Laporan Tugas Besar Tahap Satu (Clustering)

Laporan ini dibuat untuk memenuhi tugas besar

Mata kuliah Pembelajaran Mesin

Disusun oleh:

Muhamad Farell Ambiar (1301184262)

S1 INFORMATIKA

FAKULTAS INFORMATIKA

UNIVERSITAS TELKOM BANDUNG 2021

Pendahuluan

Clustering adalah proses membagi/mengelompokan suatu populasi atau titik data menjadi beberapa kelompok. Pengelompokan data dilakukan berdasarkan kemiripan nilai satau sifat dari tiap-tiap titik data pada populasi yang akan dianalisa. Pada tugas besar tahap satu ini, penerapan clustering digunakan untuk mengelompokkan data ketertarikan pada dataset kendaraan_train.csv dan kendaraan_test.csv.

Dataset ini memiliki total 11 feature atau kolom, antara lain adalah:

- Jenis_Kelamin
- Umur
- SIM
- Kode_Daerah
- Sudah_Asuransi
- Umur_Kendaraan
- Kendaraan_Rusak
- Premi
- Kanal_Penjualan
- Lama_Berlangganan
- Tertarik

	Jenis_Kelamin	Umur	SIM	Kode_Daerah	Sudah_Asuransi	Umur_Kendaraan	Kendaraan_Rusak	Premi	Kanal_Penjualan	Lama_Berlangganan	Tertarik
0	Wanita	30.0	1.0	33.0	1.0	< 1 Tahun	Tidak	28029.0	152.0	97.0	0
1	Pria	48.0	1.0	39.0	0.0	> 2 Tahun	Pernah	25800.0	29.0	158.0	0
2	NaN	21.0	1.0	46.0	1.0	< 1 Tahun	Tidak	32733.0	160.0	119.0	0
3	Wanita	58.0	1.0	48.0	0.0	1-2 Tahun	Tidak	2630.0	124.0	63.0	0
4	Pria	50.0	1.0	35.0	0.0	> 2 Tahun	NaN	34857.0	88.0	194.0	0
...	...	...	...	...	...	...	...	...	...	...	...
333465	Pria	61.0	1.0	46.0	0.0	> 2 Tahun	Pernah	31039.0	124.0	67.0	0
333466	Pria	41.0	1.0	15.0	0.0	1-2 Tahun	Pernah	2630.0	157.0	232.0	0
333467	Pria	24.0	1.0	29.0	1.0	< 1 Tahun	Tidak	33101.0	152.0	211.0	0
333468	Pria	59.0	1.0	30.0	0.0	1-2 Tahun	Pernah	37788.0	26.0	239.0	1
333469	Pria	52.0	1.0	31.0	0.0	1-2 Tahun	Tidak	2630.0	124.0	170.0	0

333470 rows × 11 columns

Hasil Observasi

Setelah melakukan pengamatan, analisis, dan design algoritma clustering untuk menentukan calon pembeli yang tertarik untuk membeli mobil baru, berikut adalah beberapa laporan mengenai analisis yang telah dilaksanakan:

Data Explorasi dan Data Praproses

- Data Merging

  Pada studi kasus clustering ini, data train dan data test digabungkan terlebih dahulu karena proses clustering tidak membutuhkan data test.

- Handling Missing Values
  Dataset awal bisa dibilang kurang berkualitas untuk dilakukan model clustering karena terdapat banyak sekali missing values pada hamper setiap feature. Oleh karena itu, perlu dilakukannya handling missing values.

```
Banyak Missing Values Tiap Column:

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
```

  Handling missing values dilakukan dengan melakukan imputasi pada missing values. Untuk missing values pada feature dengan jenis categorical akan dilakukan imputasi menggunakan Modus. Sedangkan untuk feature berjenis numerical dilakukan dengan imputasi Median atau Mean.

