Laporan Tugas Besar Tahap I (Clustering)

Laporan ini dibuat untuk memenuhi tugas besar

Mata kuliah Pembelajaran Mesin

Disusun oleh:

Mohammad Daffa Haris 1301180355

S1 INFORMATIKA

FAKULTAS INFORMATIKA

UNIVERSITAS TELKOM

BANDUNG

2021

BAB I Pendahuluan

Clustering merupakan tahapan yang perlu dilakukan sebelum melakukan klasifikasi di algoritma unsupervised learning. Clustering diperlukan karena nilai target output yang diperlukan dalam tahapan klasifikasi tidak diketahui. Hal ini terjadi karena tidak adanya label dalam dataset yang digunakan. sehingga dengan adanya clustering dapat dilakukan klasifikasi untuk melatih algoritma yang telah dibuat. Tujuan algoritma unsupervised learning adalah untuk menemukan pola yang sebelumnya tidak diketahui dalam data, tetapi pola-pola ini hanyalah sebuah perkiraan.

Dalam tugas besar tahap clustering ini didapatkan dataset berupa data ketertarikan seseorang terhadap sebuah kendaraan yang dijual. Dataset ini bernama kendaraan_test.csv dan kendaraan_train.csv, tetapi untuk tugas besar tahap I ini hanya berfokus pada dataset kendaraan_train.csv . Dataset kendaraan_train.csv ini memiliki 285.831 data dan 11 fitur. Adapun fitur-fitur dari dataset ini diantaranya adalah:

1. Jenis Kelamin
2. Umur
3. SIM
4. Kode Daerah
5. Sudah Asuransi
6. Umur Kendaraan
7. Kendaraan Rusak
8. Premi
9. Kanal Penjualan
10. Lama Berlangganan
11. Tertarik

```
 #   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
```

Gambar 1. Gambaran umum dataset kendaraan_train.csv

BAB II Hasil Penelitian

Dalam melakukan clustering kegiatan yang perlu dilakukan pertama kali adalah melakukan ekspor data ke dalam python. Untuk melakukan kegiatan ini digunakan library pandas. Dataset yang telah diekspor akan berbentuk berupa dataframe. Dengan menggunakan dataframe, akan mempermudah tahapan-tahapan yang akan dilakukan dalam melakukan clustering.

Setelah melakukan ekspor data dilakukan pre-processing data. Dalam pre-processing data terdapat beberapa tahapan yang perlukan dilakukan, diantaranya adalah:

1. Penanganan cell yang tidak memiliki nilai
   Dalam melakukan penanganan terhadap cell yang tidak memiliki nilai dilakukan imputasi mean dan modus. Dengan dilakukan imputasi, cell yang tidak memiliki nilai dapat diisi menggunakan mean atau modus fitur dari cell tersebut. Imputasi mean digunakan untuk fitur yang bersifat kontinu, adapun fitur yang bersifat kontinu adalah:
   1. Umur
   2. Premi
   3. Lama Berlangganan

   Sedangkan imputasi modus digunakan untuk fitur yang bersifat kategorikal, adapun fitur yang bersifat kategorikal adalah:
   1. Jenis Kelamin
   2. SIM
   3. Kode Daerah
   4. Sudah Asuransi
   5. Umur Kendaraan
   6. Kendaraan Rusak
   7. Kanal Penjualan
   8. Tertarik
2. Transformasi fitur
   Transformasi fitur dilakukan untuk fitur yang bersifat kategorikal dan nilainya tidak berupa angka. Terdapat dua transformasi yang perlu dilakukan yaitu transformasi pada fitur ordinal dan non-ordinal. Adapun fitur ordinal yang perlu dilakukan transformasi adalah:
   1. Umur Kendaraan
      Dalam dataset ini sudah dilakukan binning terhadap fitur umur kendaraan, yaitu melakukan pengelompokan beberapa nilai menjadi kategori baru. Tahapan transformasi yang dilakukan adalah sebagai berikut:
      - kendaraan dengan umur kurang dari satu tahun ⟶ 3
      - kendaraan dengan umur satu sampai dengan dua tahun ⟶ 2
      - kendaraan dengan umur lebih dari dua tahun ⟶ 1

      Dengan adanya transformasi terhadap fitur umur kendaraan, nilai dari fitur ini tidak lagi dalam bentuk string, melainkan dalam bentuk integer.

