APORAN AKHIR PROYEK PENAMBANGAN DATA

Customer Clustering for CRM in XYZ Store

DISUSUN OLEH:

12S17001	**Krista Lumbantoruan**
12S17042	**Aulia SL Pakpahan**
12S17049	**Paddy T Silitonga**

PROGRAM STUDI SARJANA SISTEM INFORMASI

FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO

INSTITUT TEKNOLOGI DEL

JANUARI 2020

DAFTAR ISI

Daftar Gambar

Daftar Tabel

BAB 1. BUSINESS UNDERSTANDING

Pada bab ini akan menjelaskan tentang pemahaman bisnis terhadap proyek yang akan dilakukan.

1.1 Business Objectives

Perkembangan teknologi yang semakin pesat dimanfaatkan oleh para pelaku bisnis dengan memanfaatkan teknologi digital dan internet. Salah satu contoh pelaku bisnis yaitu Toko XYZ yang merupakan toko bahan makanan yang memiliki banyak pelanggan dengan karakteristik yang berbeda-beda. Toko XYZ mengalami masalah dalam meningkatkan profit bisnis dan menentukan parameter yang tepat dalam membuat strategi pemasaran. Dalam hal ini, Toko XYZ ingin membuat keputusan dalam memperoleh strategi yang dapat meningkatkan kualitas pemasaran. Pada proyek ini, penulis akan mencoba menganalisis perilaku dari pelanggan dengan memberikan solusi untuk permasalahan Toko XYZ.

Kualitas pemasaran yang baik dapat memicu profit yang lebih besar. Salah satu cara untuk memperoleh keuntungan dalam bidang pemasaran adalah memahami pelanggan. *Customer Relationship Management* (CRM) merupakan sebuah pendekatan strategis yang berkonsentrasi untuk meningkatkan nilai terhadap perusahaan melalui penanganan hubungan yang tepat dengan pelanggan utama dan juga dengan kelompok pelanggan lainnya. Karena itu, pengelompokan pelanggan atau *customer clustering* merupakan langkah yang baik dalam pemasaran untuk menggambarkan perilaku pelanggan terhadap perusahaan.

Pada proyek ini, *customer clustering* dilakukan dengan analisis RFM (*Recency, Frequency, and Monetary*), yaitu melakukan pengelompokan berdasarkan kapan terakhir pelanggan melakukan transaksi, sering tidaknya melakukan transaksi, dan transaksi yang dilakukan pada jumlah sedikit atau besar. Dengan demikian, tujuan proyek ini adalah untuk memahami karakteristik pelanggan dengan melakukan pengelompokan berdasarkan data transaksi belanja pelanggan yang dianalisis menggunakan model RFM. Pengelompokan tersebut dilakukan dengan melihat reaksi pasar pelanggan atau transaksi pemasaran yang dilakukan agar strategi pemasaran dan layanan yang diberikan dengan tepat.

1.2 Situation Assesment

Toko XYZ merupakan toko yang menjual berbagai jenis bahan makanan. Toko ini ingin meningkat kualitas pemasaran dengan membuat strategi pemasaran yang tepat untuk

memicu profit yang lebih besar. Rekaman data pemasaran pelanggan Toko XYZ dapat digunakan untuk meninjau perilaku pelanggan. Rekaman data tersebut antara lain waktu transaksi, id pelanggan, id produk, dan lain-lain. Atribut terkait transaksi akan menjadi fokus dalam pengelompokan pelanggan di Toko XYZ berdasarkan analisis model RFM. Batasan dari proyek ini adalah rekaman data yang diguanakan pada Toko XYZ adalah pada 1 November 2000 sampai dengan 28 Februari 2001. Penentuan keputusan dalam pembuatan strategi adalah berfokus pada data transaksi pelanggan dengan menggunakan analisis RFM.

1.3 Data Mining Goal

Tujuan dari penambangan data ini adalah agar didapatkan suatu pengetahuan mengenai perilaku pelanggan sehingga dapat menjadi salah satu parameter untuk membuat keputusan atau stategi yang tepat dalam meningkatkan kualitas pemasaran yang memicu peningkatan profit.

1.4 Produce Project Plan

Proyek ini akan menggunakan algoritma K-Means pada *python*. Data pelanggan akan diekstraksi berdasarkan tiga variabel RFM setiap pelanggan pada Toko XYZ. Setiap pelanggan diakumulasikan masing-masing nilai RFM, yaitu *Recency, Frequency,* dan *Monetary. Recency* yaitu rentan terakhir kali transaksi dilakukan. Semakin kecil rentangnya, maka nilai R akan semakin besar. *Frequency* yaitu jumlah transaksi dalam satu periode. Semakin banyak *frequency*, maka nila F akan semakin besar. *Monetary*, yaitu nilai pelanggan berupa uang yang dikeluarkan pelanggan pada periode tersebut maka nilai M semakin besar. Setelah dilakukan ekstraksi data pada tahap *data preparation*, maka dilakukan penentuan jumlah *cluster* menggunakan *Elbow method.* Pemodelan data akan digunakan untuk mengoptimalkan hasil pengelompokan. *Modeling* merupakan tahap untuk melakukan pemilihan dan penerapan berbagai teknik pemodelan dan beberapa parameternya akan disesuaikan untuk mendapatkan nilai yang optimal.

