



*Matematika - Universitas Pendidikan Indonesia

Data Scientist - Sharing Session

*Slide ini juga digunakan pada webinar lainnya jika inti pokok yang akan dideliver adalah sama.

Said Al Afghani Edsa

Sabtu, 01 Oktober 2022



Said Al Afghani Edsa



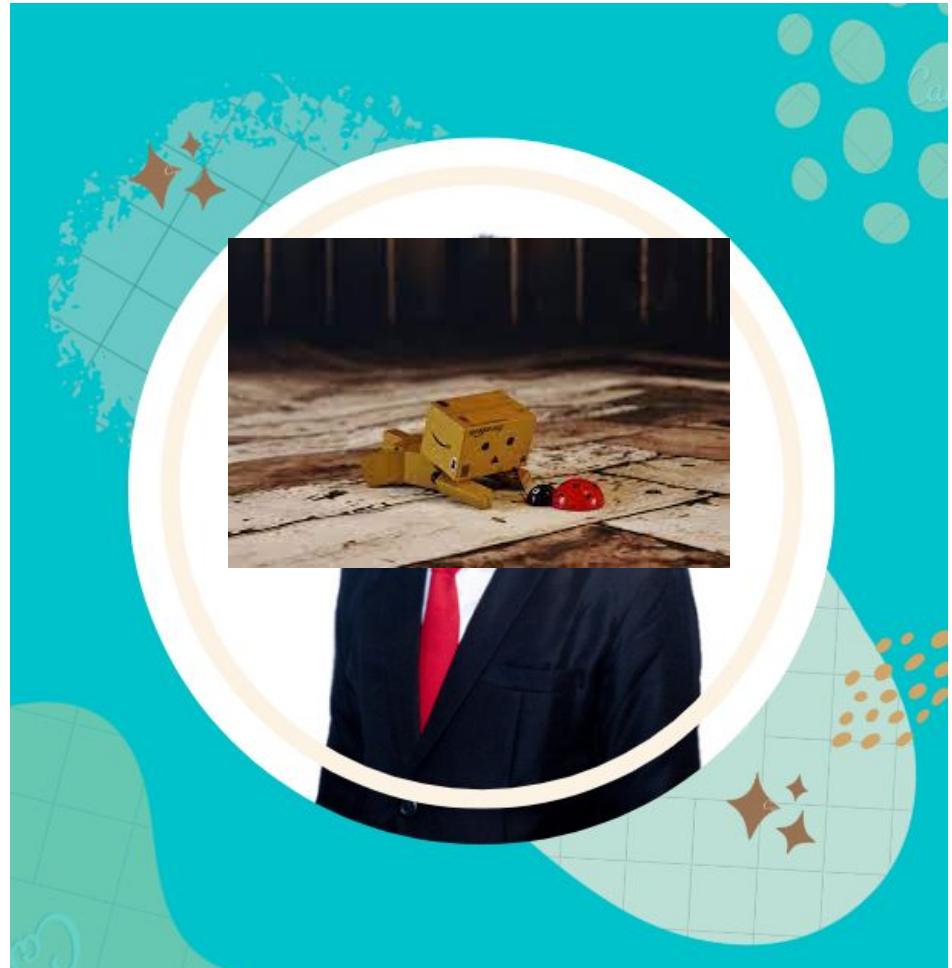
saidalaafghani.dumai@gmail.com



github.com/saiddddd



linkedin.com/in/said-al-afghani-edsa-168243147/

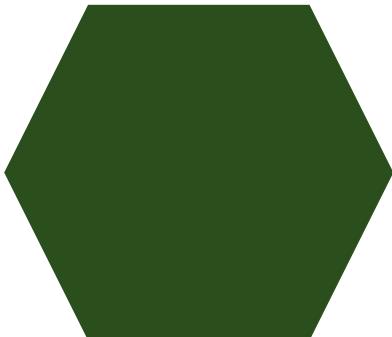




Outline Diskusi



- 1 Matematika dan Data Science
- 2 Kebutuhan bidang Data Science di Indonesia
- 3 Persiapan menjadi Data Scientist
- 4 Peluang menjadi Data Scientist bagi lulusan matematika



01. Matematika dan Data Science

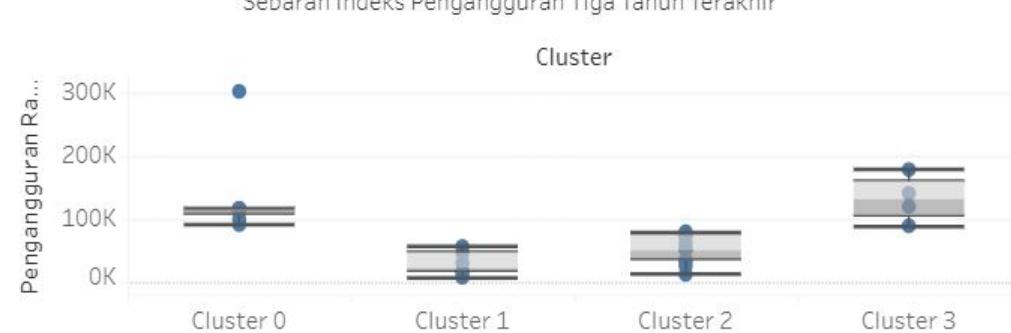
Apa ya yang menarik dari mengelola data?

Dengan data, bisa diperoleh informasi lebih lanjut terkait objek populasi yang akan ditinjau.

Rata - Rata Indeks Kedalaman Kemiskinan dan Rata - Rata Rasio Gini



Sebaran Indeks Pengangguran Tiga Tahun Terakhir



Dapat pula ditemukan pola pada data yang diamati sehingga memudahkan dalam mencari solusi terkait permasalahan yang akan disolusikan.

Penurunan nilai indeks Kedalaman Kemiskinan mengindikasikan bahwa rata-rata pengeluaran penduduk miskin cenderung makin mendekati garis kemiskinan dan ketimpangan pengeluaran penduduk miskin juga semakin menyempit.



Rasio Gini diupayakan mengalami penurunan dengan tujuan bahwa adanya pemerataan distribusi pendapatan antar penduduk

Mengolah data dapat dimulai dari yang paling sederhana, seperti :

1. Melakukan pembersihan data/data cleansing
2. Meninjau statistika deskriptif
3. Melakukan visualisasi/Exploratory Data Analysis

