Rakamin Academy X KALBE Nutritionals

Machine Learning Project

Data Scientist

Presented by
Febe Jovita

Project-based Internship *Batch August 2023*

Febe Jovita

About Me

I like to solve problems. Throughout my career, I have been driven by my intellectual curiosity to find answers to the most pressing questions. Whether it's finding a perfect song for a film or finding a rare product for a rocket, I've been able to quickly uncover a customer's pain point and identify a strategic solution.
With my refined skill set, I bring customer-centric mindfulness that enables firms to innovate and thrive. My intellectual curiosity also drives me to be a lifelong learner.

KALBE Nutritionals

Rakamin Academy

My Experience

- Project Based Virtual Intern : UI/UX Designer Nuri X Rakamin Academy Batch May 2023
- Project Based Virtual Intern : Marketing Outreach LifeVitae X Rakamin Academy Batch June 2023
- Project Based Virtual Intern : Product & Business Development Bank Muamalat X Rakamin Academy Batch July 2023

LinkedIn Febe Jovita

About Rakamin Academy

Rakamin Academy is an end-to-end platform for people to start a career in tech, providing a direct-live class with experts from top tech companies, career assessment and counseling, intensive career coaching, a virtual internship, and a job guarantee. **Rakamin Academy** represents high-quality education through the standardization of Curriculum and Pedagogical Approachment in forming Indonesian Talents who have superior quality skills and are adaptive to digital industry qualification standards.

https://www.rakamin.com/

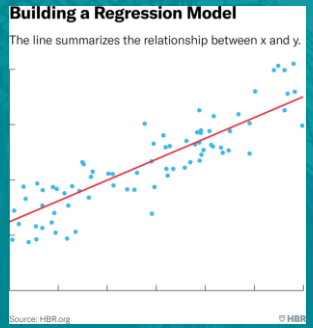
About Kalbe Nutritionals

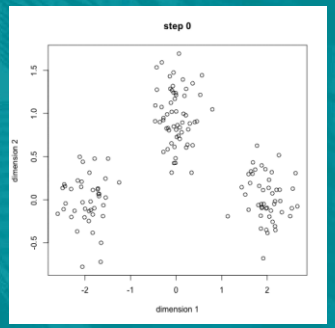
Kalbe Nutritionals (PT Sanghiang Perkasa) merupakan bagian dari PT KALBE Farma, Tbk., salah satu perusahaan farmasi terkemuka, di skala nasional maupun internasional. Pada awal pendiriannya, kami lebih dikenal dengan nama KALBE Farma Health Foods Division. Dalam rangka meningkatkan kinerja perusahaan, serta agar lebih mendekatkan diri kepada konsumen, pada tahun 2007, kami melakukan perubahan corporate brand identity, dan setelah melalui proses yang seksama, saat itu nama **KALBE Nutritionals** terpilih untuk mewakili semangat dan aspirasi kami, sekaligus mempertegas kepercayaan kami sebagai perusahaan bereputasi tinggi.

https://kalbenutritionals.com/

Outline

- Background Story
- Tools
- Challenges
- Result:
  - ✓ Saya melakukan data ingestion ke dalam dbeaver
  - ✓ Saya melakukan exploratory data analysis di dbeaver
  - ✓ Saya melakukan data ingestion ke dalam tableau public
  - ✓ Saya membuat dashboard di tableau
  - ✓ Saya membuat model prediktif menggunakan regresi dan membuat clustering
- Conclusion
- Recommendation
- Glossary
- Appendix





Source

Background Story

Saya adalah seorang Data Scientist di Kalbe Nutritionals dan sedang mendapatkan project baru dari tim inventory dan tim marketing.

Dari tim inventory, saya diminta untuk dapat membantu memprediksi jumlah penjualan (quantity) dari total keseluruhan product Kalbe.

- Tujuan dari project ini adalah untuk mengetahui perkiraan quantity product yang terjual sehingga tim inventory dapat membuat stock persediaan harian yang cukup.
- Prediksi yang dilakukan harus harian.

Dari tim marketing, saya diminta untuk membuat cluster/segment customer berdasarkan beberapa kriteria.

- Tujuan dari project ini adalah untuk membuat segment customer.
- Segment customer ini nantinya akan digunakan oleh tim marketing untuk memberikan personalized promotion dan sales treatment.

KALBE Nutritionals

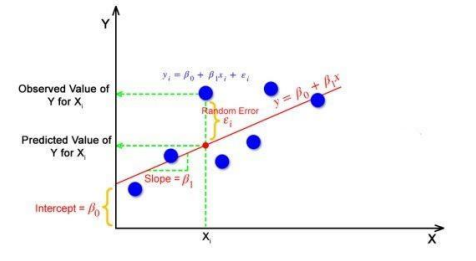
Rakamin Academy

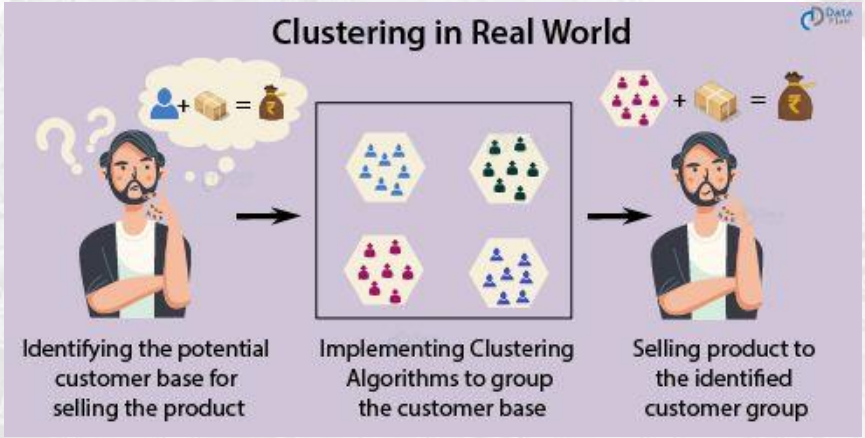
Tools

- Python
- Jupyter Notebook
- Tableau
- Dbeaver
- PostgreSQL
- Anaconda

Challenges

- Peserta dapat melakukan data ingestion ke dalam dbeaver
- Peserta dapat melakukan exploratory data analysis di dbeaver
- Peserta dapat melakukan data ingestion ke dalam tableau public
- Peserta dapat membuat dashboard di tableau
- Peserta dapat membuat model prediktif menggunakan regresi dan membuat clustering