Menurut Tsiptsis dan Chorianopoulus [1], analisis RFM digunakan untuk memahami karakteristik *customer.* Ada 5 label pelanggan yang akan digunakan pada penelitian ini yaitu, *superstars customers, golden customer, occasional customers, everyday shoppers,* dan *dormant customers.* Label ini dipisahkan berdasarkan transaksi yang dilakukan pelanggan. Pelanggan yang sering melakukan transaksi dan dengan nilai yang besar akan masuk ke

dalam label *superstars customers. Golden customer* yaitu pelanggan sering melakukan transaksi dengan nilai transaksi kedua tertinggi. *Occasional customer* yaitu label untuk pelanggan yang tidak terlalu sering melakukan transaksi namun memiliki nilai transaksi yang tinggi. *Everyday shopper* untuk pelanggan yang sering bertransaksi dengan nilai transaksi yang rendah. Dan untuk pelanggan yang memiliki nilai transaksi rendah dan sudah lama tidak bertransaksi masuk kedalam label *dormant customers.*

Hasil dari proses K-Means akan berupa informasi yang menunjukkan jumlah pelanggan setiap *cluster*, titik pusat atau *centroid*, dan nilai performa *cluster*. Hasil *cluster* akan dilakukan pengujian dengan menggunakan SSE atau *Sum of Square Error*. Selanjutnya, hasil *clustering* akan dianalisis dengan memberikan keputusan apakah teknik pemodelan yang dipergunakan dapat dijadikan standar dalam menentukan tujuan proyek. Setelah itu, akan dilakukan *deployment plan* dengan menjelaskan gambaran mengenai rencana terhadap pembuatan laporan, lalu melakukan *produce final report* memberikan visualisasi dari laporan yang telah dibuat berdasarkan pada *deployment plan*.

BAB 2. DATA UNDERSTANDING

Pada bab ini akan menjelaskan tentang pemahaman terkait data yang akan digunakan untuk menyelesaikan masalah pada proyek.

2.1 Collect Initial Data

Pengumpulan data dilakukan dengan eksplorasi di internet dan melakukan peninjauan kecocokan dengan penelitian yang dilakukan. Pada akhirnya, dataset yang akan digunakan diperoleh dari https://www.kaggle.com/chiranjivdas09/ta-feng-grocery-dataset. Data ini sebelumnya sudah pernah digunakan dalam penelitian *clustering customer segementation* berdasarkan LRFM dengan *tools* R. Bedanya pada penelitian yang akan penulis lakukan adalah penulis akan melakukan pengelompokan terhadap *customer* berdasarkan RFM menggunakan python. *Code program* untuk penelitian ini juga belum ada ditemukan di github dengan menggunakan *dataset* yang sama dan *tools* yang sama, sehingga dapat dipastikan penelitian ini tidak melakukan *plagiarism.*

2.2. *Describe Data*

Pada proyek ini, *dataset* yang digunakan adalah Ta-Feng Grocery yang merupakan kumpulan data belanja bahan makanan pada sebuah supermarket. *Dataset* tersebut merupakan hasil rekaman data pemasaran pelanggan dari 1 November 2000 sampai dengan 28 Februari 2001.

2.3. *Explore Data*

Dataset Ta-Feng Grocery terdiri dari 817.741 transaksi milik 32.266 pengguna dengan 23.812 item. *Dataset* terdiri dari 9 atribut, yaitu *tansaction_dt, customer_id, age_group, pin_code, product_subclass, product_id, amount, asset,* dan *sales_price.*

Tabel 1. Atribut dan Tipe *Dataset*

Atribut	**Tipe data pada *Ms. Excel***
TRANSACTION_ DT	*Number*

CUSTOMER_ID	*Number*
AGE_GROUP	*Number*
PIN_CODE	*Number*
PRODUCT_SUBCLASS	*Number*
PRODUCT_ID	*Number*
AMOUNT	*Number*
ASSET	*Number*
SALES_PRICE	*Number*

2.4. *Verify Data Quality*

Kualitas data pada Ta-Feng Grocery belum baik, sehingga masih dibutuhkan *data preprocessing* untuk menghasilkan kualitas data yang baik. Pada *dataset*, terdapat data yang kotor dan bernilai kosong. Untuk mengatasi masalah ini, maka tahap selanjutnya akan dilakukan *preprocessing data* untuk mendukung proses penambangan data dan menghasilkan pengelompokan yang tepat.

BAB 3. DATA PREPARATION

Pada bab ini akan menjelaskan tentang proses pemilihan dan pengolahan data yang akan dibutuhkan pada tahap pemodelan.

