TUGAS ESAI USE CASE DATA SCIENCE INDUSTRI RETAIL :

IDENTIFIKASI MENGENAI BISNIS RETAIL ONLINE K-MEANS DAN PENGELOMPOKAN HIRARKINYA

Oleh : Rumaisya Az-zahra

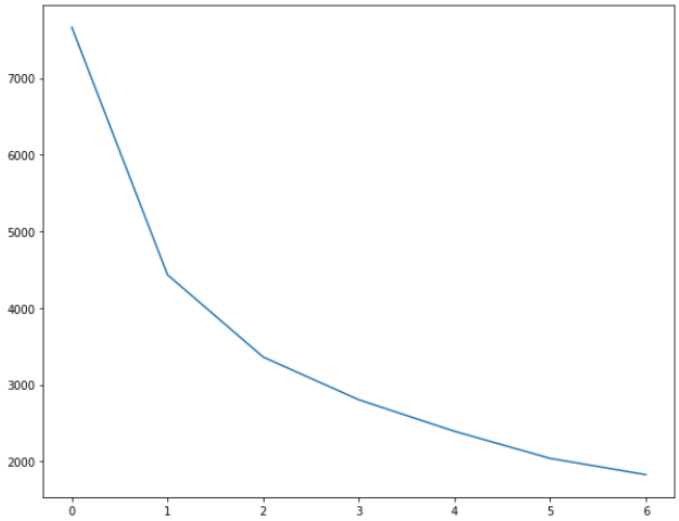
Perkembangan teknologi dan internet yang semakin pesat menjadikan online shop semakin digemari oleh masyarakat di masa revolusi industri 4.0 saat ini. Asosiasi Penyelenggara Jasa Internet (APJII) menyampaikan data statistik sebesar 143 juta atau lebih dari 50 persen orang terhubung dengan internet (Hardilawati, Binangkit, and Perdana 2019). Dari banyaknya orang yang mengakses internet tidak heran jika online shop semakin berkembang dan bersaing ketat satu sama lain untuk menawarkan produk secara menarik agar peminatnya tertarik untuk membeli produknya. Kemudian, dari persaingan antar online shop data analyst dapat membantu untuk meningkatkan keuntungan dengan melakukan pengelompokan pelanggan untuk mendapat informasi agar strategi penjualan dapat ditentukan lebih baik.

Pada bisnis online retail, terdapat jumlah pelanggan yang banyak sehingga untuk mempermudah perusahan perlu mensegmentasikan pelanggannya berdasarkan proses penganalisisan informasi data. Pada *use case business* ini, dilakukan pengelompokan pelanggan online retail dengan tujuan mensegmentasi pelanggan berdasarkan RFM (*Recency, Frequency, Monetary*) sehingga perusahaan dapat menargetkan pelanggannya secara efisien. Dari efisiensi pengklusteran pelanggan, perusahaan dapat mengidentifikasi mana kelompok pelanggan yang sering membeli produk sehingga dapat berpotensi membeli barang mewah. Kemudian, perusahaan juga dapat mengidentifikasi pelanggan yang dapat diberi sebuah promosi barang dengan harga yang cukup tinggi. Dengan demikian, perusahaan dapat memaksimalkan keuntungan dari hasil segmentasi pelanggan.

Proses segmentasi pelanggan bisnis online dapat dilakukan dengan mengumpulkan, memahami, dan mengevaluasi data yang diperoleh. Data yang akan dikelompokan dalam *use case business* ini diperoleh dari sumber Kaggle. Dalam pensegmentasian pelanggan data-data yang digunakan ialah nomor faktur, kode stock, deskripsi, jumlah, tanggal faktur, harga barang, customer ID, dan negara asal. Berdasarkan data yang didapat dilakukan proses pemahaman data sebelum mengecek missing value dari kumpulan data yang diperoleh. Pemahaman data dilakukan untuk mengkonfirmasi data terdistribusi dengan baik tanpa adanya penyimpangan yang perlu ditangani lebih lanjut.