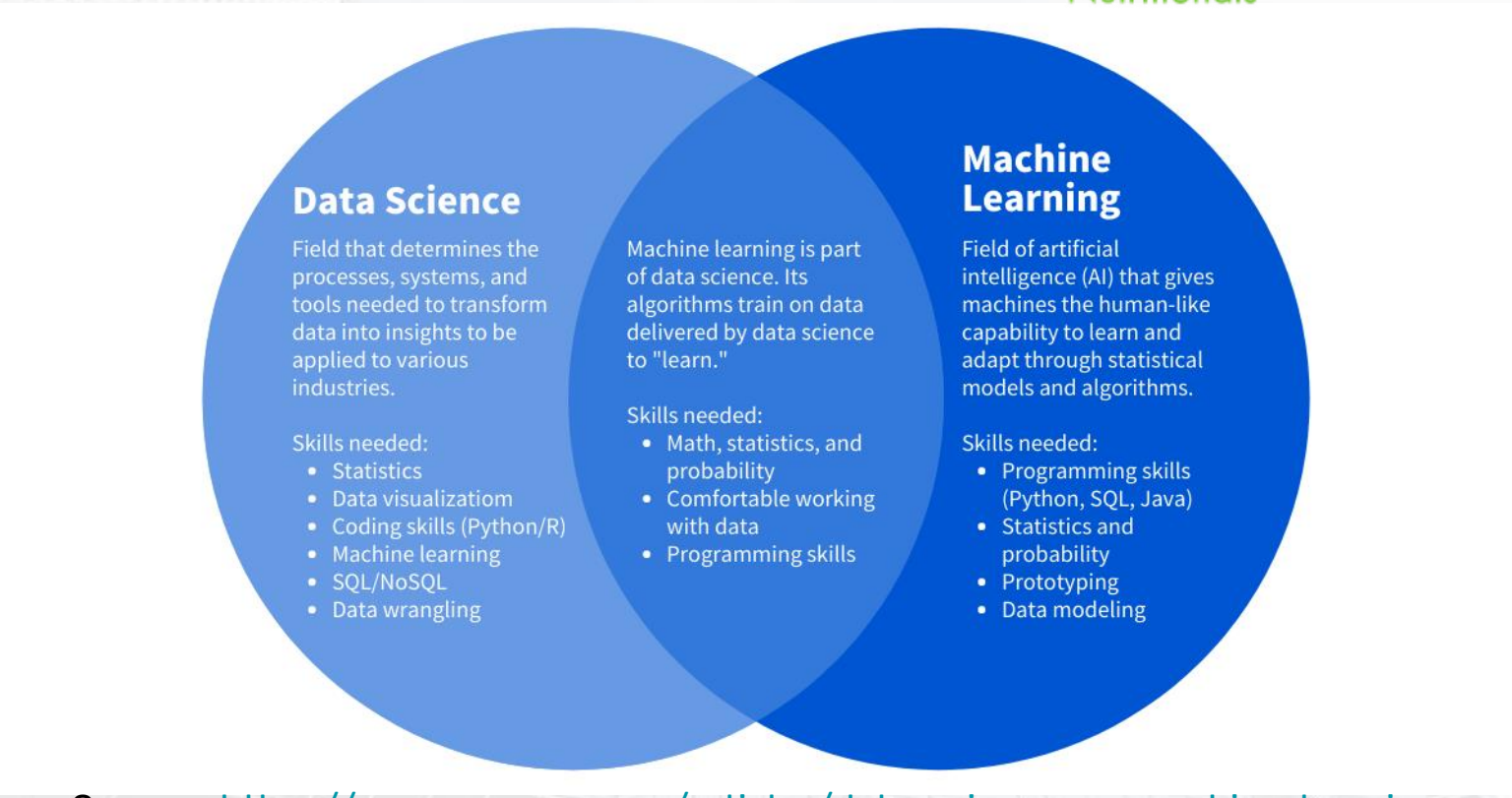




KALBE Nutritionals

Rakamin Academy

Result



Source : https://www.coursera.org/articles/data-science-vs-machine-learning

Case Study - Legend :

KALBE Nutritionals

Rakamin Academy

Dataset ini terdiri dari **4** csv file yaitu customer, store, product dan transaction. Merupakan dummy data untuk studi kasus FMCG dalam kurun waktu 1 tahun yang diambil melalui program membership.

Penjelasan

1. **Customer**
- **CustomerID** : Nomor Unik Customer
- **Age** : Usia Customer
- **Gender** : 0 Wanita, 1 Pria
- **Marital Status** : Married, Single (Belum menikah/Pernah menikah)
- **Income** : Pendapatan per bulan dalam jutaan rupiah

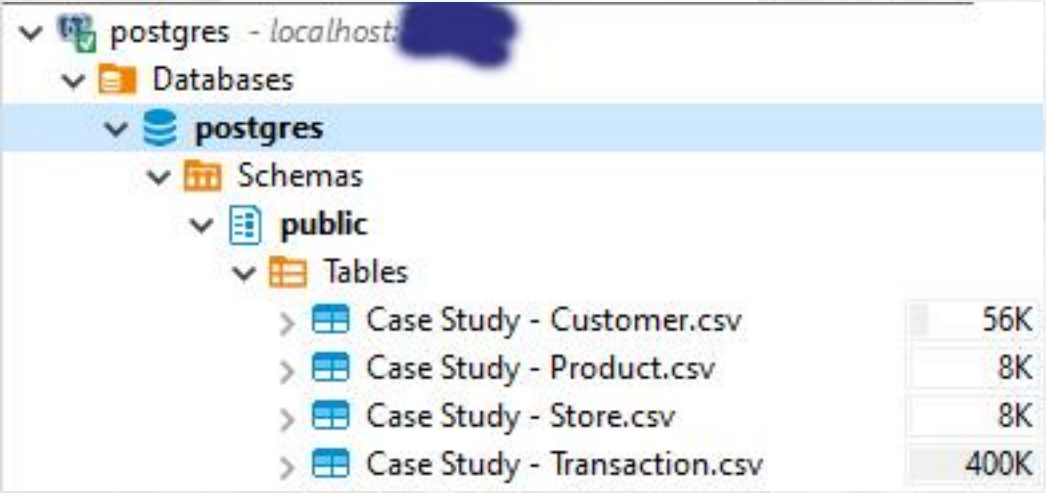
2. **Store**
- **StoreID** : Kode Unik Store
- **StoreName** : Nama Toko
- **GroupStore** : Nama group
- **Type** : Modern Trade, General Trade
- **Latitude** : Kode Latitude
- **Longitude** : Kode Longitude

3. **Product**
- **ProductID** : Kode Unik Product
- **Product Name** : Nama Product
- **Price** : Harga dalam rupiah

4. **Transaction**
- **TransactionID** : Kode Unik Transaksi
- **Date** : Tanggal transaksi
- **Qty** : Jumlah item yang dibeli
- **Total Amount** : Price x Qty