Total Nasabah baru
per Januari 2022 - Maret 2022 :
2.293.313

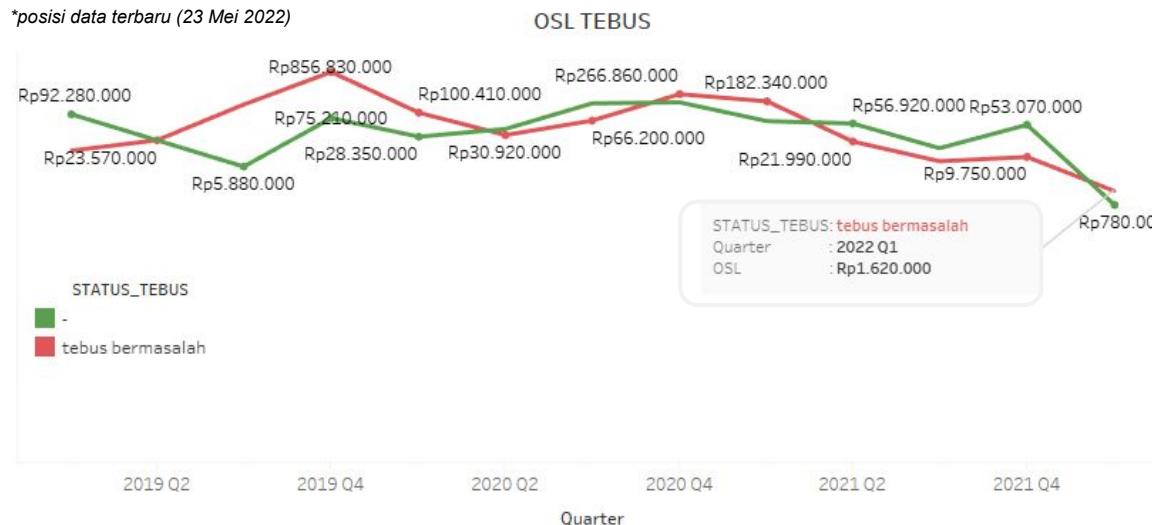
Total Kredit :
3.170.812

Kredit Aktif :
3.111.257

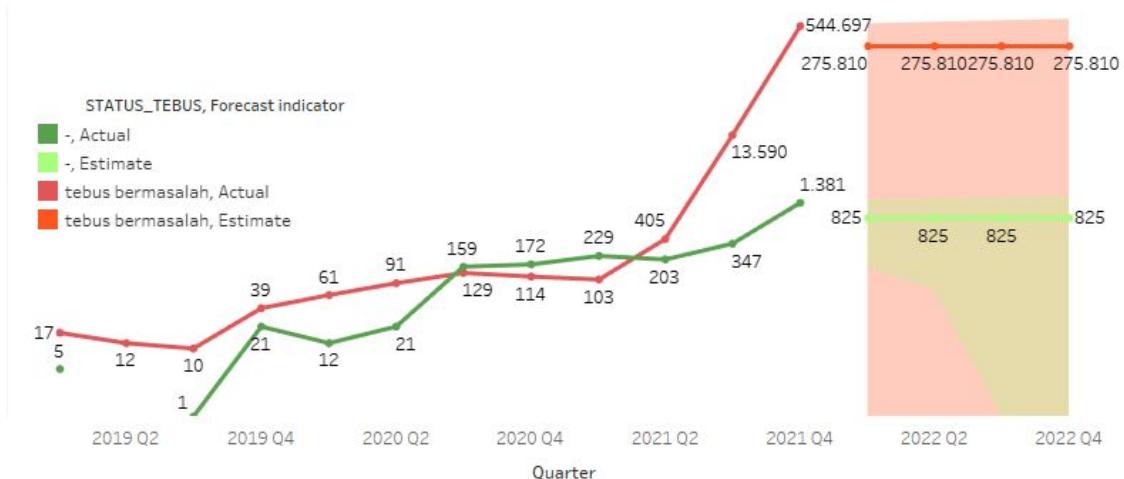
Kredit Lunas :
30.912

Note : Untuk keperluan ilustrasi

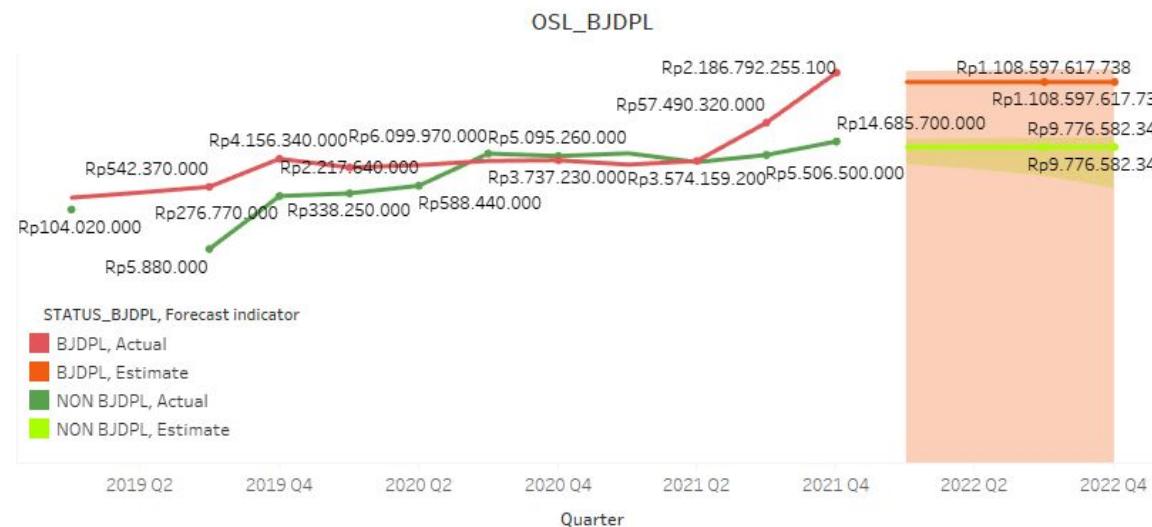
Trend Produk Gadai (All)



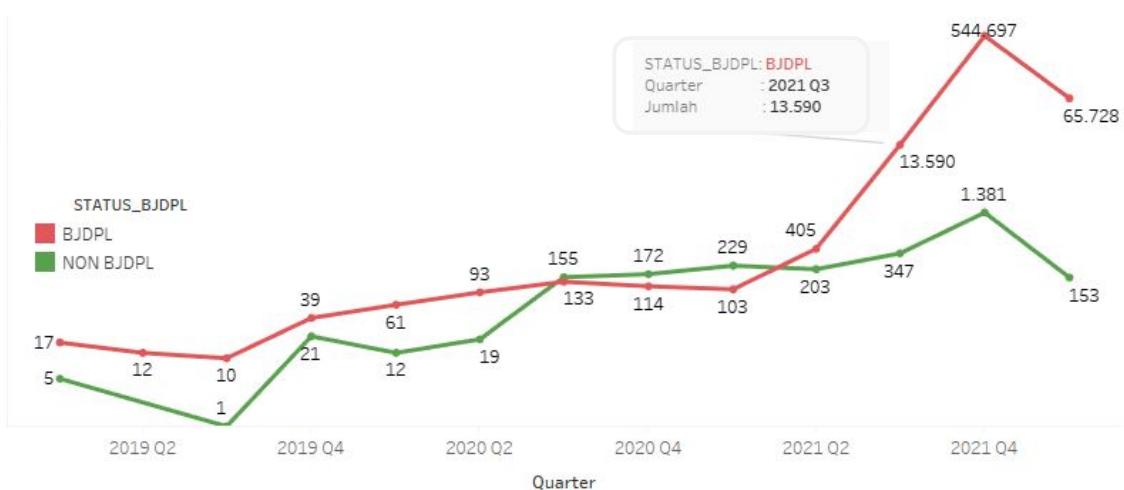
Trend Tebus Kredit Gadai



*data telah dilakukan transformasi
(untuk keperluan melihat trend)

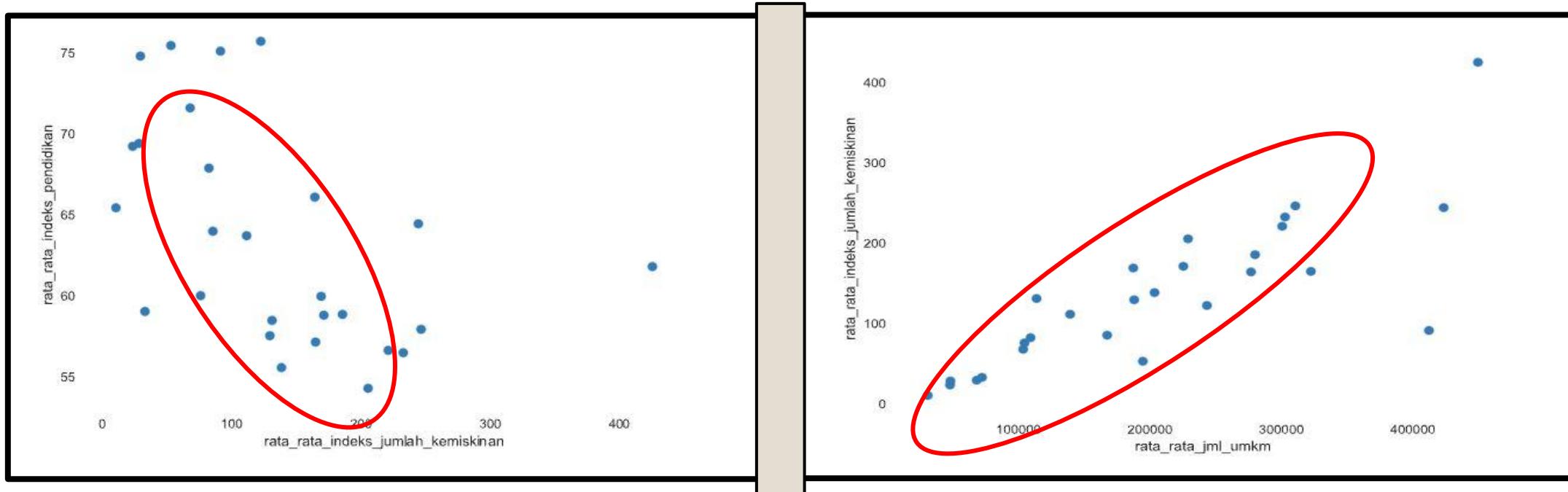


Trend BJDPL



i

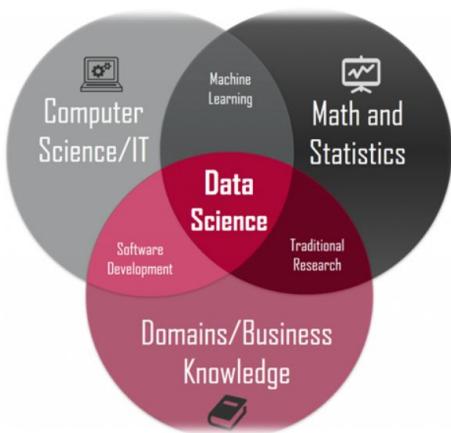
Tinjauan Korelasi



Dari paparan di atas dapat diamati bahwa untuk menekan jumlah angka kemiskinan dapat diupayakan dengan meningkatkan taraf masyarakat/meningkatkan pembangunan masyarakat. Di samping itu perlu pula dilakukan pemberdayaan terhadap UMKM yang telah ada dengan menyesuaikan terhadap kondisi covid19 yang tengah berlangsung, pemberdayaan dapat dilakukan baik melalui pengurangan jumlah UMKM atau meningkatkan kinerja UMKM yang telah ada.

Suatu keilmuan yang merupakan irisan dari bidang matematika, ilmu komputer, dan statistika dengan tujuan untuk melakukan investigasi pada suatu himpunan data dengan menggunakan prinsip data mining, mendapatkan pola, maupun melakukan prediksi pada objek kajian pada himpunan data yang sedang ditinjau.