   Selain fitur ordinal juga dilakukan transformasi terhadap fitur non-ordinal, Adapun fitur non-ordinal yang perlu dilakukan transformasi adalah:

1. Jenis Kelamin
   Pada dataset ini jenis kelamin digambarkan dalam bentuk string, yaitu wanita atau pria. Tahapan transformasi yang dilakukan adalah sebagai berikut:

Jenis Kelamin
Pria
Wanita

isPria	isWanita
1	0
0	1

2. Kendaraan Rusak
   Pada dataset ini kendaraan rusak digambarkan dalam bentuk string, yaitu pernah atau tidak. Tahapan transformasi yang dilakukan adalah sebagai berikut:

Kendaraan Rusak
Pernah
Tidak

isRusak	isTidakRusak
1	0
0	1

3. Penskalaan fitur
   Penskalaan fitur dilakukan agar setiap fitur berada pada range yang sama, Dalam tugas besar ini metode yang digunakan untuk melakukan penskalaan fitur adalah metode Min-Max Normalization. Adapun rumus dari normalisasi ini adalah:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

X_{scaled} : Hasil Normalisasi
X : Nilai Cell
X_{min} : Nilai Maksimal dari fitur
X_{max} : Nilai Minimal dari fitur

   Dengan melakukan penskalaan fitur, data-data yang ada dalam dataset akan terdistribusi dengan skala yang sesuai antara fitur satu dengan yang lainnya.

4. Feature Selection
   Feature Selection digunakan untuk mengurangi dimensi dari dataset yang ada, Pada tugas besar ini metode feature selection yang dipilih adalah menghapus fitur yang memiliki standar deviasi rendah. Standar deviasi merupakan ukuran untuk menghitung keragaman dalam kumpulan data. Maka dari itu fitur yang memiliki standar deviasi yang rendah cenderung tidak memberi pengaruh dalam perhitungan.
   Dapat dilihat jika fitur SIM dan Premi memiliki standar deviasi yang rendah, maka dari itu dilakukan drop fitur untuk keduanya.

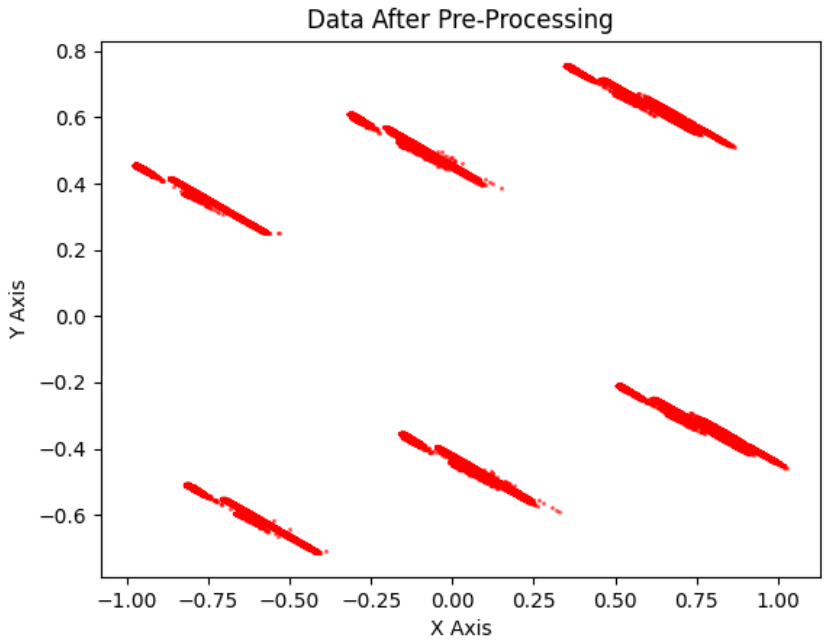
	id	Umur	SIM	Kode_Dae	Sudah_As	Premi	Kanal_Pen	Lama_Ber	Tertarik	isRusak	isTidakRus	isPria	isWanita
count	285831	285831	285831	285831	285831	285831	285831	285831	285831	285831	285831	285831	285831
mean	142916	0.289913	0.997957	0.509331	0.435939	0.051916	0.697664	0.499261	0.122471	0.529372	0.470628	0.563683	0.436317
std	82512.45	0.232794	0.045155	0.24849	0.49588	0.03109	0.330514	0.282425	0.32783	0.499137	0.499137	0.495929	0.495929
min	1	0	0	0	0	0	0	0	0	0	0	0	0
25%	71458.5	0.076923	1	0.288462	0	0.041484	0.333333	0.259516	0	0	0	0	0
50%	142916	0.276923	1	0.538462	0	0.052683	0.932099	0.499261	0	1	0	1	0
75%	214373.5	0.446154	1	0.673077	1	0.067372	0.932099	0.737024	0	1	1	1	1
max	285831	1	1	1	1	1	1	1	1	1	1	1	1

Gambar 2. Deskripsi Fitur

Selain itu karena fitur isTidakRusak dan isWanita merupakan invers dari isRusak dan isPria, maka dari itu kedua fitur ini juga didrop, karena fitur ini tidak memberikan keakurasian lebih baik.

5. Dimensionality Reduction
   Dimensionality Reduction digunakan untuk meringkas fitur-fitur yang sudah dipilih menjadi fitur yang lebih sederhana, pada kasus kali ini fitur yang awalnya berjumlah 8 (setelah SIM, Premi, isTidakRusak, dan isWanita dihapus) akan dirubah menjadi dua fitur. Metode yang digunakan adalah Principal Component analysis (PCA). Metode ini dipilih karena mudah diimplementasikan. Dengan menggunakan metode ini semua fitur yang berguna akan tetap berpengaruh, walaupun dimensi dari dataset sudah dikecilkan.

