LAPORAN TUGAS BESAR MACHINE LEARNING
K MEANS CLUSTERING

Disusun Oleh :

Rafly Athalla(1301194216)

PROGRAM STUDI INFORMATIKA
FAKULTAS INFORMATIKA
UNIVERSITAS TELKOM
2021

PENDAHULUAN

1. Persoalan

Setiap mahasiswa harus mengerjakan dua task (clustering dan classification) terhadap dataset pada link gdrive:

https://drive.google.com/drive/folders/14QPe3o6LeSjfYj-kGhCZJM4pn-I55YsJ?usp=sharing

Tugas clustering (unsupervised Learning) adalah mengelompokkan pelanggan berdasarkan data pelanggan di dealer tanpa memperhatikan label kelas apakah pelanggan tertarik untuk membeli kendaraan baru atau tidak.

2. Rumusan Masalah

Adapun Rumusan Masalah dari tugas besar ini sebagai berikut :

- Formulasi Masalah: jelaskan permasalahan yang akan diselesaikan.
- Eksplorasi dan Persiapan Data : lakukan semua teknik eksplorasi dan persiapan data yang menurut Anda perlu dilakukan.
- Pemodelan: bangunlah model menggunakan data hasil preprossecing dan lakukan proses training untuk mendapatkan hasil terbaik.
- Evaluasi: pilih metode evaluasi yang sesuai beserta justifikasinya.
- Eksperimen: lakukan berbagai eksperimen yang melibatkan tahapan Eksplorasi dan Persiapan Data, Pemodelan, dan Evaluasi untuk mendapatkan hasil terbaik.
- Kesimpulan: berikan kesimpulan dari semua proses yang dijalankan berserta hasil akhir dari berbagai eksperimen yang telah dilakukan.

PEMBAHASAN

1. Formulasi Masalah

Clustering adalah metode pengelompokan data. Clustering merupakan proses partisi satu setobjek data ke dalam himpunan bagian yang disebut dengan cluster. Objek yang di dalam cluster memiliki kemiripan karakteristik antar satu sama lainnya dan berbeda dengan cluster yang lain. Oleh karena itu, clustering sangat berguna dan bisa menemukan kelompok yang tidak dikenal dalam data

Task yang dilakukan adalah mengelompokan data pelanggan tertarik untuk membeli kendaraan baru atau tidak berdasarkan data pelanggan di dealer.K-Means dipilih sebagai metode dalam task ini. Tujuan dari clustering adalah meminimumkan jarak antara data point dan centroid , serta memaksimumkan jarak antara centroid yang dihitung menggunakan within-cluster sum of squares atau WCSS

2. Eksplorasi dan Persiapan Data

a. Data Exploration/Understanding

Pada tahap exploration/understanding saya menjawab dan menyiapkan hal hal berikut :

- Count the number of record : 285831, menurut saya ini cukup dalam mengelompokkan orang orang yang berdasarkan karakteristiknya.
- Pada setiap row data terdapat data isnull kecuali pada kolom id dan tertarik.
- Tidak ada data yang terduplikat
- Type data :
    - Id memiliki tipe data int64
    - Jenis_Kelamin memiliki tipe data categorical
    - Umur memiliki tipe data float64
    - Sim memiliki tipe data float64
    - Kode_Daerah memiliki tipe data float64
    - Sudah_Asuransi memiliki tipe data float64
    - Umur_Kendaraan memiliki tipe data float64
    - Kendaraan_Rusak memiliki tipe data categorical
    - Premi memiliki tipe data categorical
    - Kanal_Penjualan memiliki tipe data float64
    - Lama_Berlangganan memiliki tipe data float64
    - Tertarik memiliki tipe data int64
- Look for missing values Terdapat missing value pada setiap kolom , kecuali kolom tertarik.

- Validated if your data balance
    - Data tidak balance, karena lebih banyak data yang tidak tertarik dibandingkan data yang tertarik.
    - Yang tidak tertarik ada sebanyak 250825
    - Yang tertarik ada sebanyak 35006

	Jenis_Kelamin	Umur	SIM	Kode_Daerah	Sudah_Asuransi	Umur_Kendaraan	Kendaraan_Rusak	Premi	Kanal_Penjualan	Lama_Berlangganan	Tertarik
0	Wanita	49	1	8	0	1-2 Tahun	Pernah	46963	26	145	0
1	Pria	22	1	47	1	< 1 Tahun	Tidak	39624	152	241	0
2	Pria	24	1	28	1	< 1 Tahun	Tidak	110479	152	62	0
3	Pria	46	1	8	1	1-2 Tahun	Tidak	36266	124	34	0
4	Pria	35	1	23	0	1-2 Tahun	Pernah	26963	152	229	0
...	...	...	...	...	...	...	...	...	...	...	...
47634	Pria	61	1	46	0	> 2 Tahun	Pernah	31039	124	67	0
47635	Pria	41	1	15	0	1-2 Tahun	Pernah	2630	157	232	0
47636	Pria	24	1	29	1	< 1 Tahun	Tidak	33101	152	211	0
47637	Pria	59	1	30	0	1-2 Tahun	Pernah	37788	26	239	1
47638	Pria	52	1	31	0	1-2 Tahun	Tidak	2630	124	170	0