Saya melakukan data ingestion ke dalam dbeaver

- Pilih connect to database di pojok kiri dan pilih postgreSQL
- Test connection ke postgreSQL
- Di menu drop down postgreSQL plih table dan klik kanan (import data)
- Pilih CSV file dan upload 4 file dari Kalbe



Hints

Dbeaver Connection with PostgreSQL

- Sebelum menggunakan dbeaver, pastikan sudah melakukan instalasi postgreSQL terlebih dahulu.
- Jangan lupa test connection terlebih dahulu ketika mau connect ke PostgreSQL
- Delimeter dari data adalah ; bukan ,
- Untuk melakukan query pada data klik icon SQL pada dbeaver

Saya melakukan exploratory data analysis di dbeaver

- query 1 : Berapa rata-rata umur customer jika dilihat dari marital statusnya ?
- query 2 : Berapa rata-rata umur customer jika dilihat dari gender nya ?
- query 3 : Tentukan nama store dengan total quantity terbanyak!
- query 4 : Tentukan nama produk terlaris dengan total amount terbanyak!

Berapa rata-rata umur customer jika dilihat dari marital statusnya ?

```
-- Rata-rata umur customer jika dilihat dari marital statusnya
select "Marital Status", round(avg(age)) age_avg
from "Case Study - Customer.csv" cscc
group by "Marital Status"
```

	Marital Status	age_avg
1		31
2	Married	43
3	Single	29

1. Rata-rata umur customer yang married (pernah menikah) yakni 43
2. Rata-rata umur customer yang single (belum menikah) yakni 29
3. Dan ada beberapa customer yang tidak menulis marital statusnya

Berapa rata-rata umur customer jika dilihat dari gender nya ?

```
-- Rata-rata umur customer jika dilihat dari gender nya
select  gender, round(avg(age)) age_avg
from "Case Study - Customer.csv" cscc
group by gender
```

	123 gender	123 age_avg
1	0	40
2	1	39

1. Rata-rata umur customer 0 (wanita) yakni 40
2. Rata-rata umur customer 1 (pria) yakni 39

Tentukan nama store dengan total quantity terbanyak!

```
-- Nama store dengan total quantity terbanyak
select sum(qty) total_qty, cssc.storename
from "Case Study - Transaction.csv" cstc
join "Case Study - Store.csv" cssc
on cstc.storeid = cssc.storeid
group by cssc.storename
order by 1 desc
limit 1
```

	123 total_qty	ABC storename
1	2,777	Lingga

Nama store dengan total quantity terbanyak yakni Lingga

Tentukan nama produk terlaris dengan total amount terbanyak!

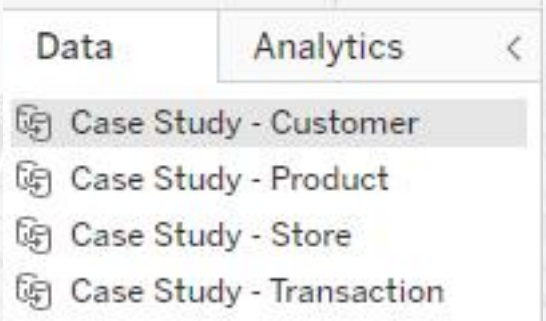
```sql
-- Nama produk terlaris dengan total amount terbanyak
select sum(totalamount) total_amount_sum, cspc."Product Name"
from "Case Study - Transaction.csv" cstc
join "Case Study - Product.csv" cspc
on cstc.productid = cspc.productid
group by 2
order by 1 desc
limit 1
```

	total_amount_sum	Product Name
1	27,615,000	Cheese Stick

Nama produk terlaris dengan total amount terbanyak yakni Cheese Stick

Saya melakukan data ingestion ke dalam tableau public

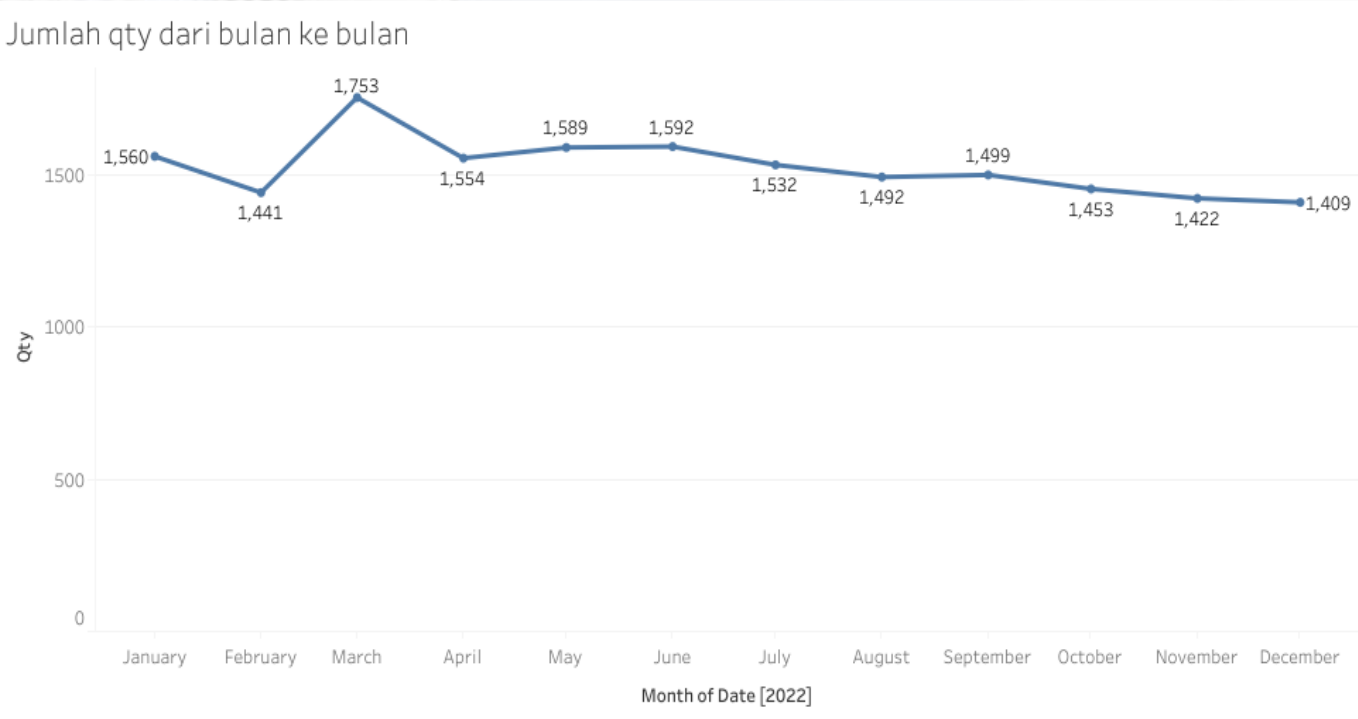
- Pertama buka tableau public. Sign in jika sudah punya akun dan sign up jika belum punya akun tableau public.
- Setelah itu klik create dan pilih web authoring
- Setelah itu akan di redirect ke halaman baru dan bisa langsung upload 4 data dari Kalbe