Setelah tahapan pre-processing data dilakukan didapatkan grafik persebaran data sebagai berikut:



Gambar 3. Grafik Persebaran Data

Setelah dilakukan pre-processing terhadap dataset, akan dilanjutkan dengan melakukan clustering terhadap dataset. Clustering dilakukan untuk membagi dataset menjadi beberapa kelompok. Metode yang digunakan dalam melakukan clustering adalah K-Means clustering, dimana dalam metode ini K merupakan jumlah cluster mula-mula yang akan dibangun. Dalam tugas besar ini

jumlah kluster yang akan dibangun adalah 2 buah, karena diharapkan kluster yang terbentuk adalah kluster orang yang tertarik untuk membeli kendaraan dan yang tidak tertarik untuk membeli kendaraan.

Mula-mulanya sejumlah K centroid akan dibangun secara random. Setelah itu setiap titik yang ada akan diassign ke centroid terdekat. Perhitungan jarak antara titik dan centroid dihitung menggunakan rumus euclidian. Adapun rumus euclidian adalah sebagai berikut:

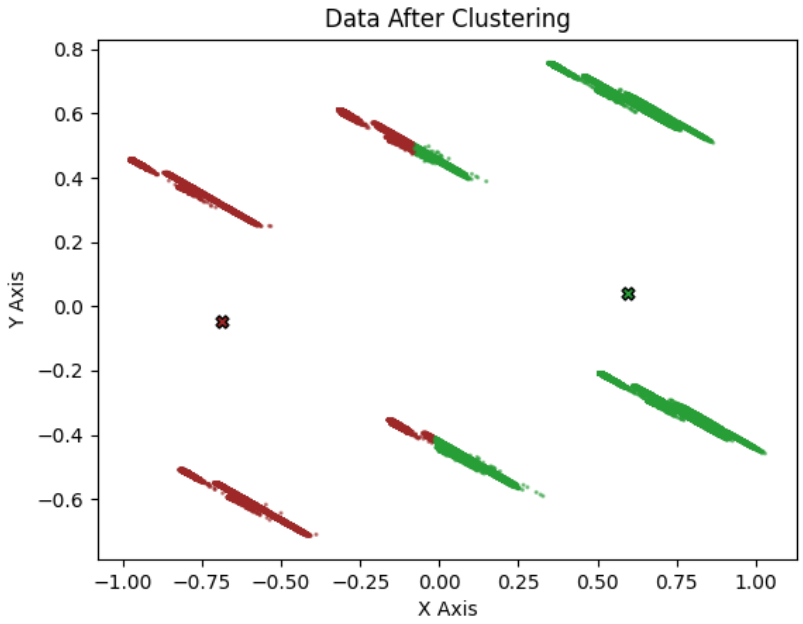
$$d_{ij} = \sqrt{\left[(x_i - x_j)^2 + (y_i - y_j)^2\right]}$$

x_i = koordinat x untuk fasilitas i

y_i = koordinat y untuk fasilitas i

d_{ij} = jarak antar fasilitas I dan j

Nantinya centroid ini akan diupdate menggunakan koordianat rata-rata dari kluster centroid tersebut. Proses update ini akan dilakukan secara berualang hingga posisi kluster awal sama dengan posisi kluster pada saat perulangan terakhir dilakukan. Berikut ini merupakan grafik persebaran data kluster iterasi terakhir.

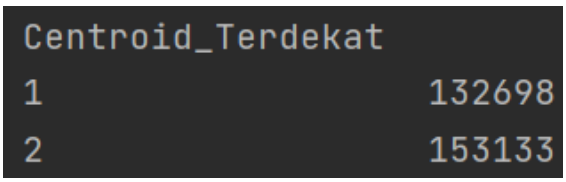


Gambar 4. Grafik Hasil Klasterisasi

BAB III Kesimpulan

Setelah melakukan klustering terhadap dataset kendaraan_train.csv didapatkan dua buah kluster dengan jumlah data pada kluster 1 sebanyak 132.698 dan data pada kluster 2 sebanyak 153.133.

Kedua kluster ini akan digunakan sebagai acuan dalam melakukan klasifikasi di tahap selanjutnya. Dengan menggunakan K-means pembangkitan kluster cenderung lebih cepat. Metode ini sangat cocok digunakan untuk pengelompokan data berukuran besar. Tetapi setelah menjalankan program beberapa kali dapat dilihat apabila K-Means tidak menghasilkan output yang stabil, hal ini terjadi karena K-Means sangat sensitif terhadap pembangkitan titik centroid awal secara random.



Centroid_Terdekat	
1	132698
2	153133

Gambar 5. Jumlah Titik di kluster 1 dan 2

LINK VIDEO: https://youtu.be/yNKnfxYdfr8