Nama : Rizky Fauzi Ramadhani

Kelas : IF-42-10

NIM : 1301184144

Formulasi Masalah

Terdapat 2 dataset yang diberikan yaitu, salju dan kendaraan. Dataset salju diminta untuk mencari tahu apakah besok akan turun salju atau tidak, sedangkan dataset kendaraan diminta untuk mencari tahu apakah pembeli tertarik untuk membeli kendaraan baru atau tidak. Dataset tersebut digunakan dengan menggunakan clustering K-Mean. Dataset yang akan digunakan yaitu dataset kendaraan untuk mencari tahu apakah pembeli tertarik untuk membeli kendaraan baru atau tidak.

Eksplorasi dan Persiapan Data

Saya menggunakan dataset Kendaraan, kemudian eksplorasi data yang saya lakukan adalah dengan melihat info data, penguraian data agar dapat melihat count, mean, std, min, max, dll. Kemudian saya mengecek missing values pada dataset, kemudian saya melakukan preprocessing pada data dengan mendrop/menghapus baris yang memiliki isi yang kosong, agar nanti dataset dapat diproses dengan sangat baik, kemudian saya melakukan encode data untuk mengubah dataset yang bersifat objek agar menjadi numerik agar lebih mudah diproses dengan menggunakan library label encoder, sehingga mengubah data yang awalnya pria, wanita menjadi 0 dan 1 kemudian encode umur_kendaraan dan juga encode kendaraan_rusak. Kemudian saya mengecek korelasi heatmap dataset dengan tujuan untuk mengambil 2 kolom yang akan dicluster, dengan melihat warna yang lebih terang di luar dari yang bernilai 1 itu merupakan korelasi heatmap yang baik untuk di cluster, dan saya mendapatkan korelasi yang bagus yaitu kendaraan_rusak dan sudah_asuransi dengan hasil 0.83. Selanjutnya sebelum ke pemodelan saya melakukan terlebih dahulu normalisasi pada dataset dengan tujuan mengubah nilai dari range 0 – 1 agar tidak terjadinya timpang tindih data.

Info dataset

```
[110] df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 285831 entries, 0 to 285830
Data columns (total 12 columns):
 #   Column             Non-Null Count   Dtype  
---  ------             --------------   -----  
 0   id                 285831 non-null  int64  
 1   Jenis_Kelamin      271391 non-null  object 
 2   Umur               271617 non-null  float64
 3   SIM                271427 non-null  float64
 4   Kode_Daerah        271525 non-null  float64
 5   Sudah_Asuransi     271602 non-null  float64
 6   Umur_Kendaraan     271556 non-null  object 
 7   Kendaraan_Rusak    271643 non-null  object 
 8   Premi              271262 non-null  float64
 9   Kanal_Penjualan    271532 non-null  float64
 10  Lama_Berlangganan  271839 non-null  float64
 11  Tertarik           285831 non-null  int64  
dtypes: float64(7), int64(2), object(3)
memory usage: 26.2+ MB
```

Sum missing value

```
id                       0
Jenis_Kelamin        14440
Umur                 14214
SIM                  14404
Kode_Daerah          14306
Sudah_Asuransi       14229
Umur_Kendaraan       14275
Kendaraan_Rusak      14188
Premi                14569
Kanal_Penjualan      14299
Lama_Berlangganan    13992
Tertarik                 0
dtype: int64
```

encode

```
['Pria' 'Wanita']
[0 1]

['1-2 Tahun' '< 1 Tahun' '> 2 Tahun']
[0 1 2]

['Pernah' 'Tidak']
[0 1]
```

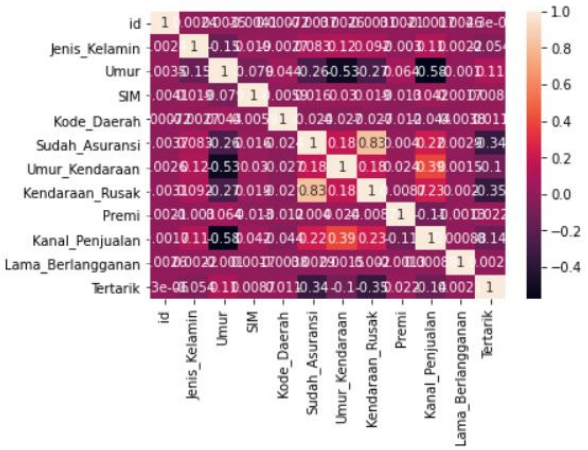
normalisasi

	0	1	2	3	4	5	6	7	8	9	10	11
0	0.000000	1.0	0.153846	1.0	0.634615	1.0	0.5	1.0	0.047251	0.932099	0.301038	0.0
1	0.000003	0.0	0.430769	1.0	0.750000	0.0	1.0	0.0	0.043104	0.172840	0.512111	0.0
2	0.000010	1.0	0.584615	1.0	0.923077	0.0	0.0	1.0	0.000000	0.759259	0.183391	0.0
3	0.000017	0.0	0.015385	1.0	0.673077	1.0	0.5	1.0	0.037402	0.932099	0.557093	0.0
4	0.000028	1.0	0.000000	1.0	0.153846	1.0	0.5	1.0	0.052380	0.981481	0.072664	0.0