Saya membuat dashboard di tableau

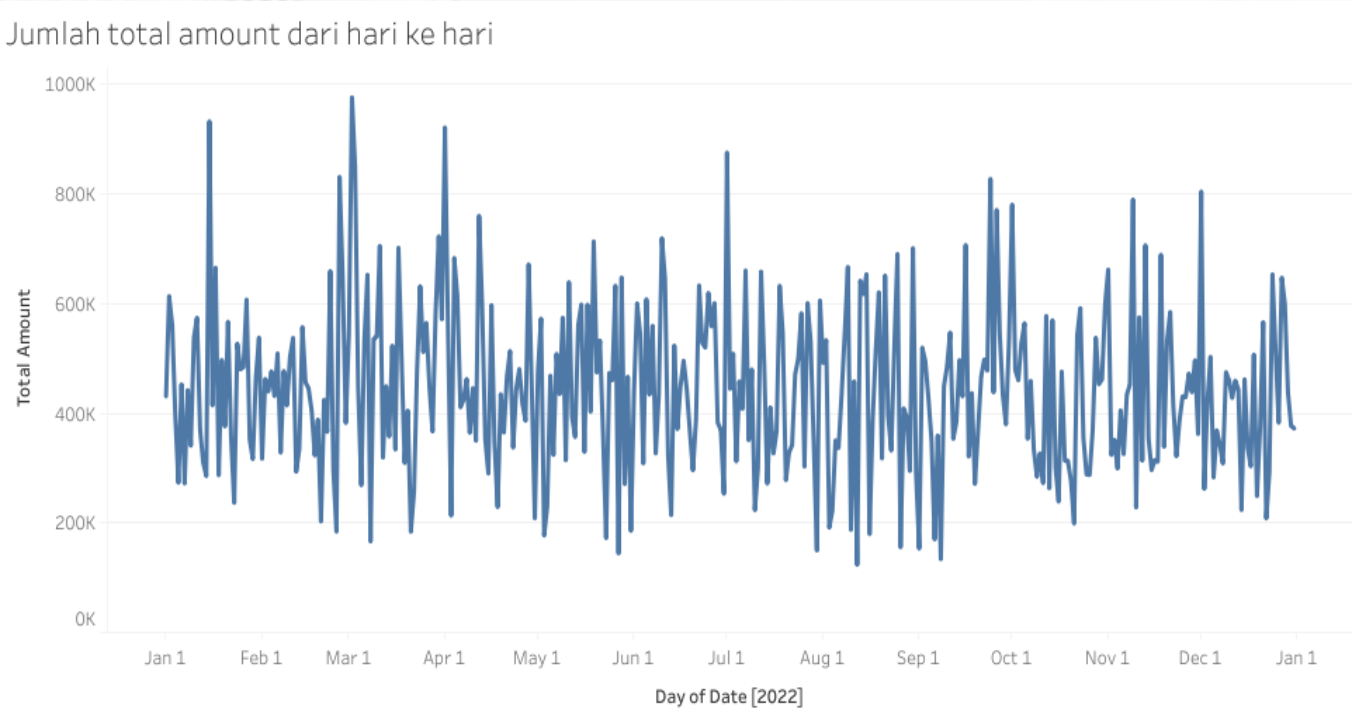
- Sebelum membuat dashboard terlebih dahulu membuat worksheet sebanyak 4.
  - Worksheet 1 Jumlah qty dari bulan ke bulan
  - Worksheet 2 Jumlah total amount dari hari ke hari
  - Worksheet 3 Jumlah penjualan (qty) by product
  - Worksheet 4 Jumlah penjualan (total amount) by store name
- Setelah itu bisa membuat dashboard dengan menggabungkan 4 worksheet.

Jumlah qty dari bulan ke bulan



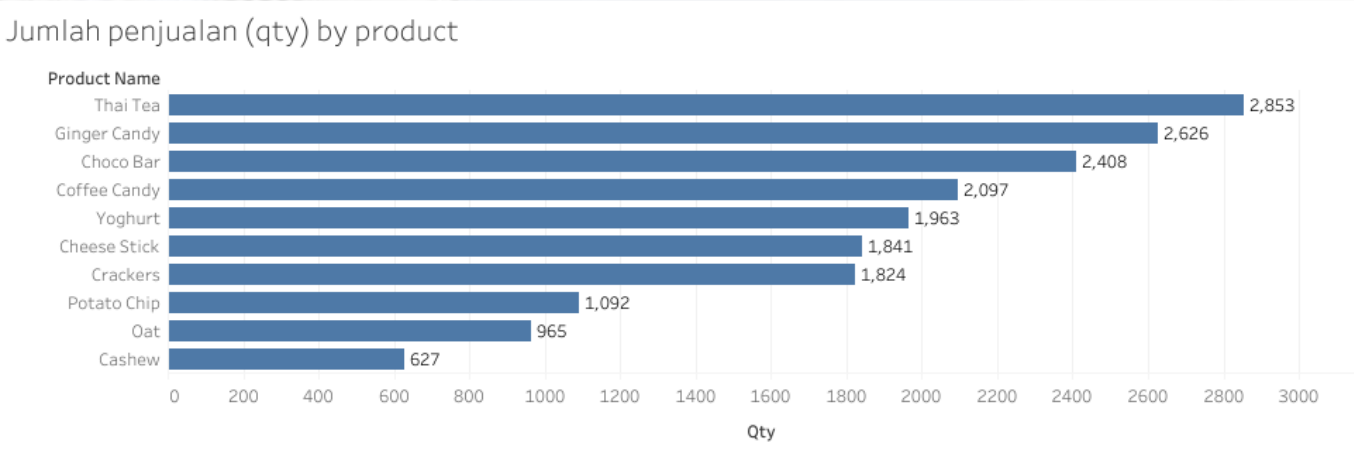
Jumlah qty tertinggi terjadi di bulan March sedangkan jumlah qty terrendah terjadi di bulan December.