Pengertian - Data Science



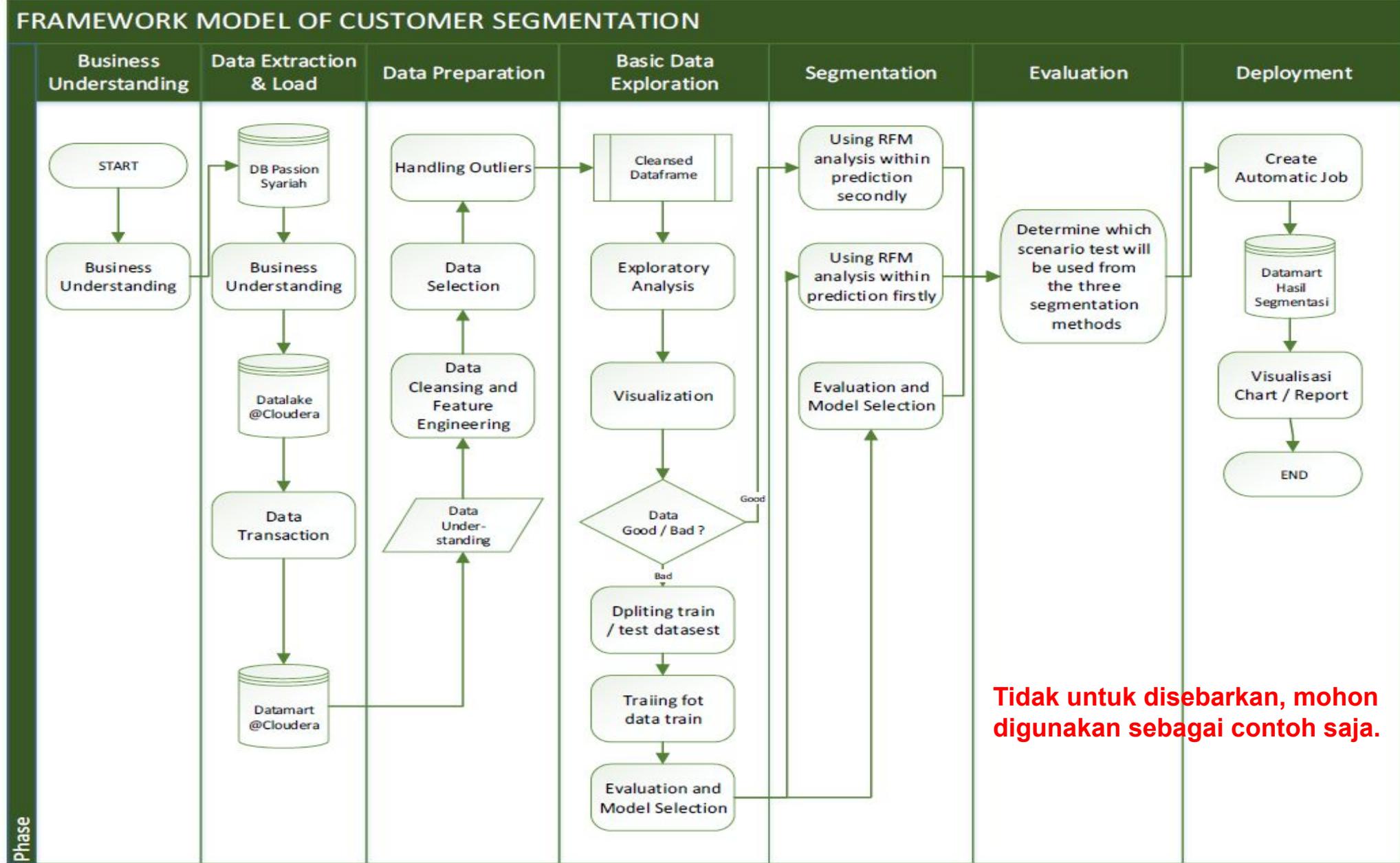
Data Science Framework : CRISP - DM

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<p>Determine Business Objectives <i>Background</i> <i>Business Objectives</i> <i>Business Success Criteria</i></p> <p>Assess Situation <i>Inventory of Resources Requirements, Assumptions, and Constraints</i> <i>Risks and Contingencies</i> <i>Terminology</i> <i>Costs and Benefits</i></p> <p>Determine Data Mining Goals <i>Data Mining Goals</i> <i>Data Mining Success Criteria</i></p> <p>Produce Project Plan <i>Project Plan</i> <i>Initial Assessment of Tools and Techniques</i></p>	<p>Collect Initial Data <i>Initial Data Collection Report</i></p> <p>Describe Data <i>Data Description Report</i></p> <p>Explore Data <i>Data Exploration Report</i></p> <p>Verify Data Quality <i>Data Quality Report</i></p>	<p>Select Data <i>Rationale for Inclusion/Exclusion</i></p> <p>Clean Data <i>Data Cleaning Report</i></p> <p>Construct Data <i>Derived Attributes</i> <i>Generated Records</i></p> <p>Integrate Data <i>Merged Data</i></p> <p>Format Data <i>Reformatted Data</i></p> <p>Dataset <i>Dataset Description</i></p>	<p>Select Modeling Techniques <i>Modeling Technique</i> <i>Modeling Assumptions</i></p> <p>Generate Test Design <i>Test Design</i></p> <p>Build Model <i>Parameter Settings</i> <i>Models</i> <i>Model Descriptions</i></p> <p>Assess Model <i>Model Assessment</i> <i>Revised Parameter Settings</i></p>	<p>Evaluate Results <i>Assessment of Data Mining Results w.r.t. Business Success Criteria</i> <i>Approved Models</i></p> <p>Review Process <i>Review of Process</i></p> <p>Determine Next Steps <i>List of Possible Actions</i> <i>Decision</i></p>	<p>Plan Deployment <i>Deployment Plan</i></p> <p>Plan Monitoring and Maintenance <i>Monitoring and Maintenance Plan</i></p> <p>Produce Final Report <i>Final Report</i> <i>Final Presentation</i></p> <p>Review Project <i>Experience Documentation</i></p>

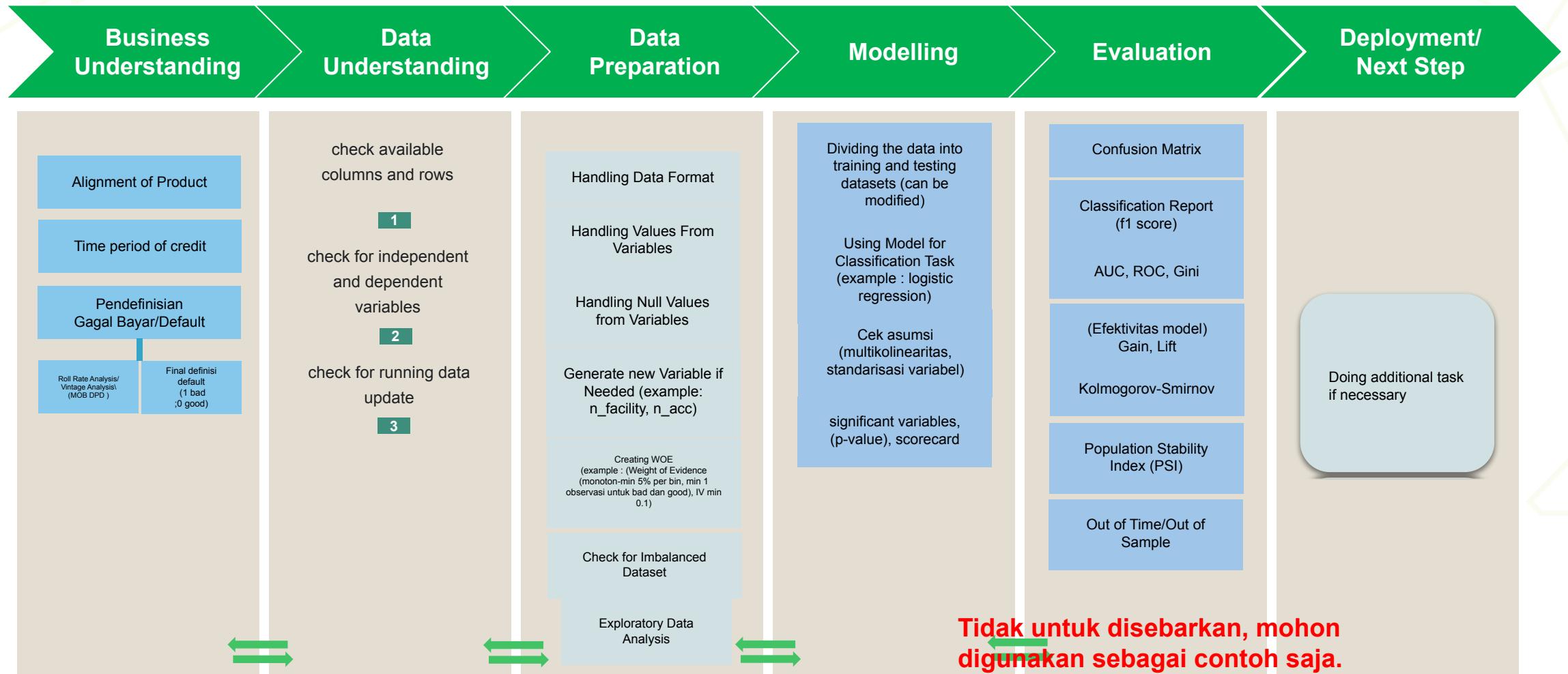
Note



1. CRISP - DM (Cross Industry Standard Process for Data Mining)
2. Poin utama dari data science ialah mampu meningkatkan nilai bisnis atau risiko/operasional perusahaan
3. Pada praktiknya tidak semua project data science memerlukan teknik pemodelan yang "advanced"
4. Perlu dipertimbangkan tahap penggunaan model (biasanya disebut tahap production) dan bentuk dari evaluasi model