47639 rows × 11 columns

```
df_Kentrain = pd.read_csv("kendaraan_train.csv")
df_Kentrain
```

✓ 0.6s

	id	Jenis_Kelamin	Umur	SIM	Kode_Daerah	Sudah_Asuransi	Umur_Kendaraan	Kendaraan_Rusak	Premi	Kanal_Penjualan	Lama_Berlangganan	Tertarik
0	1	Wanita	30.0	1.0	33.0	1.0	< 1 Tahun	Tidak	28029.0	152.0	97.0	0
1	2	Pria	48.0	1.0	39.0	0.0	> 2 Tahun	Pernah	25800.0	29.0	158.0	0
2	3	NaN	21.0	1.0	46.0	1.0	< 1 Tahun	Tidak	32733.0	160.0	119.0	0
3	4	Wanita	58.0	1.0	48.0	0.0	1-2 Tahun	Tidak	2630.0	124.0	63.0	0
4	5	Pria	50.0	1.0	35.0	0.0	> 2 Tahun	NaN	34857.0	88.0	194.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
285826	285827	Wanita	23.0	1.0	4.0	1.0	< 1 Tahun	Tidak	25988.0	152.0	217.0	0
285827	285828	Wanita	21.0	1.0	46.0	1.0	< 1 Tahun	Tidak	44686.0	152.0	50.0	0
285828	285829	Wanita	23.0	1.0	50.0	1.0	< 1 Tahun	Tidak	49751.0	152.0	226.0	0
285829	285830	Pria	68.0	1.0	7.0	1.0	1-2 Tahun	Tidak	30503.0	124.0	270.0	0
285830	285831	Pria	45.0	1.0	28.0	0.0	1-2 Tahun	Pernah	36480.0	26.0	44.0	0

285831 rows × 12 columns

Preprocessing

b. Feature Engineering

Pada tahap Feature Engineering melakukan Scaling method, yaitu menyamakan skala data dari yang kita punya dengan metode min-max normalization atau, skalanya yaitu 0 – 1, sehingga saat mencari jarak antara 2 titik, tidak ada sumbu yang mendominasi.

3. Pemodelan

Dalam pembuatan program ini, saya menggunakan metode K Means. K means merupakan salah satu algoritma clustering. Tujuan algoritma ini yaitu untuk membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan supervised learning yang menerima masukan berupa vektor (x1, y1), (x, y2),

…, (xi, yi), di mana xi merupakan data dari suatu data pelatihan dan yi merupakan label kelas untuk xi.

Kelebihan dari K-means adalah

- Mudah dilakukan saat pengimpelementasian dan di jalankan.
- Waktu yang di butuhkan untuk melakukan pembelajaran relatif lebih cepat.
- Sangat fleksibel, adaptasi yang mudah untuk di lakukan
- Sangat umum penggunaannya.
- Menggunakan prinsip yang sederhana dapat di jelaskan dalam non-statistik.

Kekurangan dari K-means adalah:

- Sebelum algoritma di jalankan, titik K diinisialisasikan secara random sehingga pengelompokan data yang di dapatkan bisa berbeda-beda. Namun apabila nilai yang diperoleh acak untuk penginisialisasi kurang baik maka pengelompokan yang didapatkn menjadi tidak optimal.
- Apabila terjebak dalam kasus yang biasanya di sebut dengan curse of dimensionality. Hal ini pun akan terjadi apabila salah satu data untuk melakukan pelatihan mempunyai dimensi yang sangat banyak, sebagai contoh; jika ada data pelatihan yang terdiri dari 2 buah atribut saja maka dimensinya ada 2 dimensi pula, namun akan berbeda jika ada 20 atribut maka akan ada 20 dimensi yang di miliki. Adapun salah satu dari cara kerja algoritma cluster ini ialah untuk mencari jarak terdekat dari antara k titik dangan titik lainnya. Apabila ingin mencari jarak untuk antar titik dari 2 dimensi hal itu masih mudah untuk di lakukan, namun bagaimana dengan 20 buah dimensi hal tersebut akan menjadi lebih sulit untuk di lakukan pencarian jarak.
- Apabila hanya ada terdapat beberapa buah titik sampel data yang ada, maka hal yang mudah untuk melakukan penghitungan dan mencari jarak titik terdekat dengan k titik yang telah di lakukan inisialisasi yang secara acak. Namun jika ada banyak titik data, misalkan satu juta data, maka perhitungan dan pencarian titik terdekat akan sangat membutuhkan waktu yang lama. Proses tersebut dapat dipercepat namun dibutuhkan sebuah struktur data yang lebih rumit seperti hashing untuk melakukan proses tersebut.
- Adanya penggunaan k buah random, tidak ada jaminan untuk menemukan kumpulan cluster yang optimal.