3.1 *Data Set*

Dataset Ta-Feng Grocery adalah kumpulan data belanja bahan makanan yang dirilis oleh ACM RecSys. *Dataset* tersebut merupakan *record* data transaksi selama 4 bulan, yaitu dari November 2000 hingga Februari 2001. Total transaksi pada *dataset* adalah 817.741 transaksi yang dimiliki oleh 32.266 pelanggan dengan 23.812 produk. Tampilan *dataset* dapat dilihat pada Gambar 1.

	TRANSACTION_DT	CUSTOMER_ID	AGE_GROUP	PIN_CODE	PRODUCT_SUBCLASS	PRODUCT_ID	AMOUNT	ASSET	SALES_PRICE
0	11/1/2000	1104905	45-49	115	110411	4710199010372	2	24	30
1	11/1/2000	418683	45-49	115	120107	4710857472535	1	48	46
2	11/1/2000	1057331	35-39	115	100407	4710043654103	2	142	166
3	11/1/2000	1849332	45-49	Others	120108	4710126092129	1	32	38
4	11/1/2000	1981995	50-54	115	100205	4710176021445	1	14	18

Gambar 1. Tampilan *dataset* Ta-Feng Grocery

Berdasarkan eksplorasi data yang telah dilakukan, *dataset* terdiri dari 10 atribut, yaitu:

1. *TANSACTION_DT*, yaitu tanggal transaksi yang dilakukan pelanggan pada Toko XYZ.
2. *CUSTOMER_ID*, yaitu id setiap pelanggan yang melakukan transaksi.
3. *AGE_GROUP*, yaitu grup yang mengelompokkan usia pelanggan. Berikut 10 kelompok usia pelanggan yang terdapat pada *dataset*.
   a. < 25
   b. 25-29
   c. 30-34
   d. 35-39
   e. 40-44
   f. 45-49
   g. 50-54

h. 55-59
i. 60-65
j. > 65

4. *PIN_CODE*, yaitu kode pos setiap pelanggan yang melakukan transaksi. Berikut 8 kode pos yang terdapat pada *dataset.*
   a. 105
   b. 106
   c. 110
   d. 114
   e. 115
   f. 221
   g. *Others*
   h. *Unknown*
5. *PRODUCT_SUBCLASS,* yaitu subkelas produk yang dibeli oleh pelanggan.
6. *PRODUCT_ID,* yaitu id produk yang dibeli pelanggan.
7. *AMOUNT,* yaitu jumlah produk yang dibeli oleh pelanggan.
8. *ASSET,* yaitu aset.
9. *SALES_PRICE,* yaitu total harga produk yang dibayar oleh pelanggan.

Dataset tidak memiliki penjelasan yang rinci tentang transaksi pelanggan, seperti mata uang apa yang digunakan untuk atribut *SALES_PRICE* (harga jual) dan unit apa yang digunakan untuk atribut *AMOUNT* atau *ASSET*.

3.2 *Select Data*

Dari total 9 atribut yang terdapat dalam *dataset* akan dipilih beberapa atribut yang akan digunakan untuk pengelompokan pelanggan. Berdasarkan kebutuhan analisis RFM, atribut yang akan digunakan adalah *TRANSACTION_DT, CUSTOMER_ID,* dan *SALES_PRICE*. Data yang dipilih akan dioperasikan dengan menggunkan *python.* Pemilihan data akan dilakukan berdasarkan *header data.* Contoh dari hasil pemilihan atribut dapat dilihat pada Gambar 2.

	TRANSACTION_DT	CUSTOMER_ID	SALES_PRICE
0	11/1/2000	1104905	30
1	11/1/2000	418683	46
2	11/1/2000	1057331	166
3	11/1/2000	1849332	38
4	11/1/2000	1981995	18

Gambar 2. Hasil pemilihan atribut

3.3 *Clean Data*

Pembersihan data dilakukan untuk menghilangkan data yang bernilai *null* atau data yang tidak lengkap. Daftar atribut yang bernilai *null* akan ditampilkan pada Gambar 3.

```
#Check for missing values in the dataset
Tafeng_data.isnull().sum(axis=0)
```

```
TRANSACTION_DT          0
CUSTOMER_ID             0
AGE_GROUP           22362
PIN_CODE                0
PRODUCT_SUBCLASS        0
PRODUCT_ID              0
AMOUNT                  0
ASSET                   0
SALES_PRICE             0
dtype: int64
```

Gambar 3. Daftar atribut bernilai *null*

Atribut TRANSACTION_DT, CUSTOMER_ID, dan SALES_PRICE tidak memiliki nilai *null* sehingga tidak dilakukan pembersihan data.