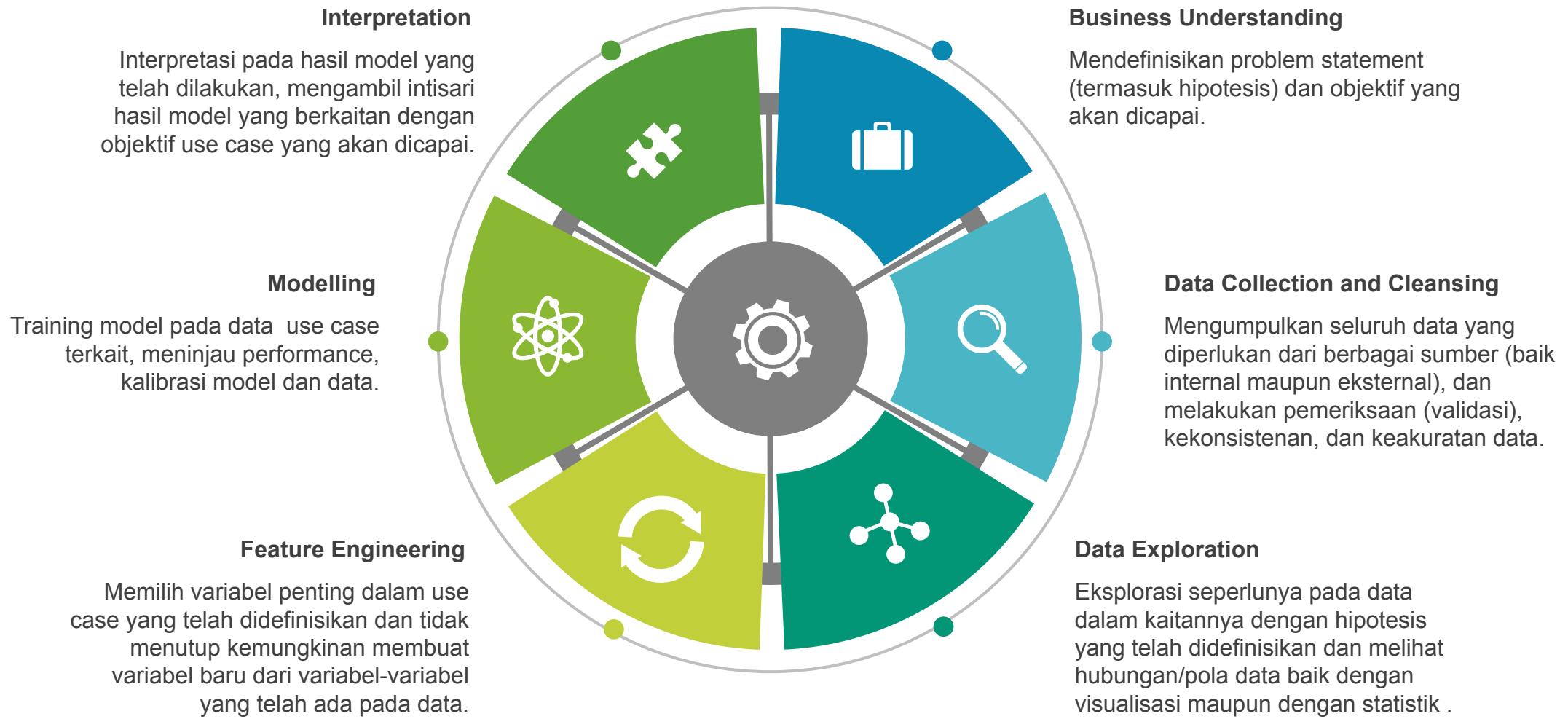
Process Flow of The Task

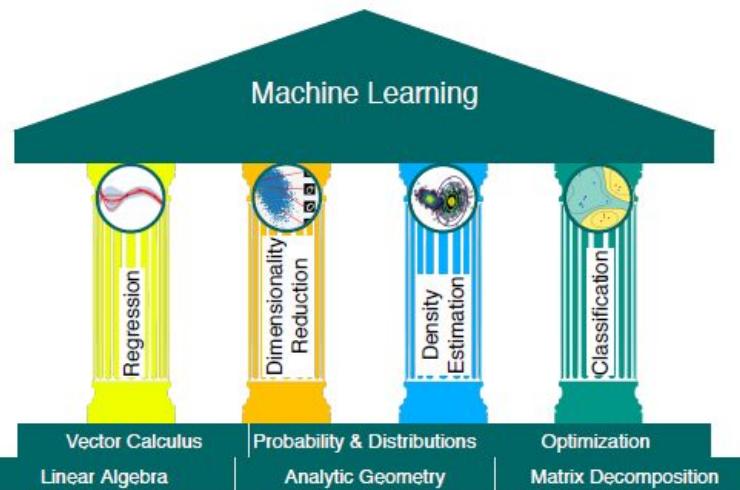


↔ : updating part if necessary

Output :
Score dari masing-masing nilai kategori variabel

DATA SCIENCE LIFECYCLE





Secara praktis, machine learning ialah suatu bentuk model dengan proses learning tanpa diprogram secara eksplisit.

Bahan baku : Matematika, Statistika, dan Pemrograman

Supervised Machine Learning

Predictive machine learning and forecasting

contoh : memprediksi indeks prestasi mahasiswa, memprediksi harga emas

Unsupervised Machine Learning

Penarikan dan pengambilan keputusan berdasarkan pola dari data

contoh : customer segmentation untuk reward program pada nasabah, segmentasi kondisi tertentu untuk efisiensi penanganan masalah sampah di suatu kota di Indonesia

Semi Supervised Machine Learning

Hybrid model dengan data terlabel dan tidak terlabel

pemanfaatan penggunaan supervised dan unsupervised learning model

Note



1. Ada 1 lagi jenis machine learning yaitu reinforcement learning yang menggunakan prinsip reward, data belajar dari proses kesalahan yang dimiliki oleh model
2. Terdapat banyak model machine learning, namun cobalah dimulai dari yang relatif sederhana untuk mendapatkan "feels" prinsip dari model machine learning
3. Tidak semua project data science menggunakan metode yang "canggih", namun aktivitas analisa/investigasi project data science selalu dimulai dengan diskusi dengan user/client bisnis/operasional kita
4. Contoh penggunaan unsupervised machine learning : [Contoh](#)



Pada kasus data customer, dapat dilakukan tinjauan pada kelompok nasabah/customer terkait mulainya diakuisisi



02.

Kebutuhan bidang Data Science di Indonesia

BERITA BISNIS

Era Digitalisasi, Kebutuhan SDM Data Science dan Kecerdasan Buatan Melonjak

Kamis, 15 Juli 2021 | 08:09 WIB



Perkembangan pembelajaran dan kebutuhan terkait profesi data terus berkembang mengalami "update" baik dari sisi penggunaan metode maupun pengembangan metode itu sendiri

Sumber :

1. <https://teknologi.bisnis.com/read/20210718/84/1419129/analisis-data-profesi-yang-dibutuhkan-untuk-kebutuhan-digitalisasi>
2. <https://insight.kontan.co.id/news/era-digitalisasi-kebutuhan-sdm-data-science-dan-kecerdasan-buatan-melonjak>
3. <https://dqlab.id/prospek-kerja-bagi-seorang-data-scientist-di-indonesia>

lock teknologi.bisnis.com/read/20210718/84/1419129/analisis-data-profesi-yang-dibutuhkan-untuk-kebutuhan-digitalisasi

BUSINESS NEWS

8 Maskapai Penerbangan Indonesia yang Bangkrut dan Setop Operasi Resmi

Menu Cari berita...

Lanjut Kuliah Harus Ngerti Bahasa Inggris Langkah Kerja

GRATIS!

Bisnis.com

Metode Belajar Bahasa Inggris Dengan Materi

Cek Kemampuan Bahasa Inggris-mu!

English Academy

Home • Teknologi • Sains

Analisis Data, Profesi yang Dibutuhkan untuk Kebutuhan Digitalisasi

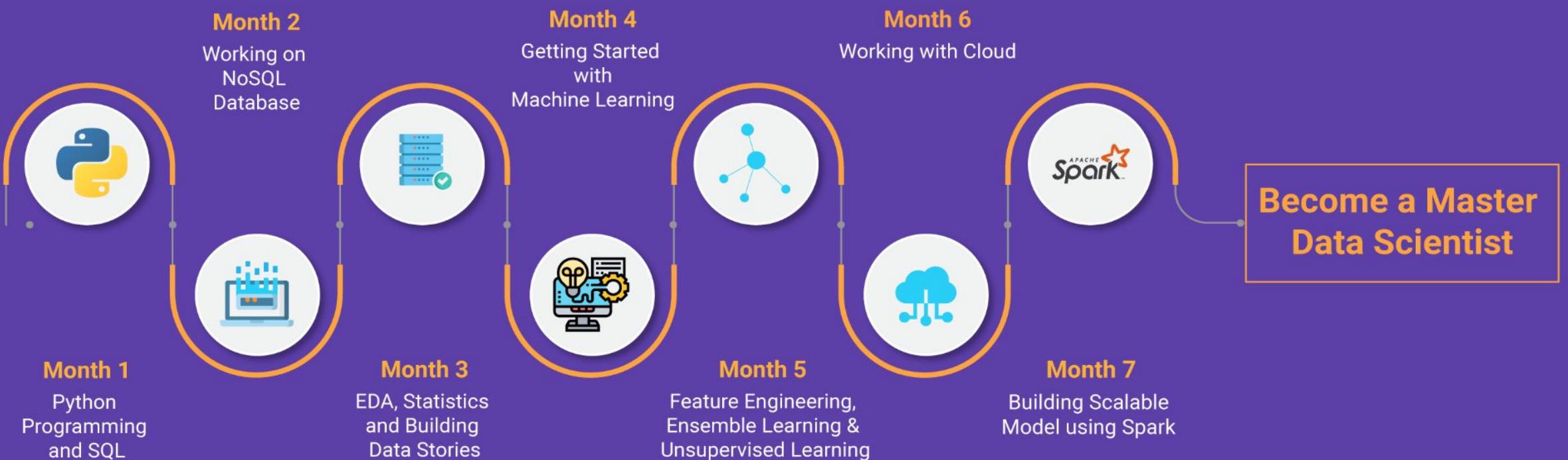
Bidang ilmu data berfokus pada analisis data untuk memberikan wawasan dan kemudian menggunakan algoritma dan pembelajaran mesin untuk membuat keputusan dan prediksi yang tepat.

03.

Persiapan menjadi Data Scientist

Bagaimana Cara Memulai?