Aplikasi K-Means Clustering sangat sering digunakan, mulai dari unsupervised learning of neural network, Pattern Recognitions, Classifications Analysis, Artificial Intelligence, Image Processing, Computer Vision dan banyak lainnya.

Langkah- Langkah dalam melakukan K-Means Method adalah :

1. Memilih objek k secara acak, setelah mendapatkan objek k tersebut data akan diproses sebagai mean pada cluster
2. Setiap objek akan dimasukkan kedalam cluster yang mempunyai kemiripan terhadap cluster. Tingkat kemiripan dapat ditentukan dengan mencari jarak objek terhadap mean atau centroid cluster tersebut
3. Lakukan perhitungan nilai centroid yang baru pada setiap cluster
4. Proses perhitungan nilai centroid tersebut dilakukan berluang-ulang hingga didapati anggota pada kelompok cluster tersebut tidak berubah

Adapun algoritma k means yang digunakan sebagai berikut :

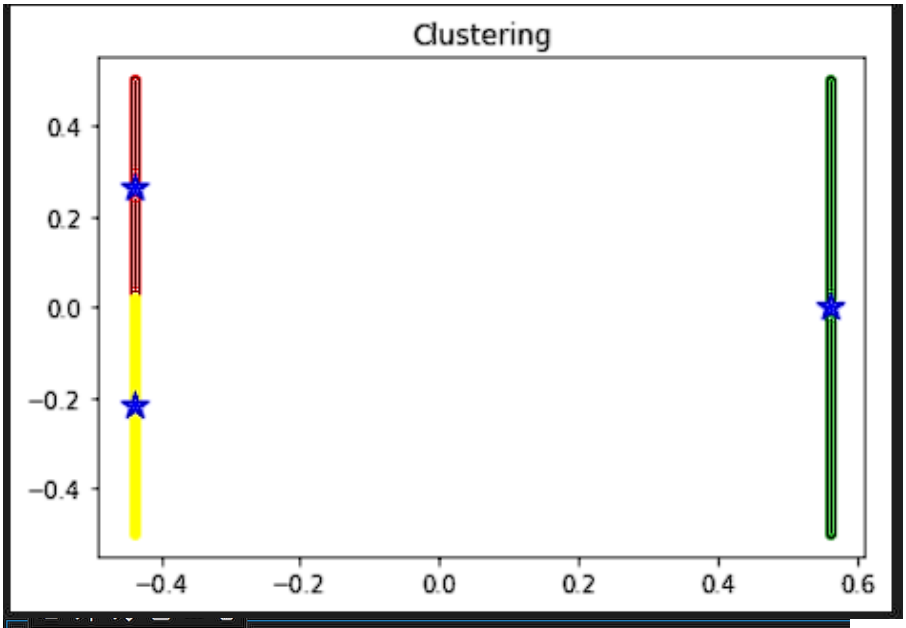
```python
def euclidean(x1,x2) :
  return np.sqrt(np.sum((x1-x2)**2))

class KMEANS :
    def __init__(self, kmeans_cluster, kmeans = 2, maksimal_iterasi = 200):
      self.kmeans = kmeans
      self.maksimal_iterasi = maksimal_iterasi
      self.Centroid = []
      self.cluster = [[] for i in range(self.kmeans)]

    def memilih_centroid_terdekat(self, row) :
      TotalJarakCentroid = [euclidean(row,centroid) for centroid in self.Centroid]
      return np.argmin(TotalJarakCentroid)

    def membuat_centroid_baru(self) :
      Centroid = np.zeros((self.kmeans, self.column))
      for indexCluster, cluster in enumerate(self.cluster) :
        centroidBaru = np.mean(self.A[cluster], axis=0)
        Centroid[indexCluster] = centroidBaru
      return Centroid

    def membuat_label(self) :
      label = np.empty(self.baris)
      for indexCluster, cluster in enumerate(self.cluster):
        for row in cluster:
          label[row] = indexCluster
      return label

    def prediksi(self, A) :
      self.A = A
      self.baris , self.column = A.shape

      # membuat centroid secara random
      index_centroid = np.random.choice(self.baris, self.kmeans, replace=False)
      for index in index_centroid :
        self.Centroid.append(self.A[index])

      # melakukan perulangan sebanyak iterasi atau centroid tidak berubah
      for i in range(self.maksimal_iterasi) :

        # mencari serta menentukan cluster atau centroid
        cluster = [[] for i in range(self.kmeans)]
        for idx, row in enumerate(self.A) :
          # mencari centroid terdekat dengan membandingkan jarak setiap centroid
          IndexCentroidTerdekat = self.memilih_centroid_terdekat(row)
          cluster[IndexCentroidTerdekat].append(idx)

        self.cluster = cluster

        # update centroid
        centroid_Lama = self.Centroid
        self.Centroid = self.membuat_centroid_baru()

        # cek centroid apakah centroid tersebut berubah atau tidak
        berubah = False
        for i,CentroidLama in enumerate(centroid_Lama) :
          # menghitung jarak dari centroid lama ke centroid baru
          jarak = euclidean(centroid_Lama, self.Centroid[i])
          if (jarak != 0) :
            berubah = True
          if (berubah == False) :
            break

        # menentukan labeling
        label = self.membuat_label()
        return label
```