Dataset describe

	id	Umur	SIM	Kode_Daerah	Sudah_Asuransi	Premi	Kanal_Penjualan	Lama_Berlangganan	Tertarik
count	285831.000000	271617.000000	271427.000000	271525.000000	271602.000000	271262.000000	271532.000000	271839.000000	285831.000000
mean	142916.000000	38.844336	0.997848	26.405410	0.458778	30536.683472	112.021567	154.286302	0.122471
std	82512.446734	15.522487	0.046335	13.252714	0.498299	17155.000770	54.202457	83.694910	0.327830
min	1.000000	20.000000	0.000000	0.000000	0.000000	2630.000000	1.000000	10.000000	0.000000
25%	71458.500000	25.000000	1.000000	15.000000	0.000000	24398.000000	29.000000	82.000000	0.000000
50%	142916.000000	36.000000	1.000000	28.000000	0.000000	31646.000000	132.000000	154.000000	0.000000
75%	214373.500000	49.000000	1.000000	35.000000	1.000000	39377.750000	152.000000	227.000000	0.000000
max	285831.000000	85.000000	1.000000	52.000000	1.000000	540165.000000	163.000000	299.000000	1.000000

Correlation



Pemodelan

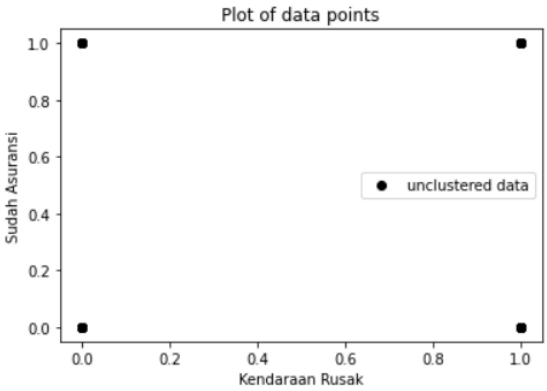
Pertama, dilakukan terlebih dahulu encode untuk mengubah nilai pada kolom yang non-numerik menjadi numerik. Kedua, dilakukan normalisasi untuk menghindari terjadinya tumpang tindih data dengan mengubah nilai dan range data dari kisaran 0 – 1. Ketiga, melakukan clustering menggunakan algoritma K-mean dengan memetakan data dengan 2 kolom berdasarkan korelasi yang didapat (kendaraan_rusak dan sudah_asuransi) sehingga membentuk node-node. Keempat, inisialisasi secara acak pusat dari cluster. Dengan ansumsi awal nilai K = 5 yang saya ambil, jadi kita memilih 5 titik data secara acak sebagai centroid. Kemudian untuk setiap titik data dihitung jarak Euclid dari semua centroid dengan rumus sebagai berikut :

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

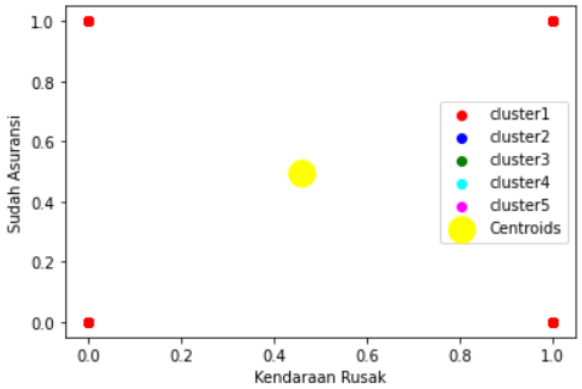
Kemudian menetapkan cluster berdasarkan jarak minimal ke semua centroid. Kemudian mengambil setiap titik hitam, lalu menghitung jarak eucliddiannya dari semua centroid. Dan kemudian mewarnai titik hitam dengan warna titik terdekatnya. Kelima, menyesuaikan pusat massa setiap cluster dengan mengambil rata-rata dari semua titik data yang termasuk dalam

cluster tersebut, setelah itu menghitung mean dari semua cluster individu untuk menetapkan semua titik data ke salah satu cluster. Sehingga didapatkan hasil cluster dari dataset kendaraan berdasarkan algoritma K-mean.