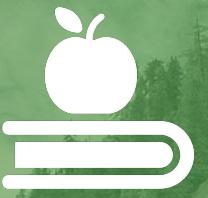
Dimulai dengan melakukan hal yang sederhana (eksplorasi dan memulai portofolio)



Memulai dengan...



**Menentukan
Tujuan**



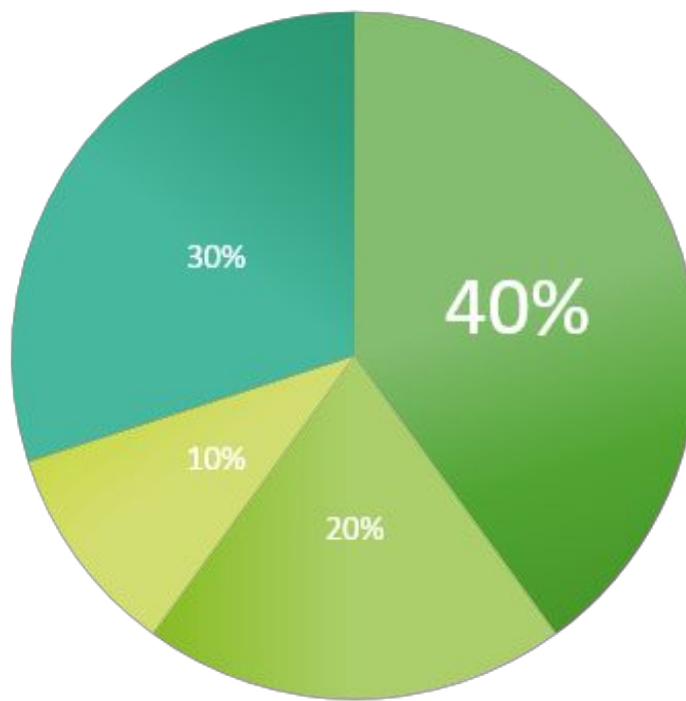
**Learning By
Doing**



Mengatur waktu

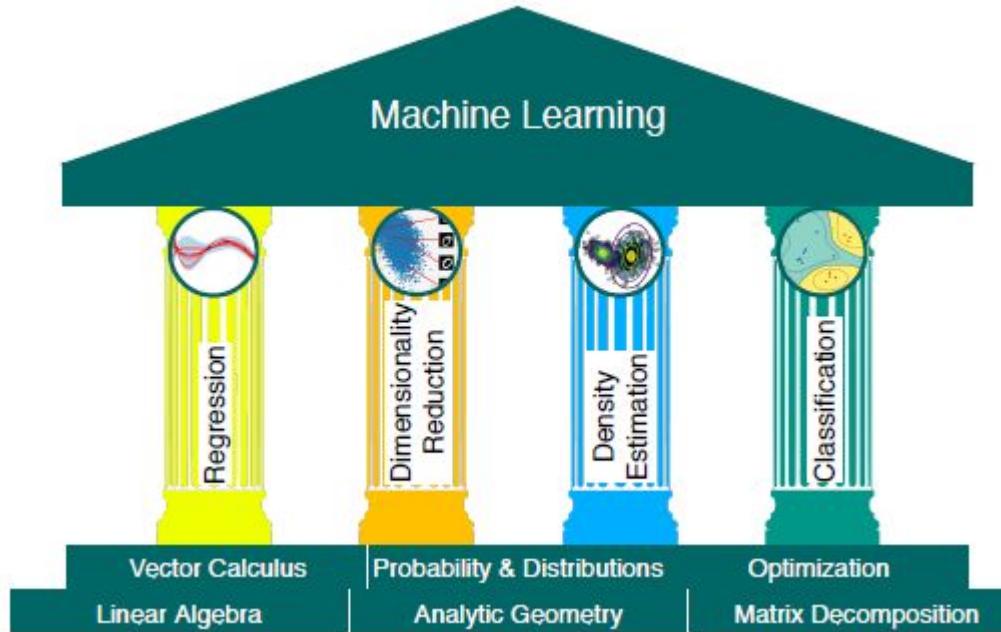


**Review dan
Arsip**



04.

Peluang menjadi Data Scientist bagi Matematikawan



Peluang terbuka lebar bagi matematika untuk data science terutama dalam pengembangan model machine learning

Salah satu konsep yang kita lihat tadi ialah jarak, khususnya jarak euclid untuk membentuk cluster.

link for study : <https://mml-book.github.io/>

Belajar dapat dimana saja dan dengan siapa...



E-book, buku, artikel (medium), dll



Platform pembelajaran:

1. DQLab : [DQLab](#)
2. Kaggle : [Kaggle](#)
3. BigQueryML : [BigQueryML](#)
4. MySql : <https://www.w3schools.com/>

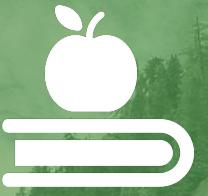


Practice and Practice!

Memulai dengan...



**Menentukan
Tujuan**



**Learning By
Doing**



Mengatur waktu



**Review dan
Arsip**

A photograph of a person sitting cross-legged on a light-colored bedsheet. They are wearing light green jeans, a white short-sleeved shirt, and a black vest with fringe details. They are using a silver laptop. The word "Lampiran" is overlaid in large, bold, black font across the center of the image.

Lampiran

Contoh 1 :
Penggunaan data
pemerintah

Data pemerintah yang bersifat open source, misalnya pada **Open Data Jabar** dapat digunakan baik untuk tinjauan ilmiah/penelitian ataupun keperluan lain yang menunjang baik kegiatan bisnis maupun penentuan strategi kebijakan

Analisis Potensi Wilayah Dengan Clustering K-Means di Jawa Barat Untuk Pemetaan Kemiskinan dan Strategi Penanggulangannya di Masa Pandemik



Topik Sosial Ekonomi
Kategori Umum

NamaTimnyaApa?

@NamaTimnyaApa

Contoh Use Case

1

Investigasi Wilayah Kabupaten/Kota dalam melakukan pemetaan strategi sesuai tujuan jangka panjang/pendek

2

Penentuan cluster wilayah prioritas dalam penanggulangan angka kemiskinan dan angka kasus covid

3

Tinjauan potensi wilayah kabupaten/kota dalam melakukan perluasan (atau lainnya) terkait UMKM dan indeks gini rasio



Meningkatnya Angka Pengangguran



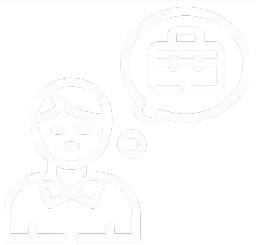
Jumlah Penduduk Bekerja
± 388 Ribu



Jumlah Pengangguran
± 603 Ribu



Tingkat Pengangguran Terbuka
± 2,42



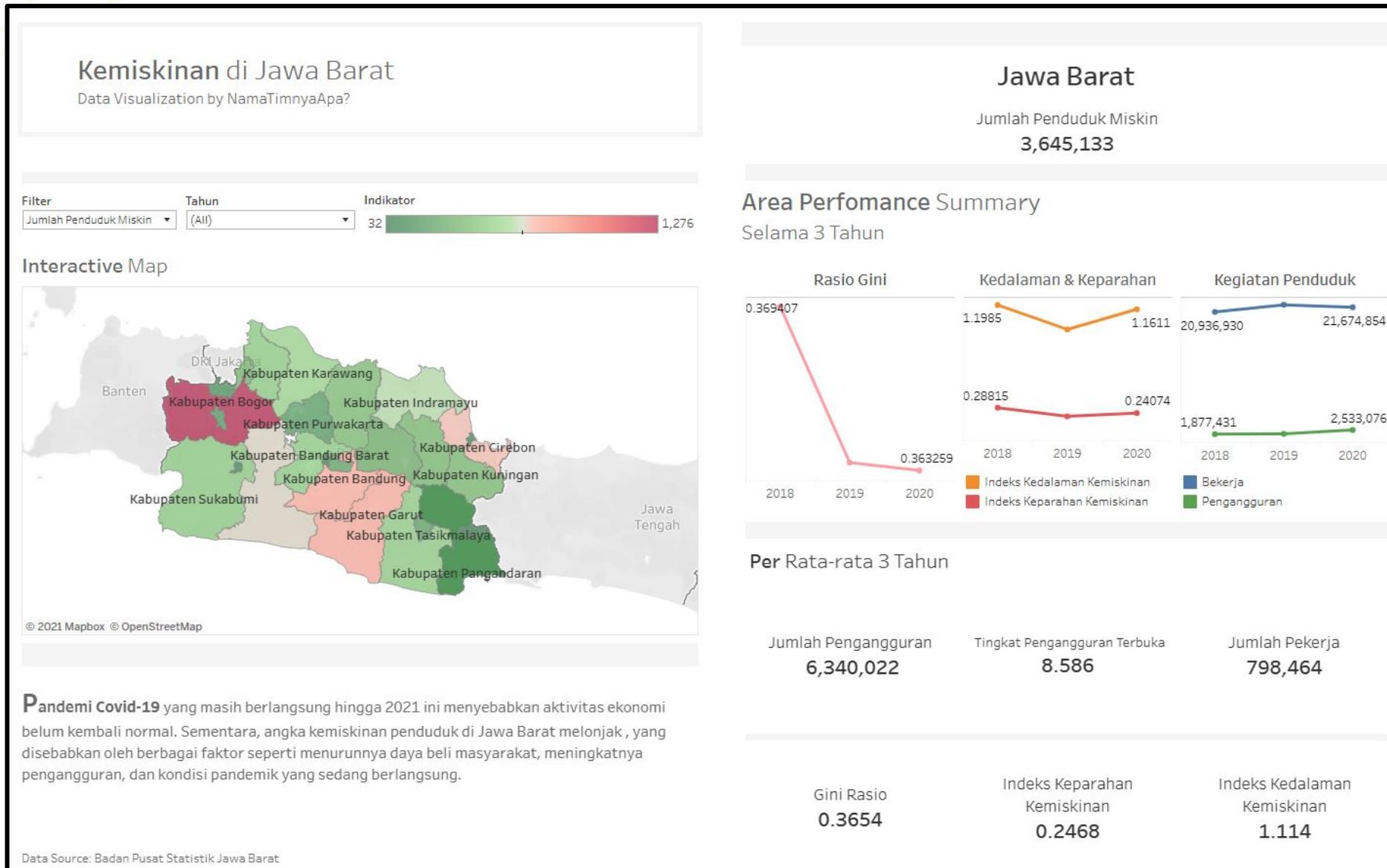
Kenaikan tingkat pengangguran dan menurunnya jumlah penduduk yang bekerja juga cukup signifikan yang diakibatkan Penurunan kinerja sektor-sektor ekonomi yang berimbas pada menurunnya kesempatan kerja dan terjadinya PHK.

