250 baris** dan **20 kolom** yang mengandung data pengguna layanan telekomunikasi yang melakukan *churn* atau tidak.

Step by step

1. Data Cleaning

Tidak ada *missing value* dan data duplikat.

2. Detect Outliers

Tanpa menghapus *outliers*, model sudah bekerja dengan baik.

3. Features Encoding

Label encoding

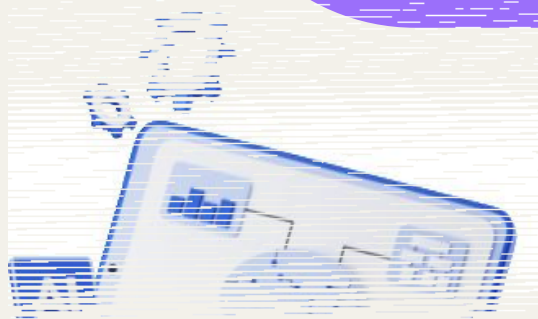
One-hot encoding

4. Standardization



04

Modeling



Algoritma yang digunakan



**Logistic
Regression**



**K-Nearest
Neighbors**



**Decision
Tree**



**Random
Forest**

Link: [Notebook Challenge Chapter 2](#)





Top 5 fitur penting model Logistic Regression

	Feature	Coefficient
19	state_CA	1.360335
70	international_plan_yes	1.343449
71	voice_mail_plan_no	1.169624
46	state_NJ	1.156521
41	state_MT	1.058861





Top 5 fitur penting model Decision Tree

	Feature	Importance
2	total_day_minutes	0.290404
14	number_customer_service_calls	0.124937
13	total_intl_charge	0.106231
5	total_eve_minutes	0.087500
69	international_plan_no	0.082645





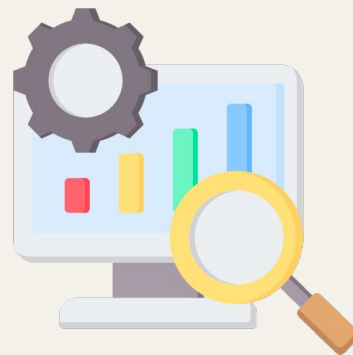
Top 5 fitur penting model Random Forest

	Feature	Importance	Std
2	total_day_minutes	0.123977	0.071983
14	number_customer_service_calls	0.123639	0.042084
4	total_day_charge	0.123092	0.065567
5	total_eve_minutes	0.054934	0.020595
69	international_plan_no	0.049977	0.040241

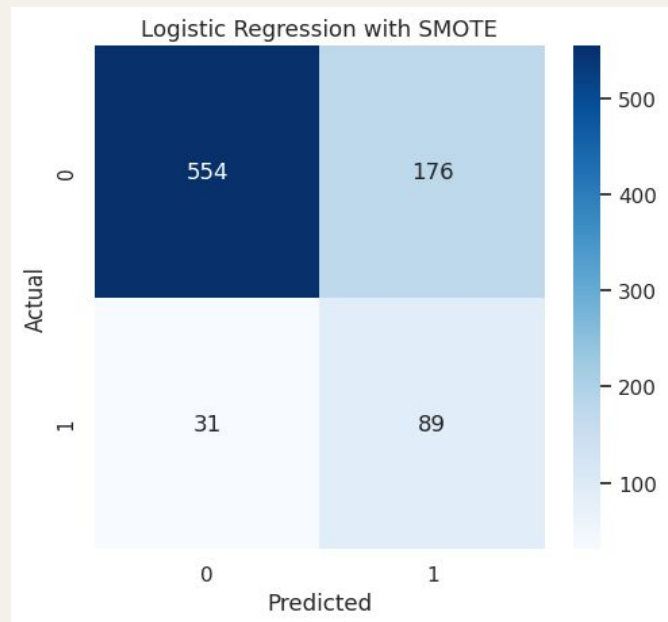
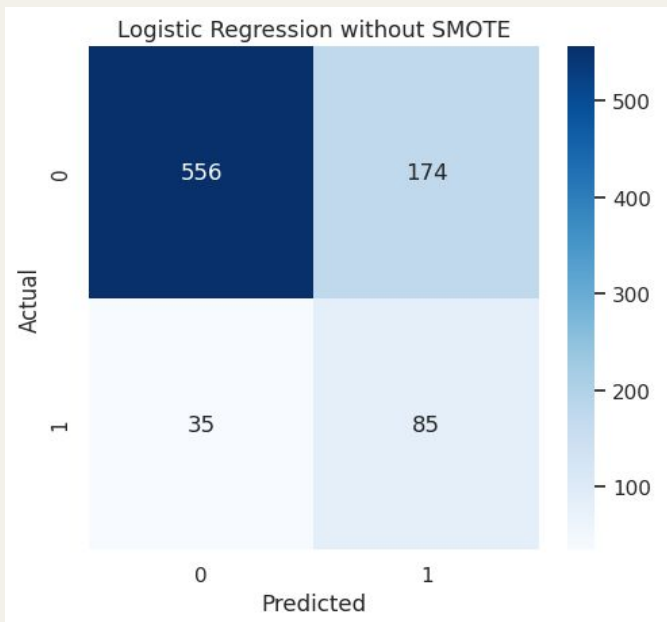