Jumlah total amount dari hari ke hari



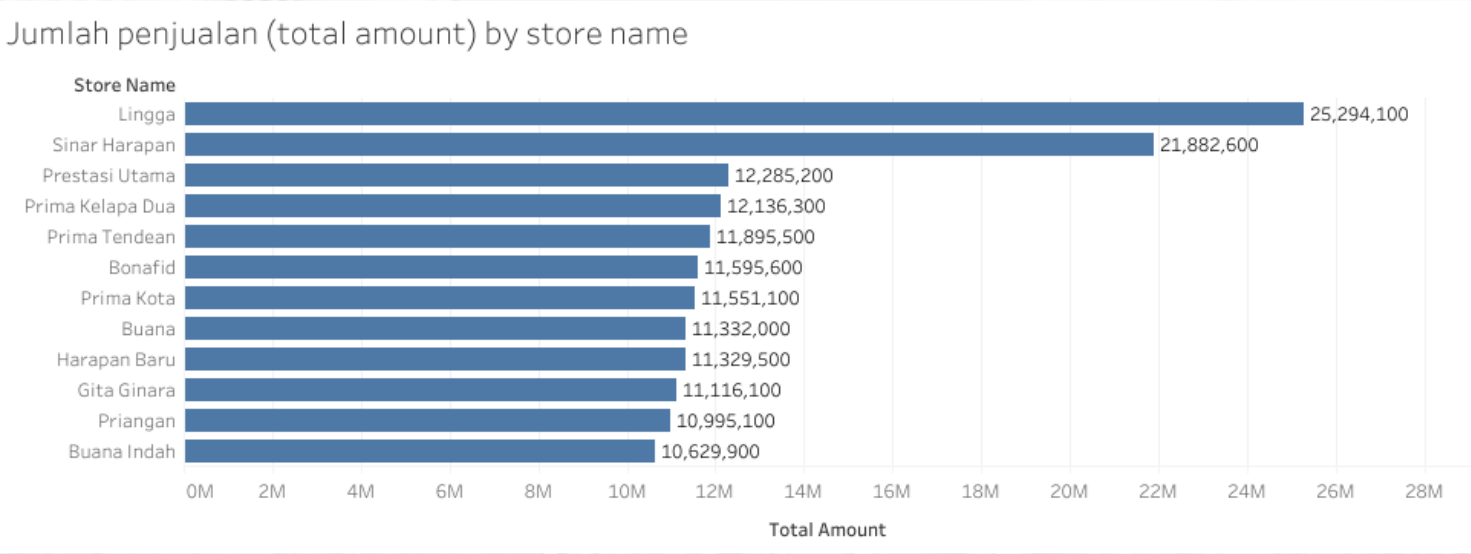
Jumlah total amount tertinggi yakni 976,500 sedangkan jumlah total amount terrendah yakni 123,600.

Jumlah penjualan (qty) by product



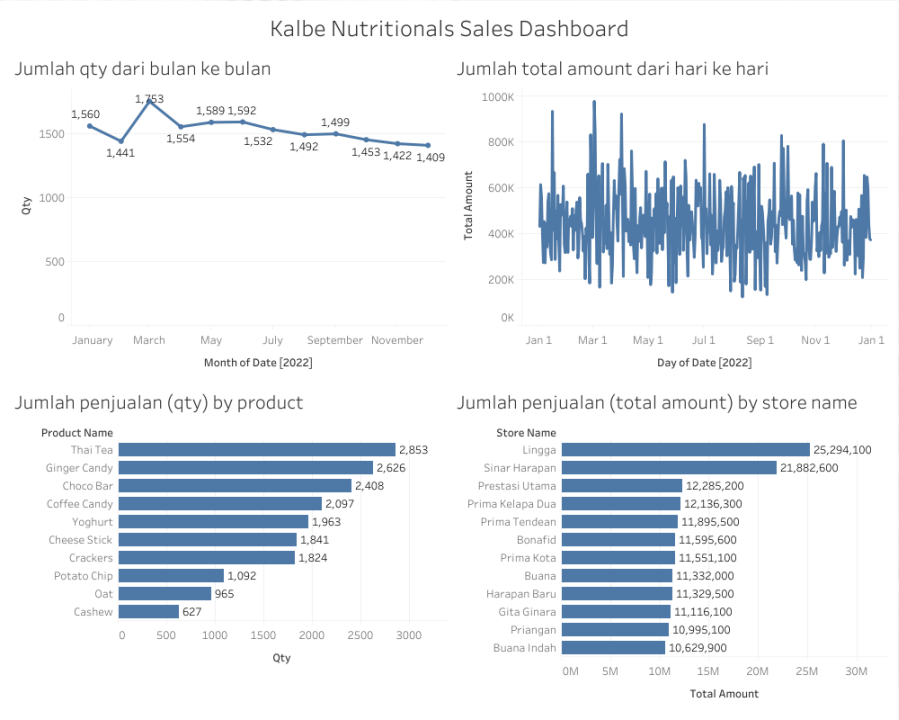
Jumlah penjualan product tertinggi yakni Thai Tea sedangkan jumlah penjualan product terrendah yakni Cashew.

Jumlah penjualan (total amount) by store name



Jumlah penjualan tertinggi dihasilkan oleh Lingga sedangkan jumlah penjualan terrendah dihasilkan oleh Buana Indah.

Dashboard



Source:
[Kalbe Nutritionals Sales Dashboard]

Saya membuat model prediktif menggunakan regresi dan membuat clustering

Machine Learning Regression (Time Series)

- Tujuan dari pembuatan model machine learning ini adalah untuk dapat memprediksi total quantity harian dari dari product yang terjual.
- Menggunakan metode time series **ARIMA**.

Machine Learning Clustering

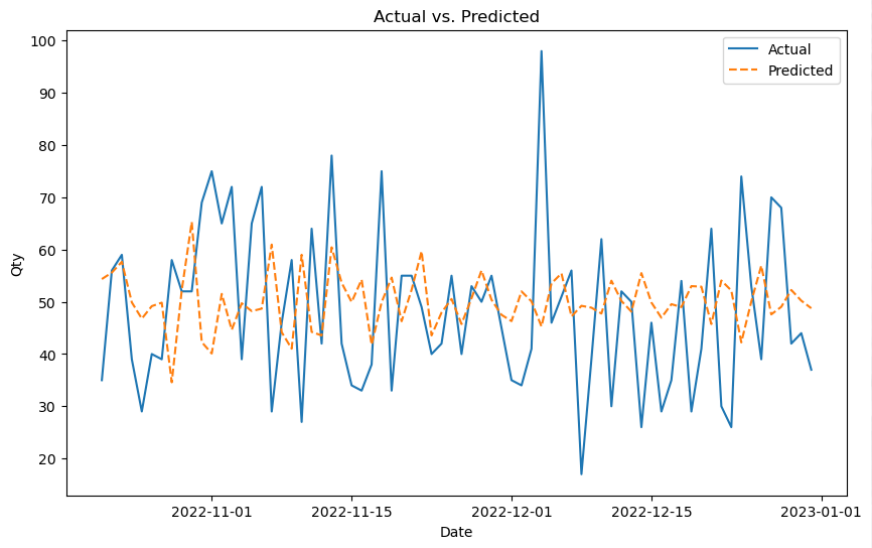
- Tujuan dari pembuatan model machine learning ini adalah untuk dapat membuat cluster customer-customer yang mirip.
- Menggunakan metode clustering **KMeans**.