3.4 *Format Data*

Untuk memudahkan pengolahan data untuk pengelompokan, dilakukan reformat tipe atribut *TRANSACTION_DT* dari *string* menjadi *datetime*. Tampilan perubahan tipe atribut akan ditampilkan pada Gambar 4.

```
#Convert the string date field to datetime
Tafeng_data['TRANSACTION_DT'] = pd.to_datetime(Tafeng_data['TRANSACTION_DT'])
```

```
[21] Tafeng_data.head()
```

	TRANSACTION_DT	CUSTOMER_ID	AGE_GROUP	PIN_CODE	PRODUCT_SUBCLASS	PRODUCT_ID	AMOUNT	ASSET	SALES_PRICE
0	2000-11-01	1104905	45-49	115	110411	4710199010372	2	24	30
1	2000-11-01	418683	45-49	115	120107	4710857472535	1	48	46
2	2000-11-01	1057331	35-39	115	100407	4710043654103	2	142	166
3	2000-11-01	1849332	45-49	Others	120108	4710126092129	1	32	38
4	2000-11-01	1981995	50-54	115	100205	4710176021445	1	14	18

Gambar 4. Tampilan reformat data pada tipe atribut *TRANSACTION_DT*

Setelah reformat tipe data, maka dilakukan penggabungan data dengan data mentah dari Ta-Feng Grocery. Hal ini bertujuan agar pemahaman data dapat dilakukan dengan dengan baik.

BAB 4. MODELLING

Pada bab ini menjelaskan tentang tahap pemodelan dengan menggunakan teknik penambangan data.

4.1 *Select Modelling Technique*

Teknik penambangan data yang dipilih adalah *Clustering* dengan algoritma K-Means. *Clustering* sangat tepat digunakan untuk mencapai tujuan awal proyek ini yaitu memeperoleh pengetahuan tentang pengelompokan pelanggan Toko XYZ dengan analisis RFM. Pemodelan data akan diawali dengan penentuan nilai *Recency*, *Frequency*, dan *Monetary* untuk pembentukan analisis RFM.

4.2 *Build Model*

Untuk melakukan pengelompokan pelanggan, maka dilakukan proses untuk pembangunan model secara bertahap.

4.2.1 RFM Model

Analisis RFM merupakan proses analisis perilaku pelanggan. Untuk menentukan kelompok pelanggan maka dilakukan pencarian nilai untuk tiga variabel, yaitu *Recency*, *Frequency*, dan *Monetary*. Nilai *Recency* merupakan selisih antara waktu transaksi pelanggan, yaitu 1 November 2000 dengan waktu terakhir melakukan transaksi. Atribut yang dibutuhkan pada pencarian nilai *Recency* adalah *TRANSACTION_DT* dan *CUSTOMER_ID*. Hasil beberapa nilai *Recency* untuk setiap pelanggan dapat dilihat pada Gambar 5.

	CUSTOMER_ID	Recency
0	1069	19
1	1113	54
2	1250	19
3	1359	87
4	1823	36

Gambar 5. Tampilan nilai *Recency*

Tampilan nilai *Recency* pada Gambar 5 diurutkan berdasarkan id pelanggan. Nilai *Recency* tersebut merupakan jumlah waktu (hari) transaksi yang dilakukan oleh pelanggan pada Toko XYZ. Semakin besar nilai rentang waktunya, maka nilai *Recency* semakin besar.

Nilai *Frequency* merupakan jumlah transaksi yang dilakukan oleh pelanggan. Atribut yang digunakan pada pencarian nilai *Frequency* adalah *CUSTOMER_ID*, nilai *Frequency* akan dihitung berdasarkan setiap id yang mucul pada *dataset*. Hasil beberapa nilai *Frequency* dapat dilihat pada Gambar 6.

	CUSTOMER_ID	Frequency
0	1069	11
1	1113	18
2	1250	14
3	1359	3
4	1823	14

Gambar 6. Tampilan nilai *Frequency*

Tampilan nilai *Frequency* pada Gambar 6 diurutkan berdasarkan id pelanggan. Nilai *Frequency* tersebut merupakan jumlah transaksi yang dilakukan atau berapa kali transaksi yang dilakukan oleh pelanggan pada Toko XYZ. Semakin banyak jumlah transaksi yang dilakukan, maka nilai *Frequency* semakin besar.

Nilai *Monetary* adalah nilai transaksi yang dilakukan pelanggan berupa uang yang dikeluarkan selama berbelanja. Atribut yang digunakan untuk mencari nilai *Monetary* adalah *CUSTOMER_ID* dan *SALES_PRICE.* Hasil beberapa nilai *Monetary* untuk setiap pelanggan dapat dilihat pada Gambar 7.

	CUSTOMER_ID	Monetary
0	1069	1944
1	1113	2230
2	1250	1583
3	1359	364
4	1823	2607

Gambar 7. Tampilan nilai *Monetary*

Tampilan nilai *Monetary* pada Gambar 7 diurutkan berdasarkan id pelanggan, dengan asumsi bahwa mata uang yang digunakan adalah dalam US$. Nilai *Monetary* tersebut merupakan nilai transaksi yang dilakukan oleh pelanggan berupa uang yang dikeluarkan pada Toko XYZ. Semakin banyak nilai uang yang dikeluarkan, maka nilai *Monetary* semakin besar.

Setelah dilakukan pencarian nilai RFM, maka dilakukan penggabungan tabel untuk nilai RFM. Hasil nilai RFM dapat dilihat pada Gambar 8.