05

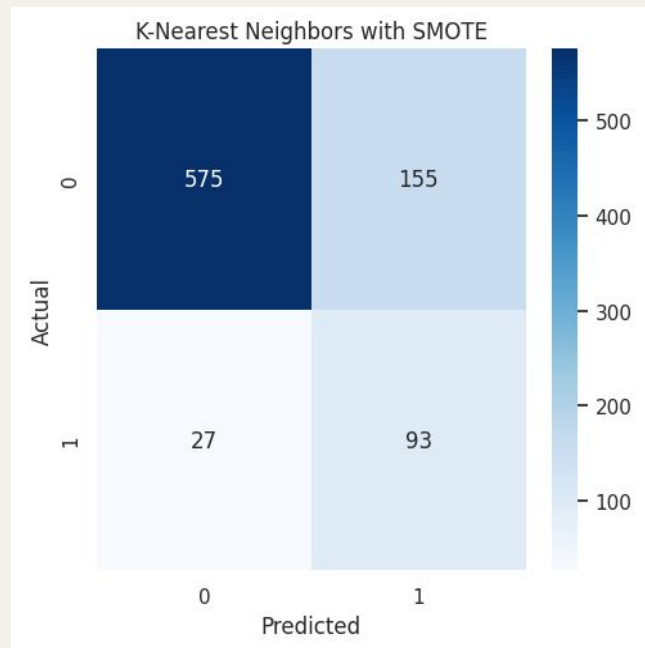
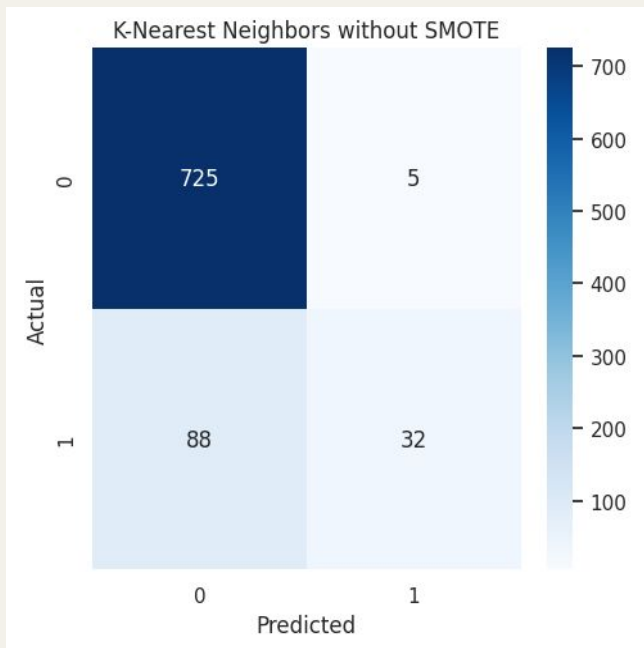
Model Evaluation