Variabel yang berkorelasi dengan Covid-19

	Vaksinasi1	Vaksinasi2	Vaksinasi1_%	Vaksinasi2_%
konfirmasi_total	0.72931019	0.709726041	-0.241595522	0.709726041
konfirmasi_sembuh	0.730726402	0.730726402	0.730726402	0.730726402
konfirmasi_meninggal	0.577743813	0.577743813	0.577743813	0.577743813
konfirmasi_aktif	0.29251015	0.29251015	0.29251015	0.29251015
konfirmasi_total_daily_growth	0.322114116	0.322114116	0.322114116	0.322114116
konfirmasi_sembuh_daily_growth	0.18041561	0.18041561	0.18041561	0.18041561
konfirmasi_meninggal_daily_growth	0.039022545	0.039022545	0.039022545	0.039022545
konfirmasi_aktif_daily_growth	-0.124708017	-0.124708017	-0.124708017	-0.124708017

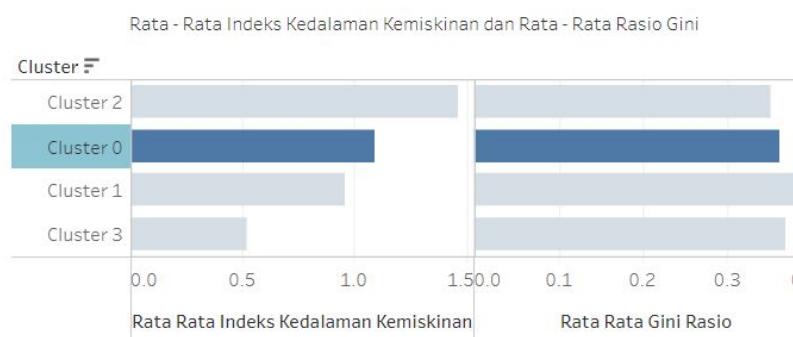
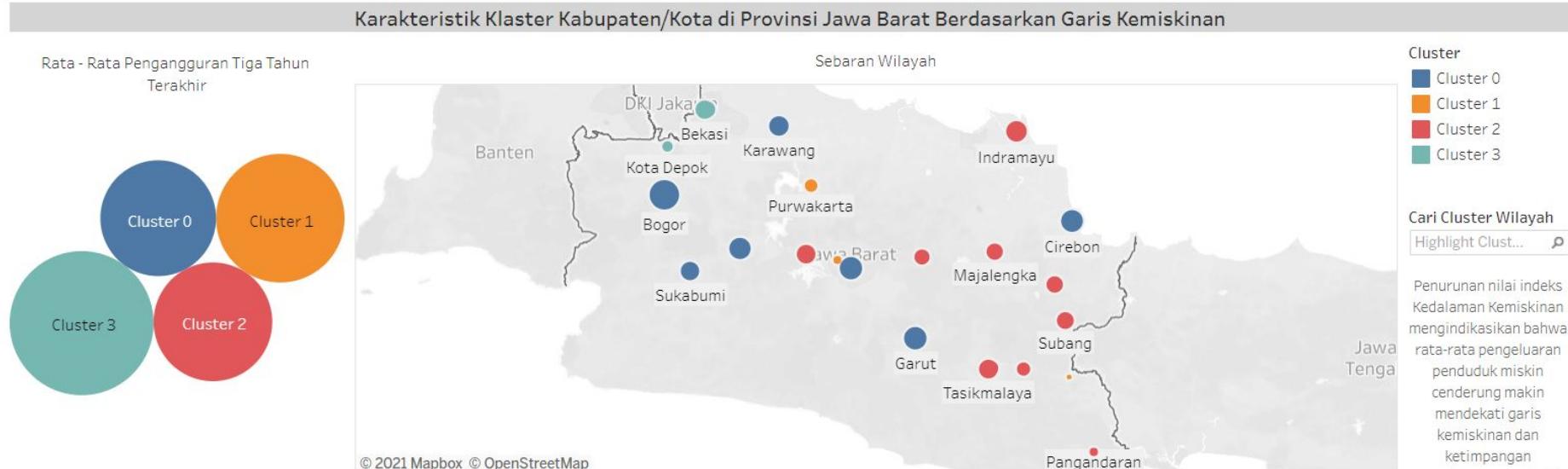
- Total penduduk Jawa Barat yang telah menerima vaksin dengan konfirmasi total daily growth tidak terlalu berkorelasi.
- Adanya Vaksinasi sangat berpengaruh kepada konfirmasi penduduk yang sembuh ataupun meninggal akibat Covid-19.
- Persentase kenaikan kasus sembuh akibat Covid-19 Meningkat seiring dengan banyaknya penduduk yang mendapatkan Vaksin.

Exploratory Data - General



[Kemiskinan di Jawa Barat | Tableau Public](#)

Klasterisasi Wilayah



Rasio Gini diupayakan mengalami penurunan dengan tujuan bahwa adanya pemerataan distribusi pendapatan antar penduduk

A photograph of a person sitting cross-legged on a light-colored bedsheet. They are wearing a white long-sleeved shirt, light green jeans, and a black vest with white fringe. They are looking down at a silver laptop keyboard. Their hands are visible; the left hand has a gold ring on the ring finger and a silver bracelet, while the right hand wears a brown leather watch. The background is a plain, light-colored wall.

Lampiran

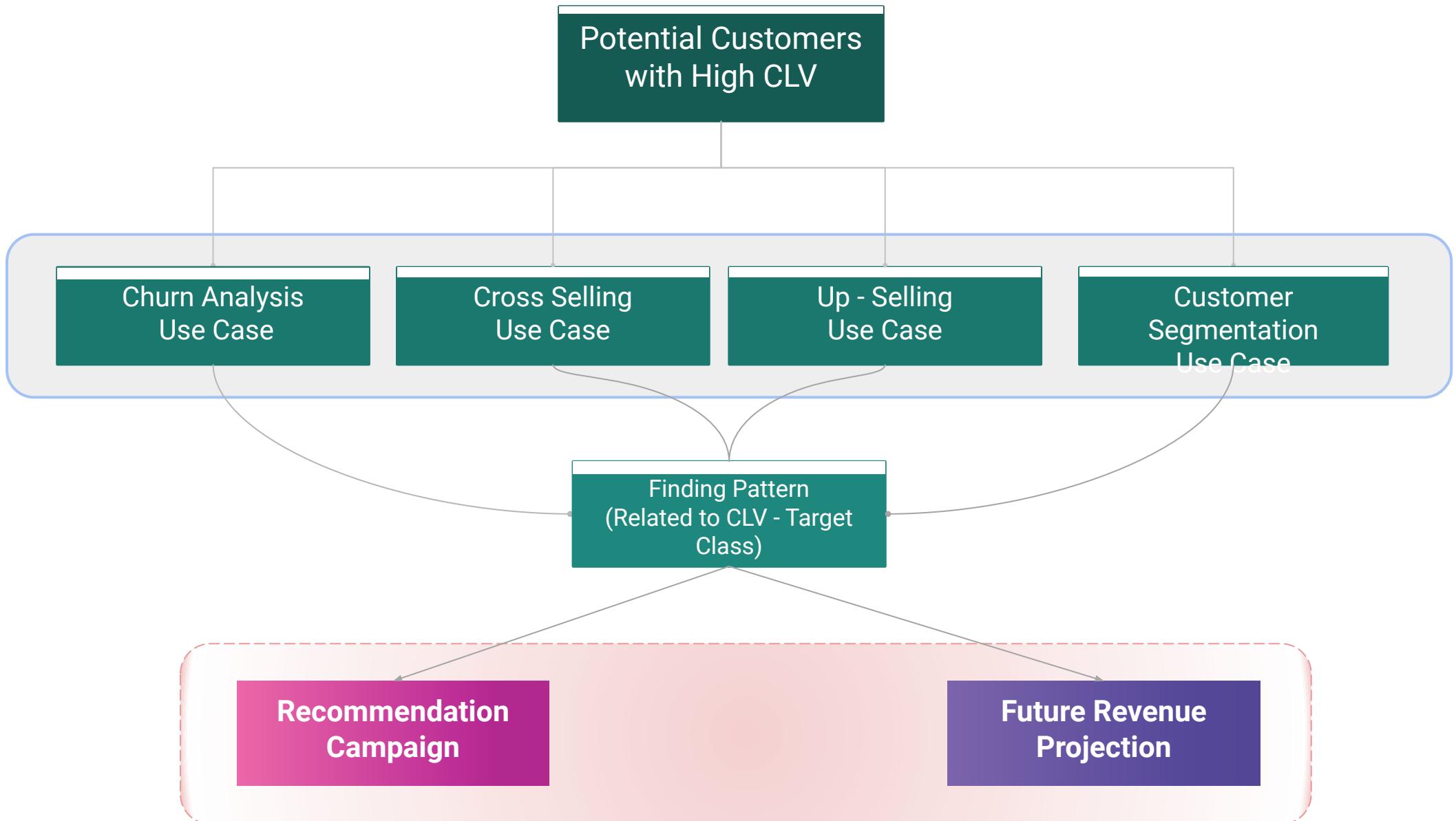
Contoh 2 :
Penggunaan data
nasabah

Customer Lifetime Value (CLV)

“CLV merupakan salah satu parameter alat bantu pembentukan strategi, bukan strategi untuk meningkatkan keuntungan.”