CUSTOMER_ID	Recency	Frequency	Monetary
1069	19	11	1944
1113	54	18	2230
1250	19	14	1583
1359	87	3	364
1823	36	14	2607

Gambar 8. Nilai RFM

Nilai RFM yang diperoleh diurutkan berdasarkan id pelanggan. Berdasarkan nilai RFM yang didapatkan dari perhitungan pada *dataset*, maka dilakukan pengelompokan nilai RFM dengan menggunakan *quantile*. Gambar 9 akan menampilkan fungsi yang mendapatkan pengelompokan nilai RFM.

```
#Functions to create R, F, and M segments
def RScoring(x,p,d):
  if x<= d[p][0.25]:
      return 1
  elif x <= d[p][0.50]:
      return 2
  elif x <= d[p][0.75]:
       return 3
  else:
        return 4
def FnMScoring(x,p,d):
  if x<= d[p][0.25]:
      return 4
  elif x <= d[p][0.50]:
      return 3
  elif x <= d[p][0.75]:
      return 2
  else:
       return 1
```

Gambar 9. Fungsi pengelompokan nilai RFM menggunakan *quantile*

Berdasarkan pengelompokan RFM yang dilakukan dengan fungsi diatas, maka didapatkan tabel pengelompokan RFM yang dapat dilihat pada Gambar 10.

CUSTOMER_ID	Recency	Frequency	Monetary	R	F	M	RFMCluster	RFMScore
1069	19	11	1944	2	3	2	232	7
1113	54	18	2230	3	2	2	322	7
1250	19	14	1583	2	3	3	233	8
1359	87	3	364	4	4	4	444	12
1823	36	14	2607	3	3	2	332	8

Gambar 10. Hasil pengelompokan RFM dengan *quantile*

Untuk memudahkan dalam menganalisis perilaku pelanggan, maka setiap pelanggan diberi label untuk mengetahui ukuran perilaku pelanggan terhadap Toko XYZ. Label yang diterapkan terdiri dari 5 label, yaitu sebagai berikut.

1. *Superstars*, yaitu pelanggan yang sering melakukan transaksi dan memiliki nilai transaksi yang tinggi.
2. *Golden*, yaitu pelanggan yang sering melakukan transaksi kedua yang tinggi.

3. *Occasional*, pelanggan yang jarang berkunjung tetapi memiliki nilai transaksi rata-rata yang tinggi.
4. *Everyday*, yaitu pelanggan yang sering melakukan transaksi tetapi nilai transaksi yang rendah.
5. *Dormant*, yaitu pelanggan dengan tingkat pembelian yang sangat rendah dan sudah lama tidak melakukan transaksi.

Perilaku pelanggan yang diterapkan dapat dilihat pada Gambar 11.

```
#Make loyalty level to each customer -- class label
loyalty_level = ['Dormant', 'Everyday', 'Occasional', 'Golden', 'Superstars']
score_cuts = pd.qcut(RFMScoring.RFMScore, q = 5, labels = loyalty_level)
RFMScoring['RFMLoyalty_Level'] = score_cuts.values
RFMScoring.reset_index().head()
```

	CUSTOMER_ID	Recency	Frequency	Monetary	R	F	M	RFMCluster	RFMScore	RFMLoyalty_Level
0	1069	19	11	1944	2	3	2	232	7	Everyday
1	1113	54	18	2230	3	2	2	322	7	Everyday
2	1250	19	14	1583	2	3	3	233	8	Occasional
3	1359	87	3	364	4	4	4	444	12	Superstars
4	1823	36	14	2607	3	3	2	332	8	Occasional

Gambar 11. Tampilan RFM Model

4.2.2 K-Means Model

Dalam proses *Clustering*, terdapat 2 tahap proses yang dilakukan yaitu menentukan nilai *k* dengan metode Elbow dan melakukan pengelompokan dengan algoritma K-Means.