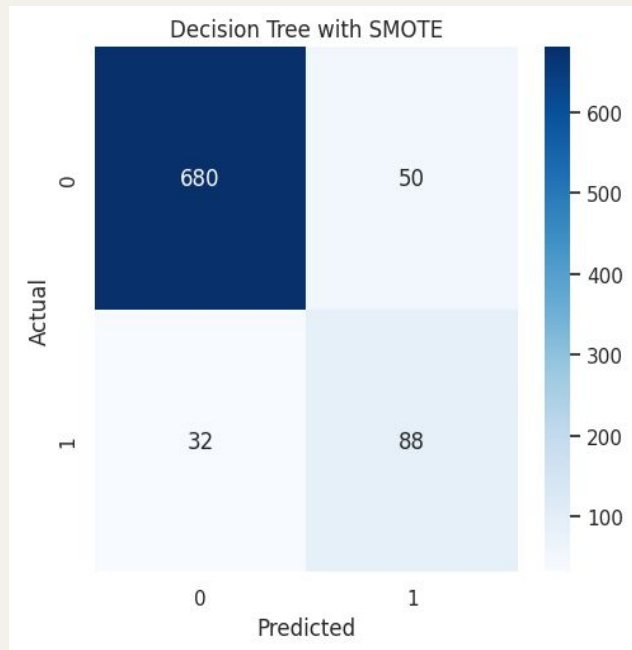
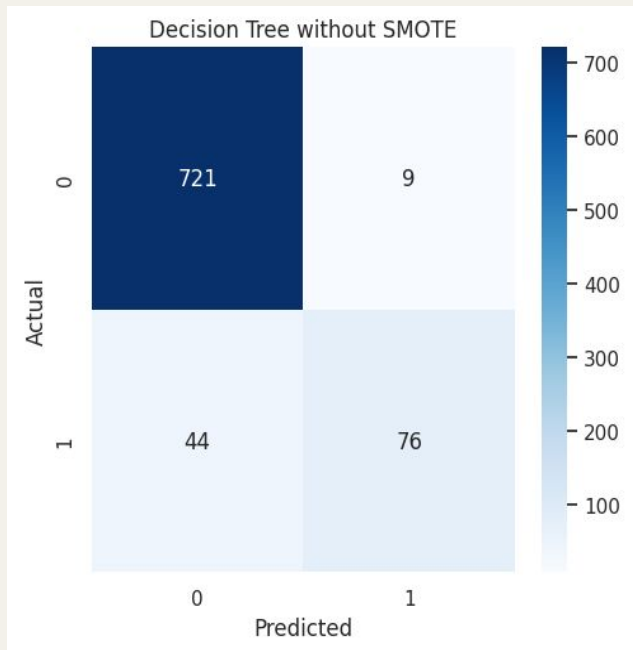
Confusion Matrix (Logistic Regression)



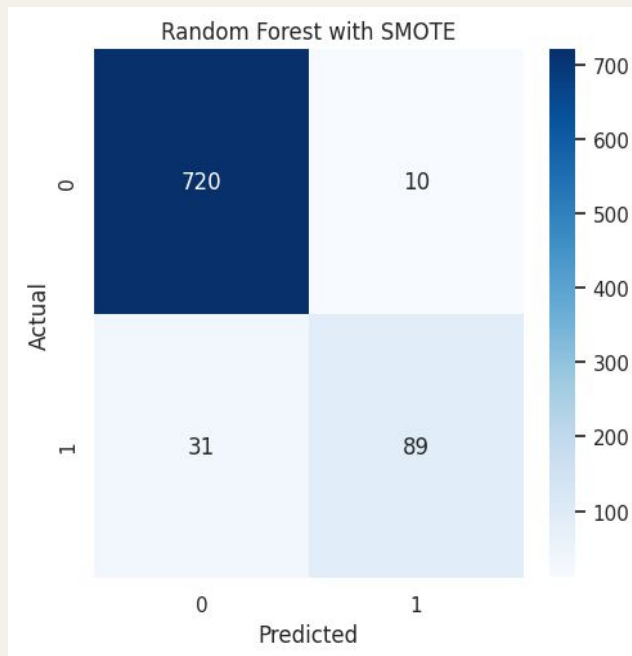
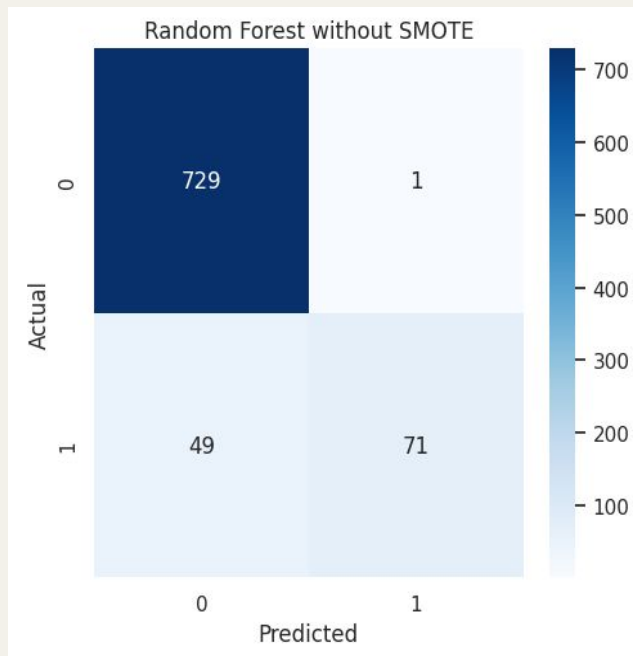
Confusion Matrix (K-Nearest Neighbors)



Confusion Matrix (Decision Tree)



Confusion Matrix (Random Forest)



Berikut disajikan tabel metrik evaluasi yang berfokus pada hasil model yang memprediksi *customers* yang *churn*.