Campaign - CLV Application

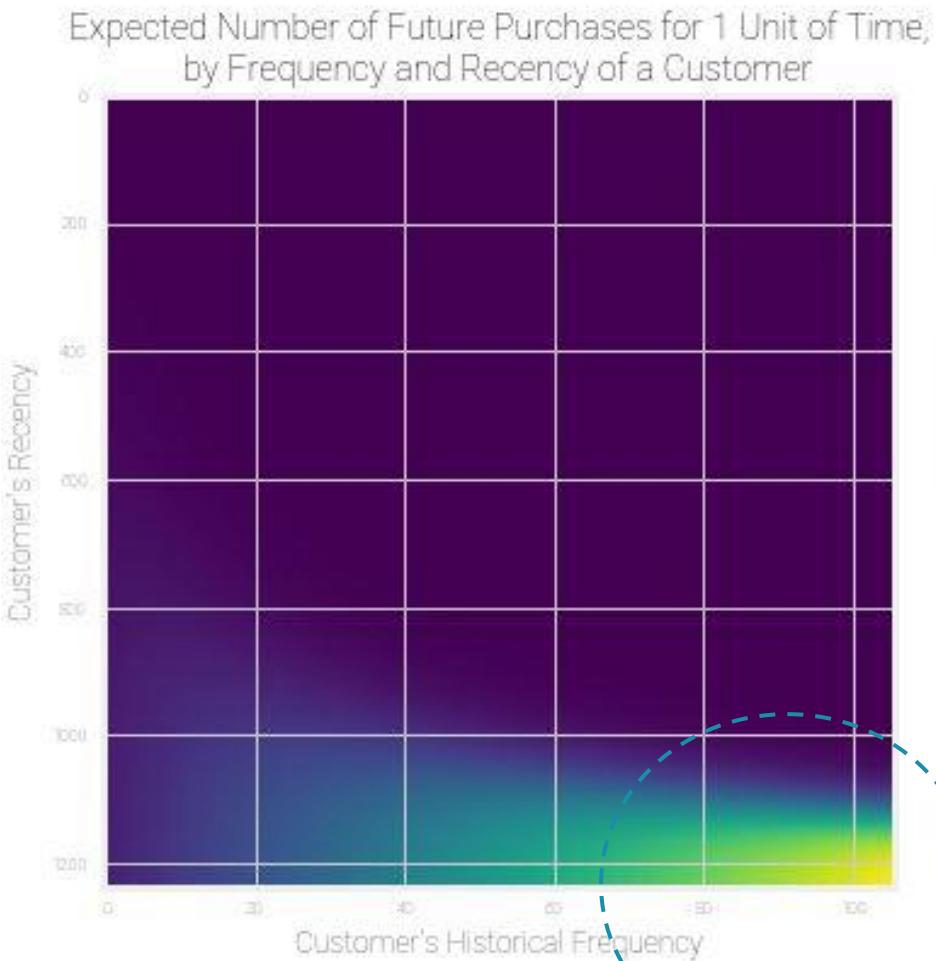


Model yang digunakan merupakan model probabilistik yaitu Negative Binomial Distribution dan Clustering (K-Means dan Gaussian Mixture Models)

Bahan bacaan : <https://benalexkeen.com/bg-nbd-model-for-customer-base-analysis-in-python/>

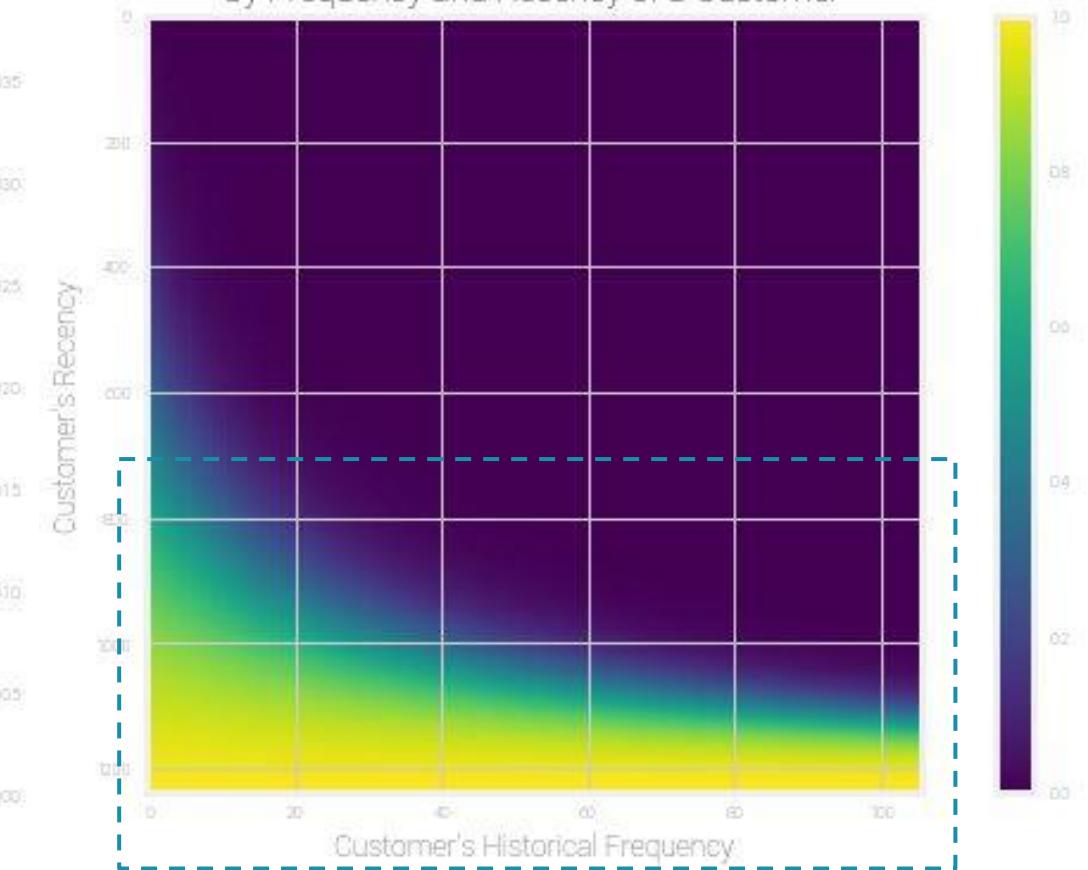


BIG PICTURE OF CUSTOMER LIFETIME VALUE - PRODUK NON ANGSURAN



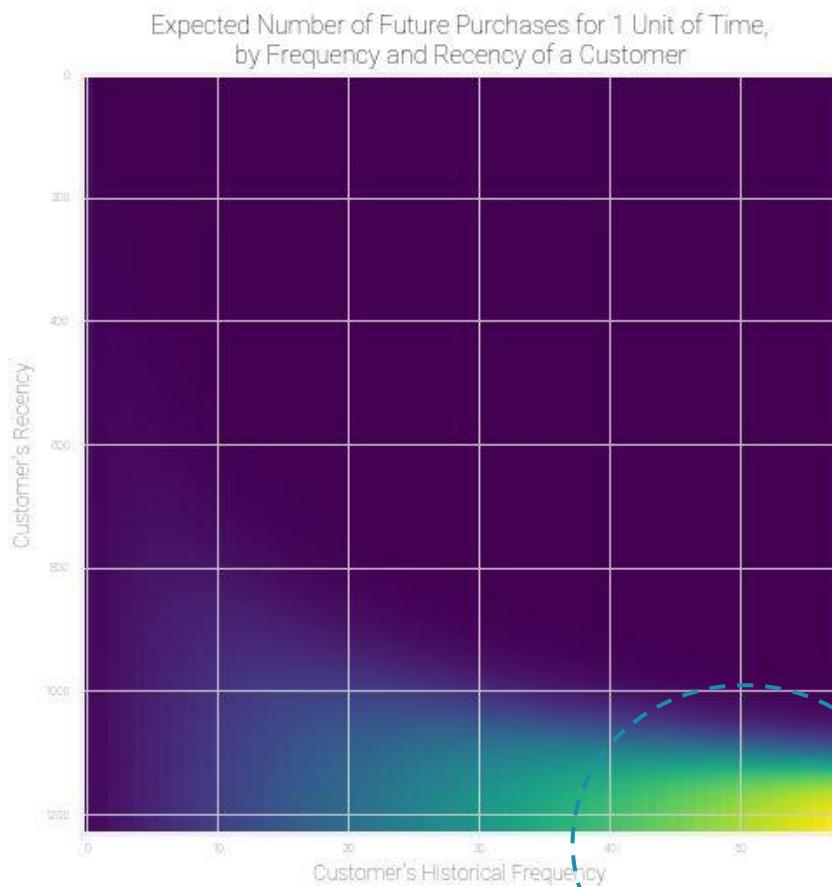
Dari grafik probability di atas dapat kita lihat bahwa pada produk non angsuran kecendrungan nasabah untuk melakukan transaksi kembali akan tinggi bila mana nasabah tersebut memiliki jumlah transaksi yang tinggi namun secara historical sudah lama jarang bertransaksi.