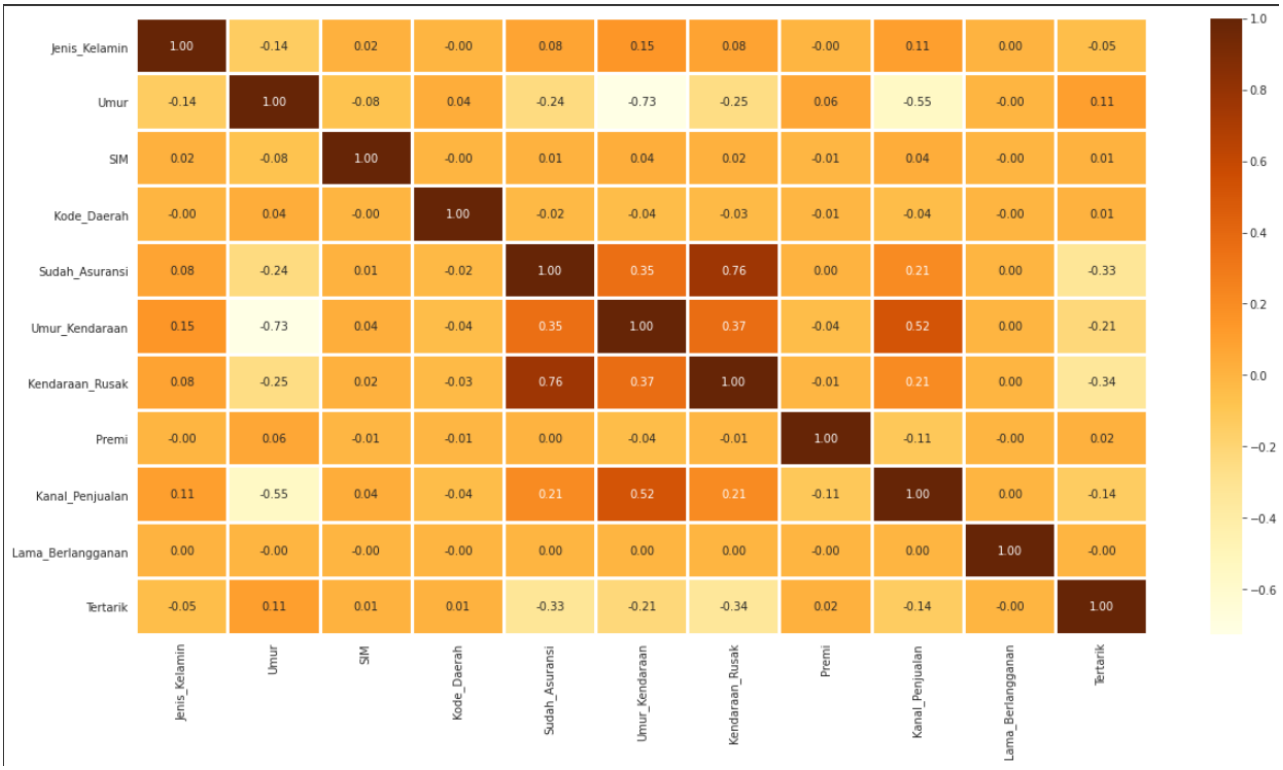
- Handling Duplicate data
  Duplicate data adalah beberapa instance yang memiliki nilai yang sama persis. Oleh karena itu dilakukan drop instance agar tidak mengganggu proses modelling nantinya.

- Data Transformation
  Melakukan encoding pada data-data yang bersifat categorical baik ordinal maupun non-ordinal menjadi data numerical. Proses ini dilakukan untuk mempermudah komputasi dalam modelling.

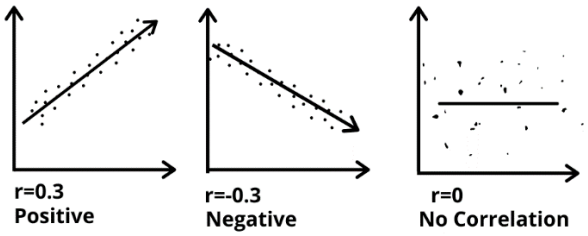
- Scaling Data
  Proses scaling data ini dilakukan agar range data tiap feature pada dataframe sama. Pada studi kasus ini digunakan method Min Max Scaler.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Feature Selection
  Setelah melakukan scaling data selanjutnya dilakuka pemilihan feature yang nantinya akan digunakan untuk proses modelling cluster. Pemilihan feature ini berdasarkan nilai correlation values setiap feature terhadap data target atau kolom “Tertarik”.