a. Penentuan nilai *k* dengan Metode Elbow

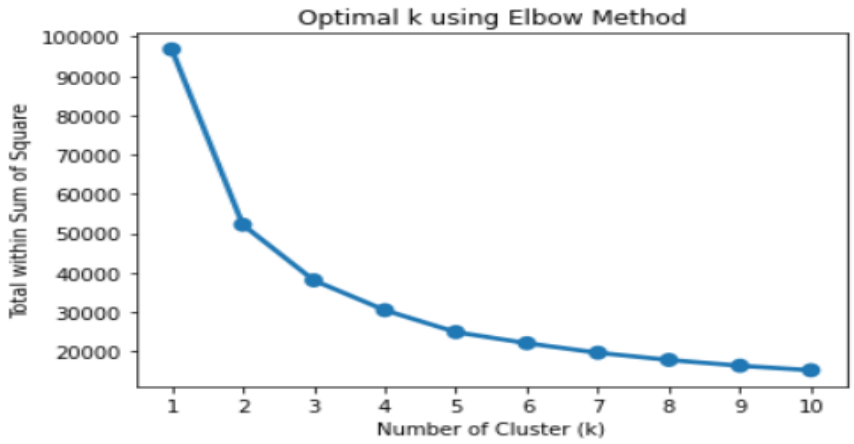
Dataset yang digunakan untuk penentuan nilai *k* adalah data yang telah melalui pra proses data. Setelah melalui pra proses data, maka dilakukan pemilihan jumlah *cluster* atau nilai *k* dengan menggunakan metode Elbow. Metode ini memilih jumlah *cluster* dengan melihat nilai SSE dan titik dari nilai SSE sudah mulai stabil (tidak turun terlalu terlalu signifikan). Titik tersebut yang menjadi titik siku pada grafik. Proses dari metode Elbow ini menggunakan percobaan jumlah *cluster* antara 1 sampai 10. Fungsi metode Elbow dapat dilihat pada Gambar 12.

```
from sklearn.cluster import KMeans
sum_of_sq_dist = {}
for k in range(1,11):
  km = KMeans(n_clusters=k, init= 'k-means++', max_iter=1000)
  km = km.fit(Scaled_Data)
  sum_of_sq_dist[k] = km.inertia_

#plot the graph for the sum of square distance values and Number of Clusters
sns.pointplot(x=list(sum_of_sq_dist.keys()), y = list(sum_of_sq_dist.values()))
plt.xlabel('Number of Cluster (k)')
plt.ylabel('Total within Sum of Square')
plt.title('Optimal k using Elbow Method')
plt.show()
```

Gambar 12. Fungsi metode Elbow

Hasil dari proses metode Elbow berupa plot yang digunakan untuk mengetahui titik siku yang terbentuk. Jumlah plot dapat dilihat pada grafik melalui Gambar 13.



Gambar 13. Plot hasil penentuan jumlah *k*

Pada tujuan awal bisnis, pengelompokan dilakukan berdasarkan 5 *cluster*. Namun, berdasarkan hasil penentuan jumlah *cluster* dengan metode Elbow, nilai *k* yang sesuai dengan kriteria metode Elbow adalah sebanyak 2 *cluster*.

b. Pengelompokan dengan K-Means

Tujuan awal bisnis adalah membangun dengan 5 *cluster*. Jumlah *cluster* tidak sesuai dengan perolehan pada metode Elbow. Namun, meskipun belum sesuai, hasil pengelompokan tetap ditampilkan pada proyek. Hasil pengelompokan dengan 5 cluster dapat dilihat pada Gambar 14.

CUSTOMER_ID	Recency	Frequency	Monetary	R	F	M	RFMCluster	RFMScore	RFMLoyalty_Level	Cluster	Color
1069	19	11	1944	2	3	2	232	7	Everyday	3	blue
1113	54	18	2230	3	2	2	322	7	Everyday	2	red
1250	19	14	1583	2	3	3	233	8	Occasional	3	blue
1359	87	3	364	4	4	4	444	12	Superstars	0	red
1823	36	14	2607	3	3	2	332	8	Occasional	3	blue

Gambar 14. Pengelompokan menggunakan K-Means dengan n=5

CUSTOMER_ID	Recency	Frequency	Monetary	R	F	M	RFMCluster	RFMScore	RFMLoyalty_Level	Cluster
1069	19	11	1944	2	3	2	232	7	Everyday	0
1113	54	18	2230	3	2	2	322	7	Everyday	1
1250	19	14	1583	2	3	3	233	8	Occasional	0
1359	87	3	364	4	4	4	444	12	Superstars	1
1823	36	14	2607	3	3	2	332	8	Occasional	0

Gambar 15. Pengelompokan menggunakan K-Means dengan *k* = 2

Setelah mendapatkan nilai *k*, maka dilakukan proses *Clustering*. Proses ini dimulai dengan menggunakan fungsi K-Means dan memasukkan nilai *k* sebanyak 2. Hasil proses K-Means dapat ditampilkan dengan menggunakan fungsi K-Means tersebut. Pengelompokan dengan K-Means dapat dilihat pada Gambar 15. Untuk mendapatkan hasil yang lebih mudah dipahami, maka dilakukan penggabungan antara data dengan hasil *clustering*.

4.3 *Assess Model*

Berdasarkan hasil pengelompokan dengan K-Means, setiap pelanggan akan dihitung jaraknya ke *centroid*. Hasil penghitungan centroid dapat dilihat pada Gambar 16.

```
#Validate the K-Means cluster with centroid value
KMeans_clust.cluster_centers_
```

```
array([[-0.53580418,  0.75398879,  0.72423099],
       [ 0.53700114, -0.75567316, -0.72584889]])
```

Gambar 16. Perhitungan nilai *centroid* dengan *k* = 2

Perhitungan ini dilakukan untuk mengetahui bahwa setiap pelanggan tepat berada pada kelompok yang dibentuk. Validasi model dilakukan dengan menggunakan rumus *Euclidean Distance*.

BAB 5. EVALUATION

Pada bab ini menjelaskan tentang evaluasi model yang dihasilkan pada tahap *Modelling*.