Metric	Logistic Regression		K-Nearest Neighbors		Decision Tree		Random Forest	
	Without SMOTE	With SMOTE	Without SMOTE	With SMOTE	Without SMOTE	With SMOTE	Without SMOTE	With SMOTE
Akurasi	75%	76%	89%	79%	94%	90%	94%	95%
Presisi	33%	34%	86%	38%	89%	64%	99%	90%
Recall	71%	74%	27%	78%	63%	73%	59%	74%
F1-Score	45%	46%	41%	51%	74%	68%	74%	81%

Metrik **F1-Score** dipilih sebagai metrik evaluasi karena menggabungkan presisi dan *recall* yang memungkinkan keduanya dipertimbangkan secara bersamaan. **Presisi** berguna untuk **meminimalkan false positive** (mengidentifikasi pelanggan yang sebenarnya tidak akan *churn* sebagai pelanggan *churn*), sementara **recall** berguna untuk **meminimalkan false negative** (mengidentifikasi pelanggan yang sebenarnya akan *churn* sebagai pelanggan yang tidak akan *churn*). Oleh karena itu, **model terbaik** untuk kasus *customer churn* ini adalah **Random Forest with SMOTE**.



Conclusion



Conclusion

1. Data dari Big Query digunakan untuk mendapatkan insight tentang situasi Covid-19 di Indonesia.
2. Jawa Barat dan DKI Jakarta memiliki jumlah kasus aktif tertinggi, sementara Maluku dan Maluku Utara memiliki jumlah kematian terendah.
3. Dashboard dirancang untuk memberikan informasi cepat tentang Covid-19, memungkinkan pengguna untuk memahami situasi, menganalisis tren, dan mengambil keputusan dalam penanganan pandemi.
4. Model Random Forest dengan teknik SMOTE menghasilkan tingkat akurasi sebesar 95%, dengan kemampuan mendeteksi pelanggan *churn* (*recall*) sebesar 74% serta mencapai F1-score sebesar 81%.

Recommendation Challenge 2

1. Perusahaan perlu cepat tanggap dan memberikan solusi yang memuaskan ketika pelanggan menyampaikan keluhan melalui *Customer Service*, agar *churn rate* semakin berkurang.
2. Perusahaan perlu memastikan layanan internasional yang diberikan berkualitas tinggi dan responsif terhadap kebutuhan pelanggan karena pengguna cenderung untuk *churn*.
3. Perusahaan dapat memberikan diskon atau bonus kepada pengguna yang telah berlangganan *Voice Mail* selama periode tertentu untuk menjaga retensi pelanggan.



Report Pembagian Tugas

Report Pembagian Tugas

Nama	Tasklist/Deliverable
An Naffila Putri Prasari	<ul style="list-style-type: none">- Melakukan <i>query</i> di Big Query- Membuat <i>chart</i> pada <i>dashboard</i> (<i>geo chart, bubble chart, line chart, scorecards, dan table with heatmap</i>)- Membuat penjelasan <i>dashboard</i>- Melakukan <i>Exploratory Data Analysis</i> (5 pertanyaan)- Membuat model klasifikasi (K-Nearest Neighbors dan Decision Tree)- Menentukan model terbaik dengan metrik evaluasi- Membuat kesimpulan dan saran- Menyusun <i>deck presentation</i>
Rifqi Mufiddin	<ul style="list-style-type: none">- Melakukan <i>query</i> di Big Query- Membuat <i>chart</i> dan <i>filter</i> pada <i>dashboard</i> (<i>month year filter, province filter, island filter, column bar chart</i>)- Membuat penjelasan <i>dashboard</i>- Melakukan <i>styling</i> dan <i>finishing</i> pada <i>dashboard</i>- Melakukan <i>Exploratory Data Analysis</i> (variabel numerik dan kategorik, analisis hubungan)- Melakukan <i>data preprocessing</i>- Membuat model klasifikasi (Logistic Regression dan Random Forest)- Menentukan model terbaik dengan metrik evaluasi- Membuat <i>model interpretation</i>



Thanks!

CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)