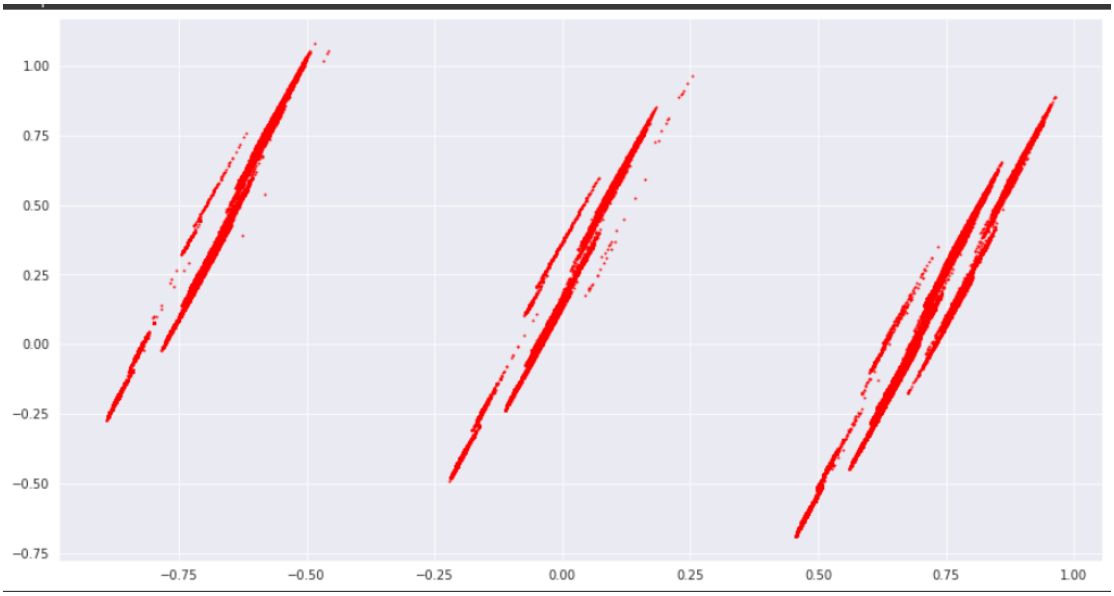


Dengan berdasarkan prinsip koefisien korelasi, maka feature yang memiliki correlation values paling mendekati 0 akan di drop. Dalam hal ini yang didrop adalah kolom Jenis_Kelamin, SIM, Kode_Daerah, Premi, dan Lama_Berlangganan.



- Demensional Reduction
  Tahap ini perlu dilakukan karena dalam proses pemodelan akan melibatkan beberapa feature yang sebelumnya sudah dipilih pada tahap feature selection. Ditahap ini data feature akan digabungkan menjadi hanya 2 feature. Proses ini dilakukan dengan menggunakan metode Principal Components Analysis (PCA).

- Visualisasi Distribusi Data
  Setelah dilakukan Praposes beriku adalah visualisasi sebaran datanya:



Cluster Modelling

Setelah selesai melakukan data eksplorasi dan data praproses, data yang sudah bersih siap untuk diproses kedalam model clustering. Berikut adalah tahap clustering yang dilakukan:

- Menentukan jumlah centroid
  Menentukan jumlah K berdasarkan grafik sebaran data.
- Inisialisasi letak centroid awal secara random berdasarkan grafik
- Lakukan grouping setiap data point ke centroid terdekat dengan perhitungan Euclidean distance

- Kemudian perbaharui centroid lama dengan centroid baru berdasarkan rata-rata data point pada cluster/group tsb.
- Lakukan hal tersebut diatas berulang-ulang hingga centroid tidak lagi berpindah.

Evaluasi Model

Evaluasi model unutk mengetahui apakah model yang dibuat sudah baik, menggunakan Silhouette method score. Hal ini karena method ini cocok untuk mengukur fitness dari cluster model.

```
Nilai Silhouette Method untuk n_cluster = 2 adalah 0.6112094467816658
Nilai Silhouette Method untuk n_cluster = 3 adalah 0.5438697301338059
Nilai Silhouette Method untuk n_cluster = 4 adalah 0.5689183446401981
```