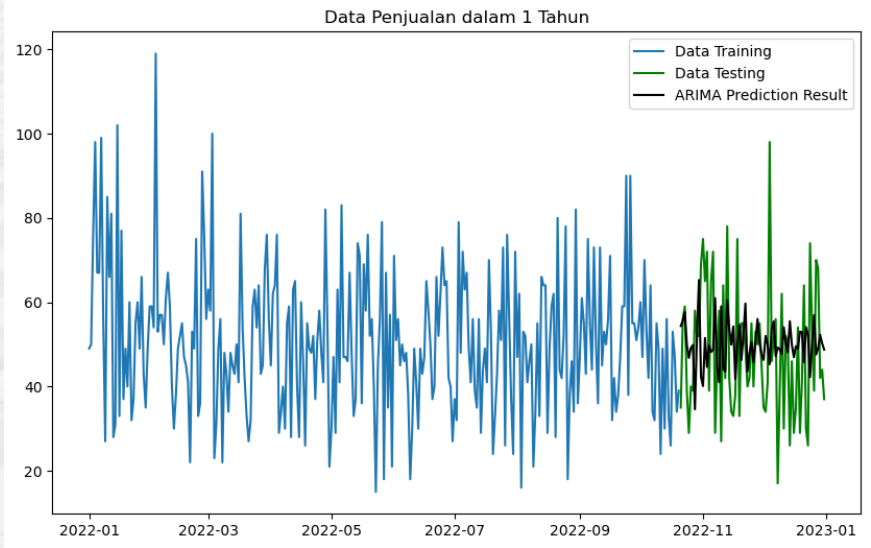
Machine Learning Regression (Time Series)

- Membaca data csv
- Data cleansing terlebih dahulu, merubah tipe data supaya sesuai
- Data merge untuk menggabungkan semua data
- Membuat data baru untuk regression, yaitu groupby by date lalu yang di aggregasi adalah qty di sum
- Menggunakan metode time series **ARIMA**

Forecasting Result

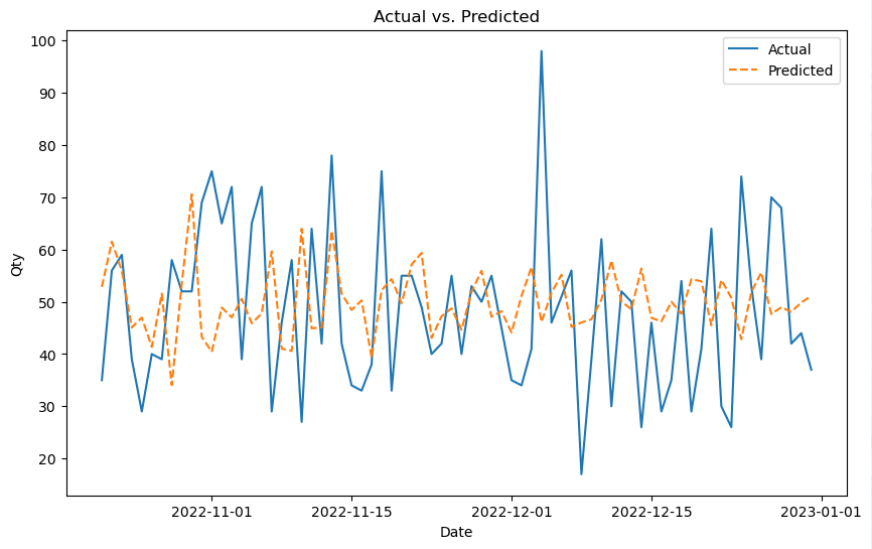
Model forecasting jumlah total produk terjual harian yang dihasilkan masih jauh dari yang diharapkan. Karena hasil forecasting tidak mendekati data asli dengan Mean Absolute Error (MAE) : 14.33.

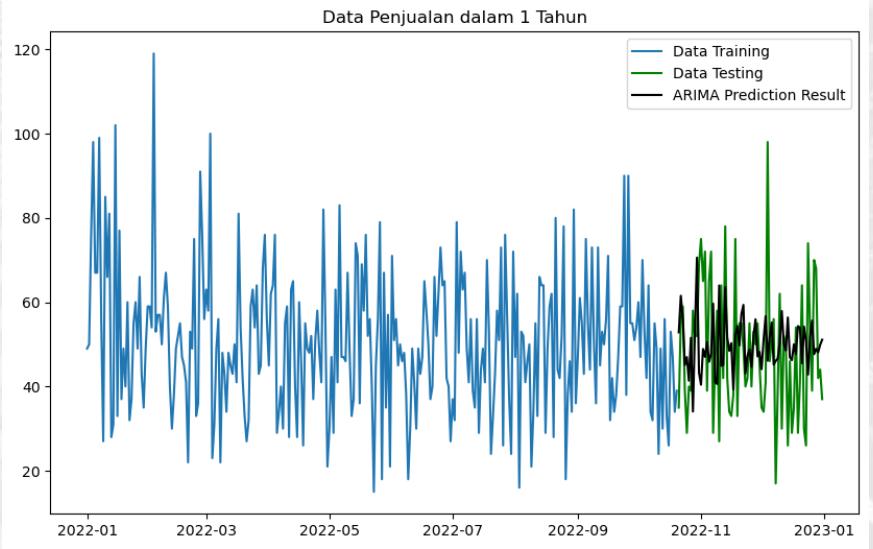




Forecasting Result (Hyperparameter Tuning)

Meskipun sudah menggunakan hyperparameter tuning, model yang dihasilkan masih belum mengalami perubahan yang signifikan. Namun terjadi penurunan nilai MAE menjadi 14.15.