4. Evaluasi

Melakukan evaluasi dengan menggunakan Elbow Method dengan menggunakan nilai distortion. Distortion adalah rata-rata jarak kuadrat dari titik tengah cluster dari masing-masing cluster. Dihitung untuk nilai K dari 2 sampai 10 agar dapat membuat grafik kelandaian antara nilai distortion terhadap nilai K. Nantinya akan ditentukan nilai K terbaik berdasarkan grafik tersebut yang dimana ditentukan dengan cara memilih nilai K yang merupakan titik dimana nilai distortion/inertia dari data tersebut mulai menurun secara linear

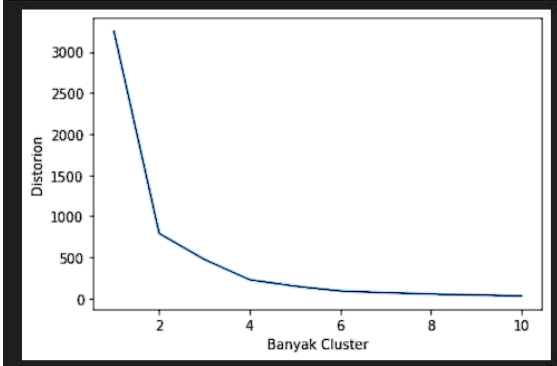
5. Eksperimen

a. Umur dan kendaraan yang rusak




```
#umur dan kendaraan_rusak
dataEksperimen = normalisasi.iloc[:, [1,10]].values
dataEksperimen = dataEksperimen[:10000]

from sklearn.cluster import KMeans
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 50)
    kmeans.fit(dataEksperimen)
    wcss.append(kmeans.inertia_)
plt.plot(range(1, 11), wcss)
plt.xlabel('Banyak Cluster')
plt.ylabel('Distorion')
plt.show()
```

✓ 1.6s



PENUTUP

1.Kesimpulan

Dari proses reading data, exploratory data analysis, dan clustering yang telah dijalani, dapat disimpulkan bahwa:

- Berdasarkan elbow method, untuk clustering ini K = 3 adalah jumlah k cluster yang efektif
- Untuk visualisasi, hasil clustering harus direduksi dengan PCA terlebih dahulu agar terlihat jelas sebaran clusternya
- Apabila dilihat dari elbow method, makin banyak K yang digunakan saat KMeans maka total minimum jarak makin sedikit karena centroid yang menyebar dapat memperkecil jarak antar data dan centroid
- Dataset yang diberikan kurang baik untuk di clustering.

Link Youtube : https://youtu.be/kBNg3gFlRDg

DAFTAR PUSTAKA

[1] V. Handayani, A. dan A. P. kurniati, "Analisa Clustering Menggunakan Algoritma K-Modes," Telkom University, pp. 1-8, 2010.

[2] J. O. Ong, "Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University," Jurnal Ilmiah Teknik Industri, vol. 12, no. 1, pp. 10 - 20, 2013

[3] Purnamasari, S.B. 2014. Pemilihan Cluster Optimum Pada Fuzzy C-Means (Studi kasus: Pengelompokan Kabupaten/Kota di Jawa Tengah berdasarkan Indikator Indeks Pembangunan Manusia). Jurnal Gaussian. Vol.3 No.3. Semarang: Universitas Diponegoro