Targetan pelanggan yang efektif dapat meningkatkan bisnis online perusahaan dengan mengelompokan pelanggannya sesuai ketentuan RFM. Dari hal tersebut, dapat dilakukan proses data preparation dengan dilakukan *check missing value* yaitu membuang variabel yang memiliki jumlah observasi missing value yang sangat besar. Selanjutnya, menyesuaikan tipe data sesuai dengan variabel serta melakukan exploratory data secara visualisasi untuk melihat persebaran data untuk memperoleh proporsi dari target variabel. Pada proses ini dibutuhkan variabel *derived* dalam proses algoritma dengan dibuatnya 3 variabel pengelompokan yaitu *Recency, Frequency,* dan *Monetary*. *Recency* dikelompokan berdasarkan seberapa baru pelanggan membeli produk yang mana berarti jumlah hari sejak pelanggan melakukan pembelian terakhir. *Frequency* diklusterisasi berdasarkan seberapa sering pelanggan membeli dalam jangka waktu tertentu. Dari hal tersebut, dapat dipahami sebagai seberapa sering atau berapa banyak pelanggan menggunakan produk suatu perusahaan. *Monetary* dikelompokan dari jumlah total uang yang dihabiskan pelanggan dalam periode tertentu. Oleh karena itu, pembelanja dengan pembelian terbesar akan dibedakan dengan pelanggan lain seperti VIP.

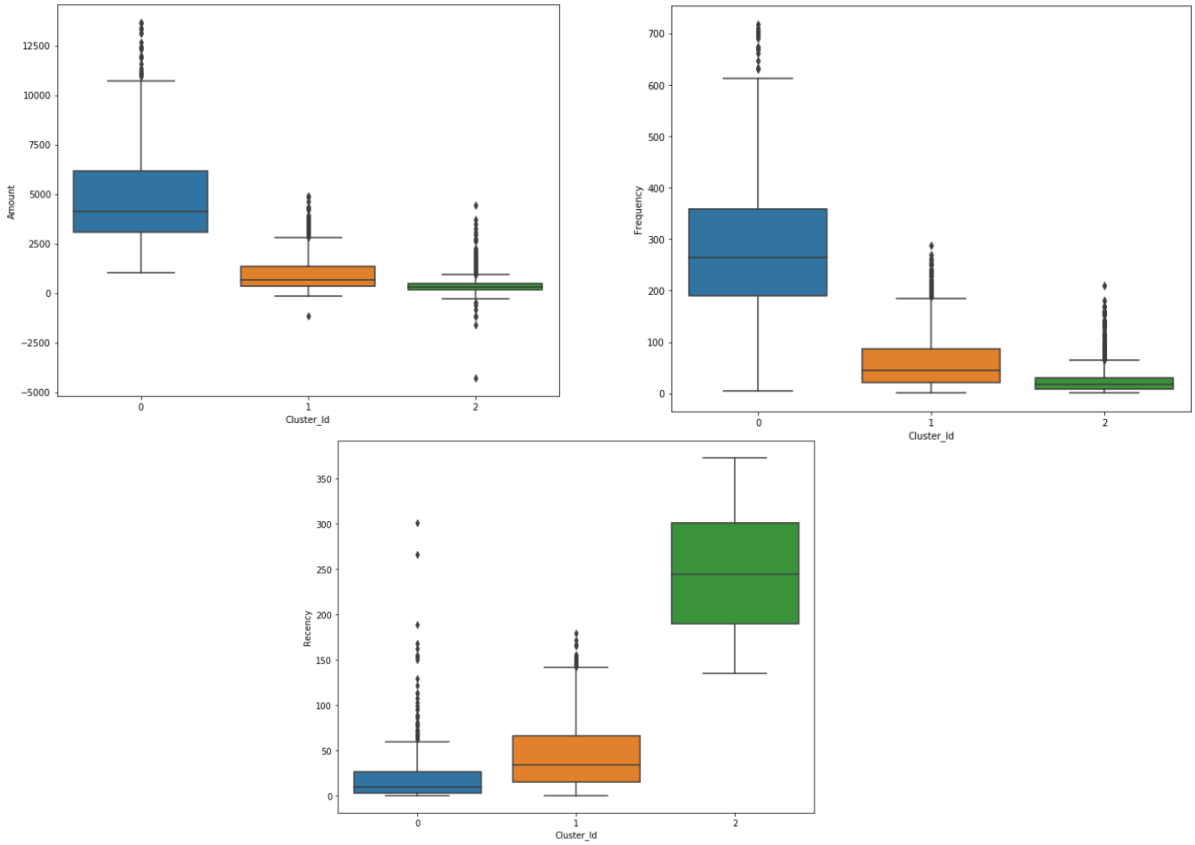
Pada tahap modelling dilakukan dengan cara K-Means clustering. K-Means clustering didefinisikan sebagai penentuan kelompok-kelompok dalam set dapa dengan mengelompokan data menurut kesamaan fitur. Peluang dalam satu kelompok akan semakin besar jika fitur yang dimiliki data semakin mirip. Pengelompokan K-means juga termasuk ke dalam salah satu algoritma yang paling sederhana dan populer. Cara kerja di tahap pemodelan K-Means dalam retail online yaitu pertama menginisialisasi k poin, yang disebut mean (rata-rata), secara acak. Kemudian mengkategorikan setiap item ke rata-rata terdekatnya dan memperbarui koordinat rata-rata, yang merupakan rata-rata item dan dikategorikan dalam mean tersebut. Selanjutnya, mengulangi proses untuk jumlah iterasi (urutan) tertentu. Di proses modelling ini dilakukan proses penemuan nomer kluster yang optimal dengan cara *Elbow Curve.* Proses ini dilakukan dengan langkah mendasar untuk setiap algoritme tanpa pengawasan dalam menentukan jumlah klaster optimal tempat data dapat dikelompokkan.