5.1 *Evaluate Results*

Data hasil pengelompokan dengan menggunakan K-Means yang awalnya membagi pelanggan pada 5 segmen diuji untuk mengetahui apakah pengelompokan tersebut telah optimal. Metode yang diterapkan dalam dalam pengujian performa yaitu metode *Silhouette Coefficient*. Dengan pengujian diambil 100 data secara acak dinilai secara manual dan dibandingkan dengan hasil pemodelan yang dilakukan. Fungsi *Silhouette Coefficient* dan hasil evaluasi pengelompokan dapat dilihat pada Gambar 17.

```
#Calculating the silhouette score
#k=5
km5 = KMeans(n_clusters=5, init='k-means++', max_iter=1000)
y_means = km5.fit_predict(z)
silhouette_scores = silhouette_score(z, y_means)
print(f'Silhouette Score (n=5) : {silhouette_score(z, y_means)}')
```

```
Silhouette Score (n=5) : 0.40308754956423665
```

Gambar 17. Fungsi *Silhouette Coefficient* dan hasil evaluasi model dengan 5 *cluster*

Hasil evaluasi model dengan 5 *cluster* sebesar 0,40. Hasil ini menunjukkan bahwa model yang dihasilkan belum optimal untuk pengelompokan data. Berdasarkan metode Elbow yang digunakan untuk penentuan nilai *k*, maka jumlah *cluster* yang tepat adalah sebanyak 2 *cluster*. Fungsi *Silhouette Coefficient* dan hasil evaluasi pengelompokan dengan 2 *cluster* dapat dilihat pada Gambar 18.

```
#Calculating the silhouette score
#k=2
km2 = KMeans(n_clusters=2, init='k-means++', max_iter=1000)
y_means = km2.fit_predict(z)
silhouette_scores = silhouette_score(z, y_means)
print(f'Silhouette Score (n=2) : {silhouette_score(z, y_means)}')
```

```
Silhouette Score (n=2) : 0.82528433985263
```

Gambar 18. Fungsi *Silhouette Coefficient* dan hasil evaluasi model dengan 2 *cluster*

Hasil evaluasi model dari 2 *cluster* sebesar 0,82. Hasil dari pengujian menghasilkan nilai yang baik karena pada nilai *Silhouette Coefficient* dengan prinsip semakin besar nilainya (mendekati 1) maka semakin baik suatu *centroid* digunakan sebagai solusi *clustering*.

5.2 *Review Process*

Pemodelan yang dilakukan belum sesuai dengan tujuan pengimplementasian penambangan data. Pengelompokan pelanggan dengan 5 *cluster* menghasilkan evaluasi yang kurang baik. Sehingga, perbaikan jumlah *cluster* ditentukan dengan menggunkan metode Elbow. Nilai *k* yang diperoleh dari metode tersebut adalah sebanyak 2 *cluster*. Hasil pembagian pelanggan dengan 5 dan 2 *cluster* dapat dilihat pada Gambar 19 dan Gambar 20.

```
Tafeng_data = pd.DataFrame(Scaled_Data)
Tafeng_data['Cluster'] = pred
Tafeng_data['Cluster'].value_counts()
```

```
1    8986
4    8079
2    5437
0    5057
3    4707
Name: Cluster, dtype: int64
```

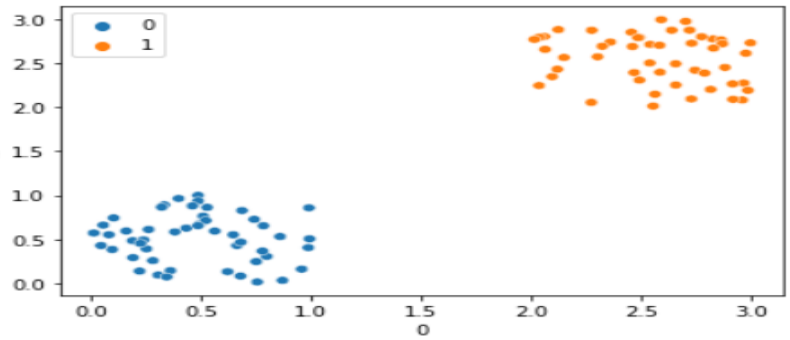
Gambar 19. Total pelanggan pada masing-masing *cluster* dengan *k* = 5

Pada hasil perhitungan dengan 5 *cluster*, jumlah pelanggan pada label *Superstars* = 5057 pelanggan, *Golden* = 8986 pelanggan, *Occasional* = 5437 pelanggan, *Everyday* = 4707 pelanggan, dan *Dormant* = 8079 pelanggan.

```
Tafeng_data = pd.DataFrame(Scaled_Data)
Tafeng_data['Cluster'] = pred
Tafeng_data['Cluster'].value_counts()
```

```
0    18750
1    13516
Name: Cluster, dtype: int64
```

Gambar 20.Total pelanggan pada masing-masing cluster dengan *k* = 2



Gambar 21. *Scatter plot* untuk pengelompokan pelanggan dengan 2 *cluster*

Hasil pembagian menunjukkan total pelanggan pada masing-masing *cluster*. *Cluster* pertama sebanyak 18.750 pelanggan dan *cluster* kedua sebanyak 13.516 pelanggan. Penyebaran data dengan 2 *cluster* tersebut dapat dilihat dengan *plot* pada Gambar 21. Berdasarkan hasil tersebut, maka tidak ada faktor penting dalam proses yang terabaikan atau terlewati.