Sebelum clustering :

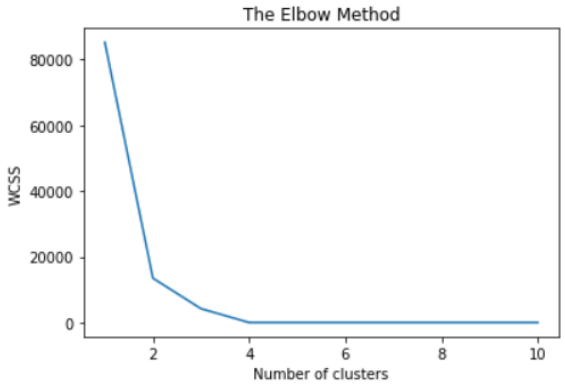


Sesudah clustering :



Evaluasi

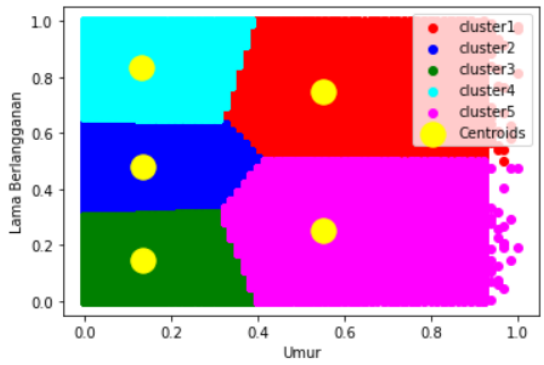
Evaluasi yang saya lakukan adalah dengan Teknik siku (elbow method). Langkah-langkah yang dilakukan oleh elbow method yaitu pengelompokan pada nilai K yang berbeda mulai dari range 1- 10. Untuk setiap K, dihitung WCSSnya. Kemudian memplot nilai WCSS sesuai dengan jumlah kluster. Gambar pada plot yang terdapat siku/elbow merupakan cluster yang paling sesuai. Yang dimana saya mengambil nilai K = 5 pada elbow method menunjukkan bahwa nilai K = 3 lebih optimal.



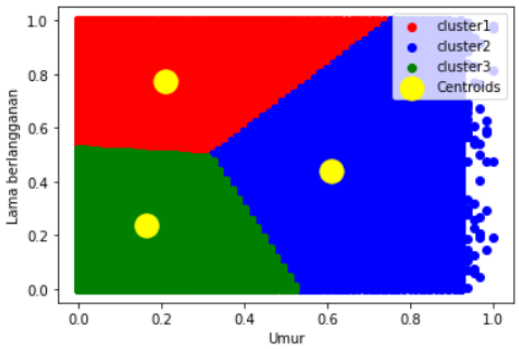
Eksperimen

Eksperimen yang saya lakukan mengubah nilai K dan mengubah kolom yang akan dicluster, dan mendapatkan hasil sebagai berikut :

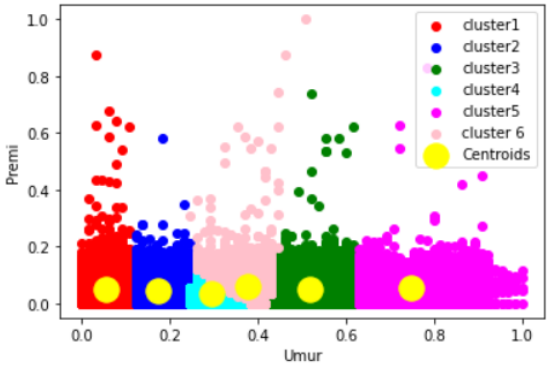
Nilai K = 5 , Kolom Umur dan Lama Berlangganan



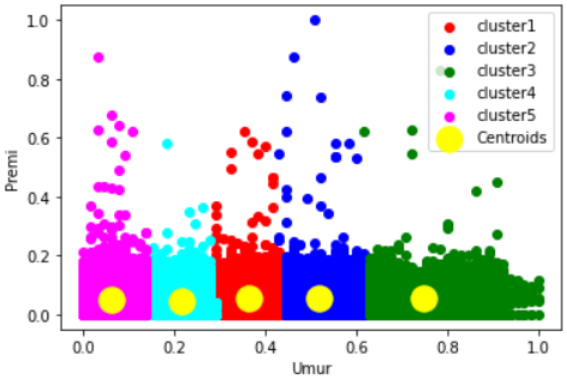
Nilai k = 3, kolom umur dan lama berlangganan



Nilai k = 6, kolom umur dan premi



Nilai k = 5, kolom umur dan premi



Kesimpulan

Kesimpulan yang saya dapatkan berdasarkan eksperimen clustering yang saya lakukan, disimpulkan bahwa kolom dengan nilai yang unik akan lebih optimum yaitu adalah kolom umur dan lama berlangganan.

Link Video Presentasi

https://youtu.be/cH7H6UgHdik

Link Colab

https://colab.research.google.com/drive/18VEihICP2PE_BreZjiGnIoCBo96sVv3L?usp=sharing