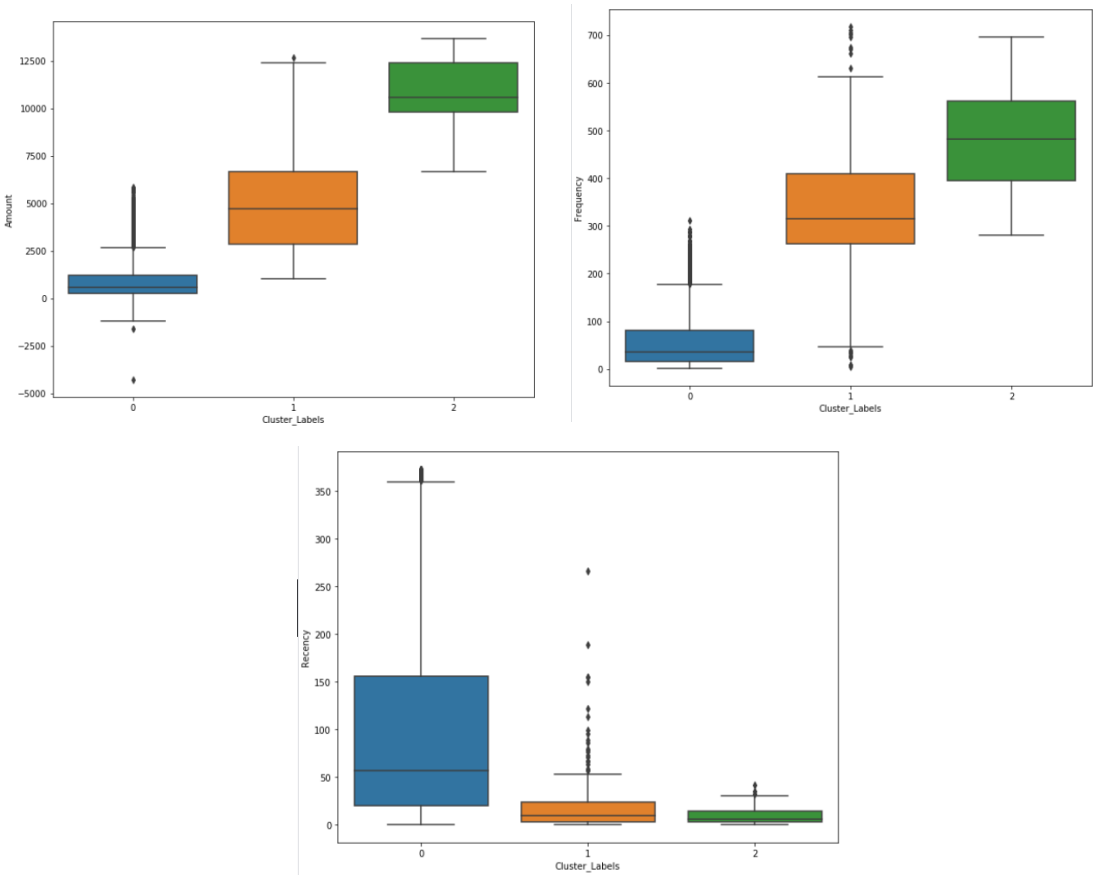
Gambar 1. Elbow Curve

Pada proses modelling, dilakukan juga pengujian *silhouette coefficient* untuk mengetahui seberapa dekat relasi antara objek di sebuah kluster serta seberapa jauh kluster tersebut terpisah dari kluster lainnya. *Silhouette coefficient* pada modelling ini dilakukan dengan perumusan berikut $\frac{P-Q}{\max(P,Q)}$. P adalah jarak rata-rata ke titik-titik di cluster terdekat yang bukan merupakan bagian dari titik data. Sedangkan, Q adalah jarak intra-cluster rata-rata ke semua titik dalam clusternya sendiri. Nilai rentang skor siluet terletak antara -1 hingga 1. Skor yang mendekati 1 menunjukkan bahwa titik data tersebut sangat mirip dengan titik data lain dalam klaster tersebut, Skor yang mendekati -1 menunjukkan bahwa titik data tidak mirip dengan titik data dalam klasternya.

Berdasarkan hasil dari K-Means yang digunakan dari pengklusteran data diperoleh plot visualisasi cluster ID dan RFM. Kemudian, pada proses pengelompokan hierarki melibatkan pembuatan kelompok yang memiliki urutan yang telah ditentukan sebelumnya dari atas ke bawah. Ada dua jenis pengelompokan hierarkis yaitu *divisive* dan *agglomerative.* Dari hasil data statistik yang diperoleh dibuat juga visualisasi plot antara cluster labels dan RFM. Berikut merupakan data visualisasi atau plot antara cluster ID dan RFM serta cluster labels dan RFM.



Gambar 2 Plot Cluster ID dan RFM (*Recency, Frequency, dan Monetary*)



Gambar 3 Plot Cluster labels dan RFM (*Recency, Frequency, dan Monetary*)