Conclusion & Recommendation

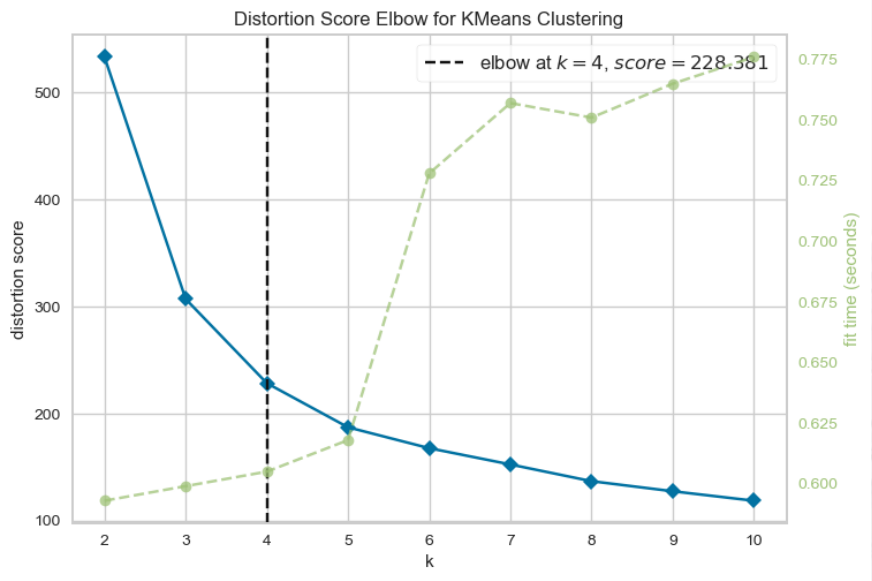
- Harus mengatur order value pada metode ARIMA dengan benar sehingga dapat menghasilkan forecasting values yang mendekati data aslinya.
- Identifikasi Pola Musiman: Jika terdapat pola musiman pada data, pastikan untuk memperhatikan pola tersebut dalam memilih model dan order values yang sesuai.
- Mencoba menggunakan metode time series lain yang mungkin lebih sesuai dengan data yang digunakan.

Machine Learning Clustering

- Membaca data csv
- Data cleansing terlebih dahulu, merubah tipe data supaya sesuai
- Data merge untuk menggabungkan semua data
- Membuat data baru untuk clustering, yaitu groupby by customerID lalu yang di aggregasi adalah :
  - Transaction id count
  - Qty sum
  - Total amount sum
- Menggunakan metode clustering **KMeans**

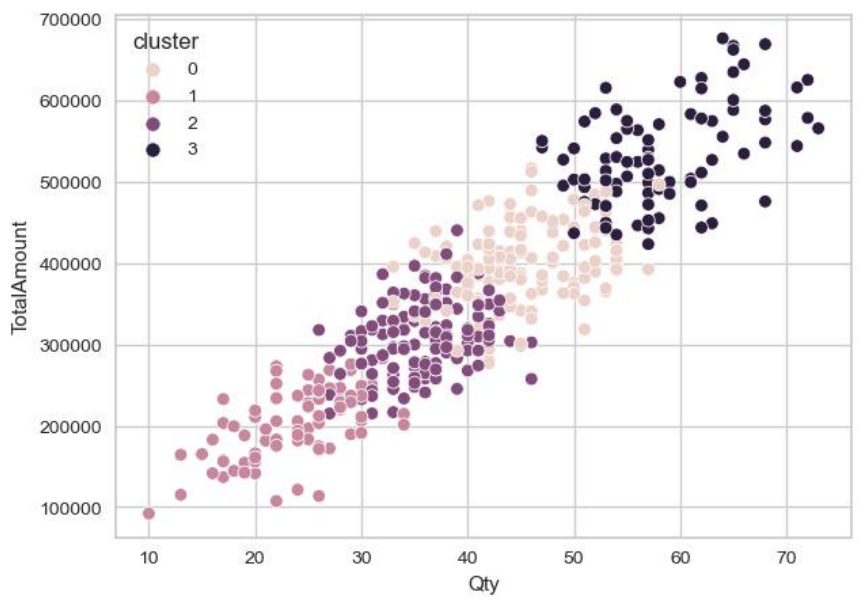
Clustering

Mencari k value terbaik menggunakan metode Elbow. Berdasarkan grafik di bawah ini, k value terbaik adalah k = 4 untuk pembentukan cluster pada Kmeans.



Clustering

Berikut hasil clustering dengan total terbentuknya 4 cluster yang diambil dari Qty dan TotalAmount.



Clustering Result

- **Cluster 0** = pelanggan dengan pembelian sedang hingga banyak dalam hal Quantity dan Total Amount. Ini adalah pelanggan yang menghabiskan banyak uang.
- **Cluster 1** = pelanggan dengan pembelian lebih sedikit dan menghabiskan uang relatif lebih sedikit. Ini adalah pelanggan dengan pembelian lebih sedikit dengan skala lebih terjangkau.
- **Cluster 2** = pelanggan dengan pembelian moderat dan membelanjakan uang dalam jumlah moderat. Ini adalah pelanggan dengan kebutuhan atau preferensi moderat dalam hal kuantitas dan pengeluaran.
- **Cluster 3** = pelanggan dengan pembelian banyak dan menghabiskan lebih banyak uang. Mewakili pelanggan dengan kebutuhan atau preferensi yang lebih tinggi dalam hal kuantitas dan pengeluaran.

Recommendation

Cluster 0 : Penggemar Kesehatan

- Fokus pada kampanye pemasaran yang menekankan manfaat produk nutrisi Kalbe Nutritionals dalam mendukung kesehatan dan gaya hidup aktif pelanggan di Cluster 0 yang memiliki pembelian dalam jumlah sedang hingga banyak. Mengkomunikasikan bahwa produk-produk Kalbe Nutritionals memberikan solusi terbaik untuk mendukung gaya hidup sehat dan aktif.
- Menawarkan paket produk nutrisi yang terdiri dari produk-produk dari berbagai tahapan kehidupan, seperti persiapan kehamilan, nutrisi untuk bayi, anak-anak, remaja dan orang dewasa. Dengan demikian, pelanggan dapat memilih produk yang sesuai dengan tahapan kehidupannya.
- Melakukan kerjasama dengan apotek atau toko kesehatan yang mempunyai pembelian dalam jumlah banyak untuk memperluas distribusi produk nutrisi Kalbe Nutritionals di Cluster 0.

Recommendation

Cluster 1 : Pembeli yang hemat

- Fokus pada kampanye promosi khusus produk nutrisi Kalbe Nutritionals yang memiliki harga terjangkau dan dapat memenuhi kebutuhan nutrisi pelanggan di Cluster 1 yang memiliki pembelian lebih sedikit. Komunikasikan bahwa produk-produk Kalbe Nutritionals memberikan nilai tinggi dengan harga terjangkau.
- Meningkatkan kesadaran akan manfaat produk nutrisi Kalbe Nutritionals melalui program edukasi tentang pentingnya nutrisi yang baik bagi kesehatan, terutama pada tahapan kehidupan tertentu. Dengan begitu, pelanggan bisa lebih memahami pentingnya mengonsumsi nutrisi yang berkualitas.
- Menawarkan diskon atau voucher belanja khusus sebagai insentif bagi pelanggan di Cluster 1 yang melakukan pembelian berulang atau membeli beberapa produk sekaligus.