Probability Customer is Alive,
by Frequency and Recency of a Customer

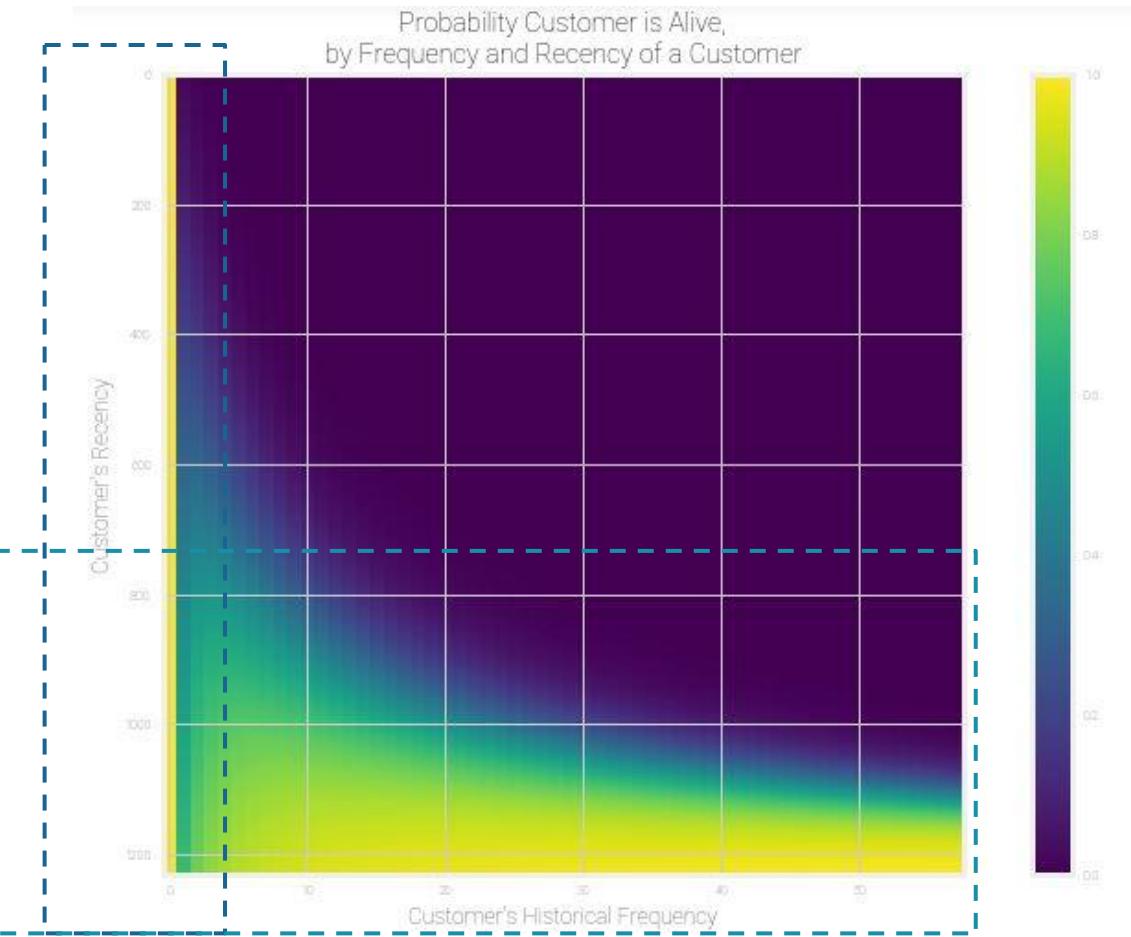


Peluang nasabah pada periode berikutnya masih tetap akan aktif (namun belum tentu bertransaksi) akan tinggi bilamana memiliki umur transaksi yang tinggi dengan kata lain keterbaruan nasabah untuk melakukan transaksi tidak menjamin nasabah tersebut akan aktif pada periode berikutnya.

BIG PICTURE OF CUSTOMER LIFETIME VALUE - PRODUK ANGSURAN

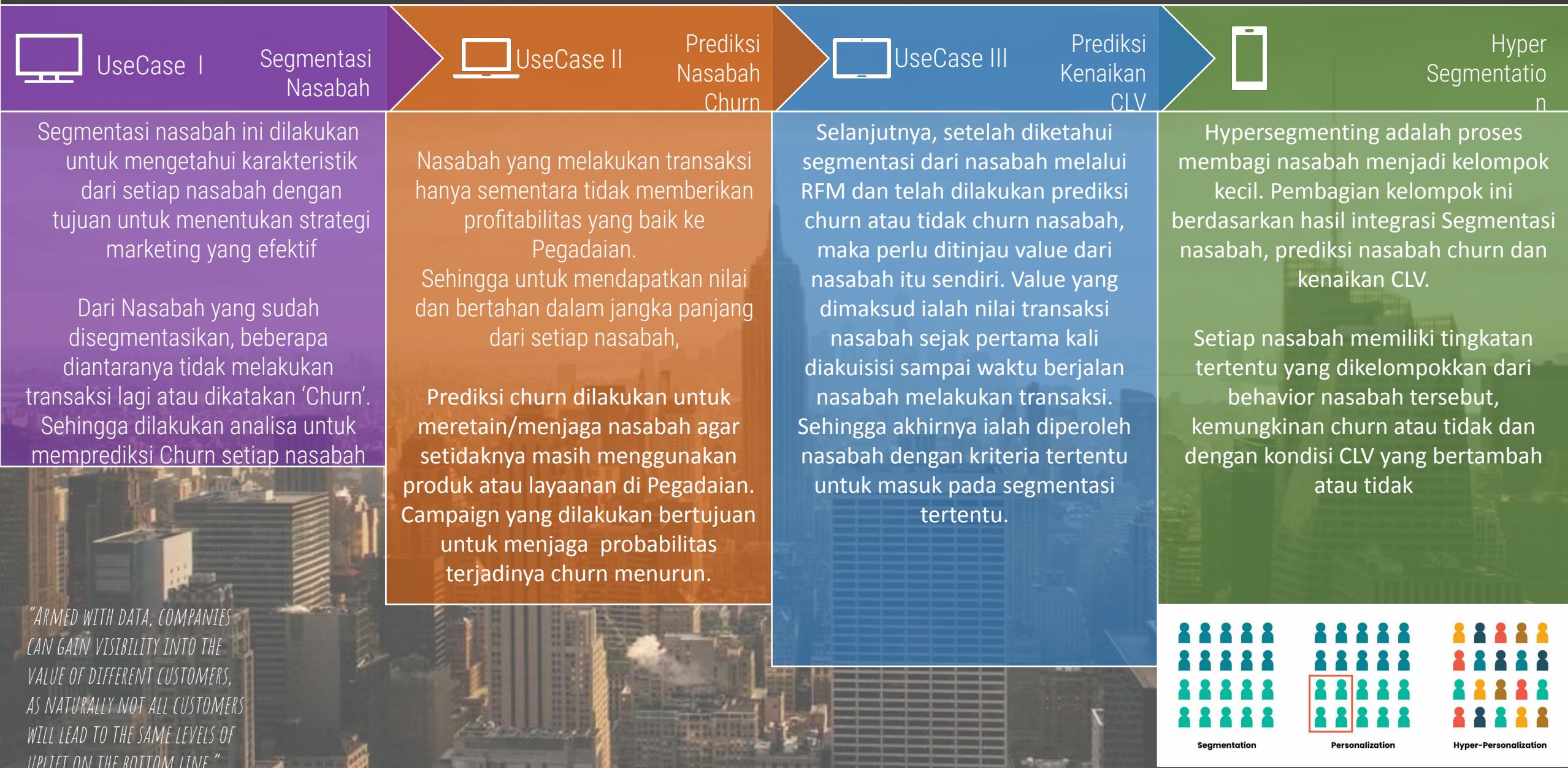


Dari grafik probability di atas dapat kita lihat bahwa pada produk non angsuran kecendrungan nasabah untuk melakukan transaksi kembali akan tinggi bila mana nasabah tersebut memiliki jumlah transaksi yang tinggi namun secara historical sudah lama jarang bertransaksi.



Peluang nasabah pada periode berikutnya masih tetap akan aktif (namun belum tentu bertransaksi) akan tinggi bilamana memiliki umur transaksi yang tinggi dengan kata lain keterbaruan nasabah untuk melakukan transaksi tidak menjamin nasabah tersebut akan aktif pada periode berikutnya namun ada nasabah yang akhir-akhir ini bertransaksi namun frekuensi transaksinya yang rendah memiliki peluang aktif kembali pada periode berikutnya.

Integrasi Use-Case



Pengkategorian

(berdasarkan slide 25)

GROUP 1

PLATINUM



KRITERIA

CUSTOMER TYPE : Best

PREDICTION : Not Churn, Increasing CLV

GROUP 2

GOLD



KRITERIA

CUSTOMER TYPE : Shopper, FirstTime
, dan Valuable

PREDICTION : Not Churn, Decreasing CLV

GROUP 3

SILVER



KRITERIA

CUSTOMER TYPE : Spender and
AboutToSleep

PREDICTION : Churn, Increasing and
Decreasing CLV

GROUP 4

BRONZE



KRITERIA

CUSTOMER TYPE : Uncertain, Frequent

PREDICTION : Churn dan Not Churn, Increasing
dan Decreasing CLV

*penamaan kategori dapat disesuaikan oleh tim pemasaran dengan kondisi eksisting yang sedang berjalan



*Loyalty program
(customer are rewarded with deals and gift)*



Promotion/discount, up sell



Product Promotion



Product promotion

**Treatment diberikan berdasarkan syarat dari masing masing kategori yang dibentuk*





Terima Kasih