Berdasarkan hasil dari metode K-Means diperoleh penjelasan akhir pengelompokan. K-Means clustering dengan 3 cluster ID yaitu nasabah dengan cluster ID 1 merupakan nasabah dengan jumlah transaksi yang tinggi dibandingkan dengan nasabah lainnya. Pelanggan dengan cluster ID 1 sering menjadi pembeli. Pelanggan dengan cluster ID 2 bukanlah pembeli baru dan karenanya paling tidak penting dari sudut pandang bisnis. Kemudian, clustering Hirarkis dengan 3 Label cluster yaitu Pelanggan dengan Cluster_Labels 2 merupakan pelanggan dengan jumlah transaksi yang tinggi dibandingkan dengan pelanggan lainnya. Pelanggan dengan Cluster_Labels 2 sering menjadi pembeli. Pelanggan dengan Cluster_Labels 0 bukanlah pembeli baru dan karenanya paling tidak penting dari sudut pandang bisnis. Dengan demikian, skor pelanggan didapat dari pengklusteran cluster ID dan cluster labels berdasarkan model dari RFM (*Recency, Frequency, dan Monetary*)

REFERENSI :

Hardilawati, Wan Laura, Intan Diane Binangkit, and Riky Perdana. 2019. "Endorsement: Media Pemasaran Masa Kini." *Jimupb* 7(1):88–98.

Binus. (2020). Cross-Industry Standard Process for Data Mining (CRISP-DM). https://mmsi.binus.ac.id/2020/09/18/cross-industry-standard-process-for-data-mining-crisp-dm/

Nurhayati, N. (2018). Pengujian Silhouette Coefficient. http://nopi-en.blogspot.com/2018/11/pengujian-silhouette-coefficient.html

Kumar, M. (2020). Online Retail K-Means & Hierarchical Clustering. *Kaggle*. https://www.kaggle.com/code/hellbuoy/online-retail-k-means-hierarchical-clustering#Step-4-:-Building-the-Model

LINK GITHUB & LINKEDIN

GITHUB PROFILE : https://github.com/rumaisyaa

GITHUB PAGE : https://rumaisyaa.github.io/

LINKEDIN : www.linkedin.com/in/rumaisya-az-zahra