Recommendation

Cluster 2 : Pencari Kualitas

- Meningkatkan ketersediaan produk nutrisi Kalbe Nutritionals terpopuler dan banyak dicari pelanggan di Cluster 2. Pastikan produk tersebut selalu tersedia dalam stok yang cukup.
- Meningkatkan komunikasi mengenai kualitas dan teknologi terkini yang digunakan dalam produksi produk nutrisi Kalbe Nutritionals, sehingga pelanggan di Cluster 2 lebih yakin terhadap kualitas produk yang dibelinya.
- Fokus pada pengembangan produk nutrisi Kalbe Nutritionals yang membantu memperkuat sistem kekebalan tubuh dan kesehatan secara umum, mengingat pelanggan di Cluster 2 mencari produk yang dapat diandalkan untuk kesehatannya.

Recommendation

Cluster 3 : Penggemar Nutrisi

- Fokus pada pengembangan produk nutrisi Kalbe Nutritionals yang lebih terspesialisasi dan kaya akan nutrisi, terutama bagi orang dewasa yang memiliki kebutuhan atau preferensi dari segi kuantitas dan pengeluaran. Komunikasikan bahwa produk Kalbe Nutritionals di Cluster 3 merupakan pilihan terbaik untuk memenuhi kebutuhan nutrisi tertentu.
- Membangun hubungan dengan dokter atau klinik besar yang mempunyai klien dengan kebutuhan nutrisi lebih tinggi, sehingga produk Kalbe Nutritionals dapat direkomendasikan kepada pasien.
- Mengadakan acara khusus yang melibatkan tenaga kesehatan untuk mengedukasi pelanggan di Cluster 3 mengenai manfaat dan keunggulan produk nutrisi Kalbe Nutritionals dalam memenuhi kebutuhan nutrisi lanjutan.

Glossary

- Clustering adalah algoritma machine learning unsupervised yang mengelompokkan titik data bersama berdasarkan kesamaan mereka. Tujuan dari clustering adalah untuk menemukan kelompok titik data yang mirip satu sama lain dan berbeda dari titik data dalam kelompok lain.
- Dashboard digunakan untuk mengorganisir dan visualisasi data perusahaan secara holistik.
- Data ingestion adalah proses pengumpulan, pemrosesan, dan pembebanan data dari berbagai sumber ke dalam sistem penyimpanan data yang sentral dan terstruktur. Tujuan dari data ingestion adalah memastikan data tersedia dan siap digunakan untuk analisis lebih lanjut.

Glossary

- Exploratory data analysis digunakan oleh data scientist untuk menganalisis dan menyelidiki kumpulan data dan meringkas karakteristik utamanya, seringkali menggunakan metode visualisasi data.
- Machine Learning adalah sub bidang Artificial Intelligence, yang secara luas didefinisikan sebagai kemampuan mesin untuk meniru perilaku manusia yang cerdas.
- Regresi merupakan algoritma machine learning yang tergolong dalam supervised learning, Regresi adalah metode statistik yang digunakan untuk menganalisis hubungan antara variabel dependen dan satu atau lebih variabel independen. Variabel dependen adalah variabel yang diprediksi, dan variabel independen adalah variabel yang digunakan untuk memprediksi variabel dependen.

ARIMAX atau Regresi ARIMA

Tentang ARIMAX

ARIMAX atau Regresi ARIMA merupakan perpanjangan dari model ARIMA. Dalam peramalan, metode ini juga melibatkan variabel independen [4]. Model ARIMAX merepresentasikan komposisi rangkaian waktu keluaran menjadi komponen-komponen berikut: *autoregressive* (AR), *moving average* (MA), terintegrasi (I), dan facktor eksternal (X) [5]. Faktor eksternal (X) mencerminkan penggabungan tambahan dari nilai sekarang $u_i(t)$ dan nilai masa lalu $u_i(t-j)$ dari input faktor eksternal (variabel independen) ke dalam model ARIMAX[6].

Rumus Multiple linear regression models:

$$Y = \beta_0 + \beta_1*x_1+...+\beta_i*x_i+\varepsilon$$

Di mana Y merupakan sebuah variabel dependen dari variabel prediktor x_i dan ε biasanya diasumsikan sebagai error/white noise. Kami akan mengganti ε dengan n_t pada persamaan. Error ϕ_t diasumsikan mengikuti hasil dari model ARIMA. Sebagai contoh, jika n_t mengikuti model ARIMA (1,1,1), dapat kita tuliskan

$$Y = \beta_0 + \beta_1x_1+\beta_2x_2+...+\beta_ix_i+\eta_t$$

$$(1-\phi_1B)(1-B)\eta_t = (1+\phi_1B)\varepsilon_t$$

Di mana ε_t, merupakan seri white noise. Model ARIMAX memiliki two error terms; the error dari model regresi yang dinotasikan dengan ϕ_t dan error dari model ARIMA model yang dinotasikan dengan ε_t.

Source : https://algoritmaonline.com/2021-09-09-regression-arima-arimax/

K-Means Clustering

- K-Means Clustering adalah suatu metode penganalisaan data atau metode Data Mining yang melakukan proses pemodelan unsupervised learning dan menggunakan metode yang mengelompokkan data berbagai partisi.
- Metode:
  - Pengguna menentukan jumlah kluster k.
  - Secara acak memilih k titik data sebagai pusat awal dari klaster.
  - Mengelompokkan setiap titik data ke kluster dengan pusat terdekat.
  - Menghitung kembali nilai rata-rata centroid berdasarkan semua titik data yang telah dikelompokkan.
  - Mengulangi langkah 3 dan 4 samapi pusat tidak berubah lagi.
- Menentukan banyaknya K (klaster / kelompok)
  - Mendefinisikan rentang nilai K untuk menjalankan *K-Means Clustering*.
  - Mengevaluasi *Sum of Squared Errors* (SSE) untuk model menggunakan setiap jumlah klaster yang telah ditentukan.

Appendix

- Result : https://drive.google.com/drive/folders/1xiIzfIDjSQp8DwHWkXNhwVHTfjd2pJVA?usp=sharing
- Github : https://github.com/FebeJovita/PBI_Data-Scientist_Kalbe-_Nutritionals
- My portfolios : https://github.com/FebeJovita

Thank You

Let's collaborate.

LinkedIn : www.linkedin.com/in/febe-jovita-7b1572246
Email : febejovita@gmail.com
Instagram : https://instagram.com/fjovitaj
Github (My portfolios) : https://github.com/FebeJovita

Rakamin Academy X KALBE Nutritionals