5.3 *Determine Next Steps*

Untuk selanjutnya, inisial label yang diterapkan pada *dataset* akan ditinjau kembali. Kemudian label pada pengelompokan pelanggan dengan 2 *cluster* akan ditentukan oleh *marketing* Toko XYZ dalam membantu keputusan strategi CRM untuk meningkatkan pemasaran. Langkah yang akan dilakukan selanjutnya adalah lanjut ke tahap *Deployment*.

BAB 6. DEPLOYMENT

Bab ini akan menjelaskan rencana penyebaran, *monitoring,* dan *maintanance* berdasarkan proyek yang telah dikerjakan.

6.1 *Deployment Plan*

Deployment merupakan fase penyusunan laporan atau presentasi dari pengetahuan yang didapat dari evaluasi pada proses penambangan data. Rencana penyebaran dapat diidentifikasi sebagai tahap konsolidasi mengenai langkah apa yang diambil setelah pemodelan penambangan data diperoleh. Laporan akhir disusun dengan menyusun langkah-langkah pengerjaan proyek dengan metode CRISP-DM. Kemudian hasil yang diperoleh mengenai pengetahuan yang didapat atau pengenalan pola pada data dalam proses penambangan data dan dipresentasikan dalam bentuk visualisasi gambar atau deskripsi yang mudah dipahami.

6.2 *Plan Monitoring and Maintanance*

Pada proyek ini diberikan hasil bahwa dengan menggunakan konsep *clustering* pada penambangan data yang diterapkan dalam aktivitas *customer* pada Toko XYZ dengan menggunakan CRM dapat memberikan kemampuan untuk melakukan pengelompokan berdasarkan kapan terakhir kali pelanggan melakukan transaksi, sering atau tidak melakukan transaksi dan jumlah transaksi yang dilakukan. Pada proyek ini, peneliti menggunakan algoritma K-Means dan berhasil mengelompokkan *customer* berdasarkan kategori yang sudah ditentukan dengan hasil akhir model mencapai 0,82 yang artinya nilai tersebut baik dengan mengikuti prinsip *Silhouette Coefficient.* Dengan mengunakan metode Elbow, peneliti mendapatkan hasil untuk *cluster* yaitu 2 *cluster* dengan nilai dari masing-masing *cluster* adalah 18.750 pelanggan dann 13.516 pelanggan. Terdapat kelemahan pada sistem ini, yaitu hasil kalkulasi yang berubah setiap dilakukan di-*run.* Sehingga dibutuhkan perbaikan dan pemeliharaan pada sistem agar mendapatkan hasil yang akurat.

BAB 7. KESIMPULAN DAN SARAN

Pada bab ini menjelaskan hasil yang diperoleh dan saran yang diberikan pada pengerjaan proyek.

7.1 Kesimpulan

Berikut adalah kesimpulan dan hasil yang diperoleh.

1. Berdasarkan analisis RFM dengan 5 *cluster*, jumlah pelanggan pada label *Superstars* = 5057 pelanggan, *Golden* = 8986 pelanggan, *Occasional* = 5437 pelanggan, *Everyday* = 4707 pelanggan, dan *Dormant* = 8079 pelanggan.
2. Evaluasi model menggunakan *Silhouette Coefficient* dengan jumlah *cluster* sebangank 5 adalah 0,40. Hasil ini masih kurang baik untuk digunakan sebagai solusi pengelompokan pelanggan.
3. Berdasarkan penerapan metode Elbow, jumlah *k* yang sesuai digunakan adalah sebanyak 2 dan evaluasi model dengan jumlah *cluster* tersebut adalah 0,82. Evaluasi menggunakan *Silhouette Coefficient*.
4. Pada hasil perhitungan dengan 2 *cluster*, *cluster* pertama sebanyak 18.750 pelanggan dan *cluster* kedua sebanyak 13.516 pelanggan.
5. Pengelompokan pelanggan dengan hasil analisis RFM dapat ditinjau kembali untuk menentukan strategi pemasaran berdasarkan perilaku pelanggan terhadap transaksi pada Toko XYZ.

7.2 Saran

Saran yang dapat diterapkan dari hasil proyek ini adalah untuk proyek selanjutnya diharapkan peneliti dapat menggunakan beberapa algoritma *clustering* sebagai perbandingan untuk mendapatkan metode yang memberikan hasil paling baik.

DAFTAR PUSTAKA

[1] E. Haddi, X. Liu, and Y. Shi, "Sentiment Analysis of Hotel Review using Naïve Bayes Algorithm and Integration of Information Gain and Genetic Algorithm as Feature Selection Methods," *Int. Semin. Sci. Issues Trends Bekasi*, 2014.

[2] Subburathinam, K. (2016). *Customer Segmentation Framework using Redefined*. *May 2015*.