# 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**

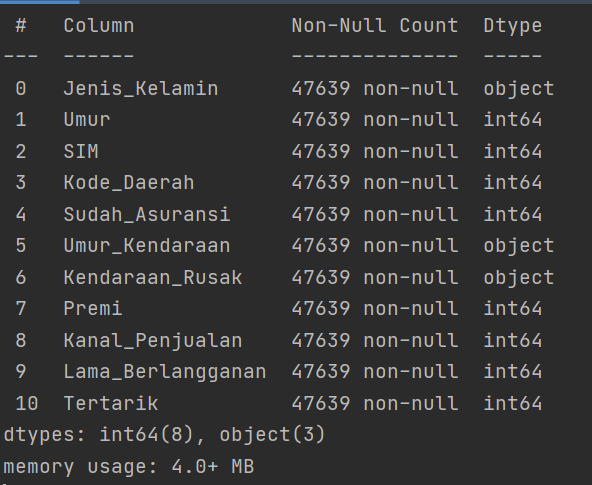
**UNIVERSITAS TELKOM**

**BANDUNG**

**2021**

# BAB I Pendahuluan

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\_test.csv . Dataset kendaraan\_test.csv ini memiliki 47639 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

Gambar 1. Gambaran umum dataset kendaraan\_test.csv

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 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.

# 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, nilai dari cell yang tidak memiliki nilai dapat diperkirakan menggunakan mean atau modus dari fitur 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
9. 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 dalam bentuk string lagi, 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 |

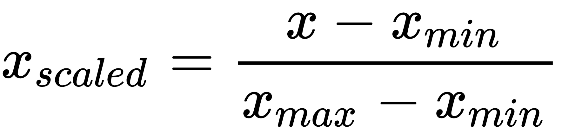
1. Kendaraan Rusak

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 |

1. Penskalaan fitur

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

Xscaled : Hasil Normalisasi

X : Nilai Cell

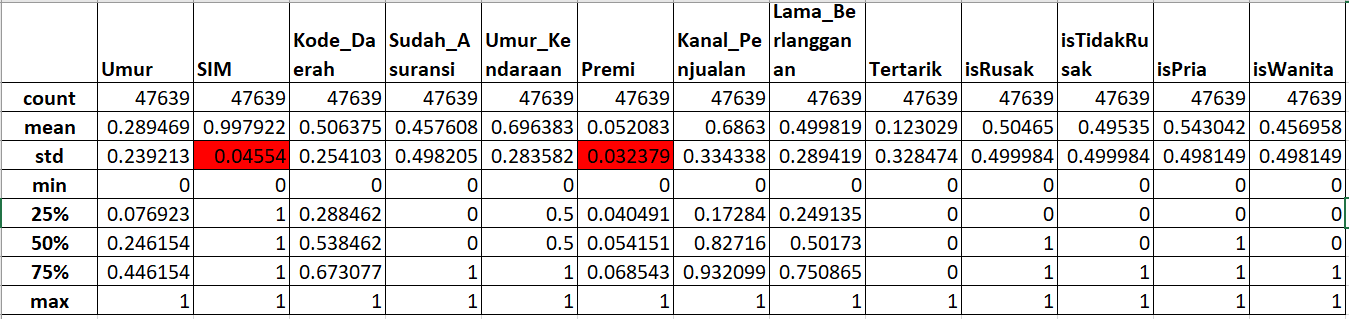
Xmin : Nilai Maksimal dari fitur

Xmax : 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.

1. 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 yang rendah. Standar deviasi merupakan ukuran dalam memperhitungkan keragaman dalam kumpulan data. Maka dari itu fitur yang memiliki standar deviasi yang rendah cenderung tidak memberi pengaruh dalam perhitungan.

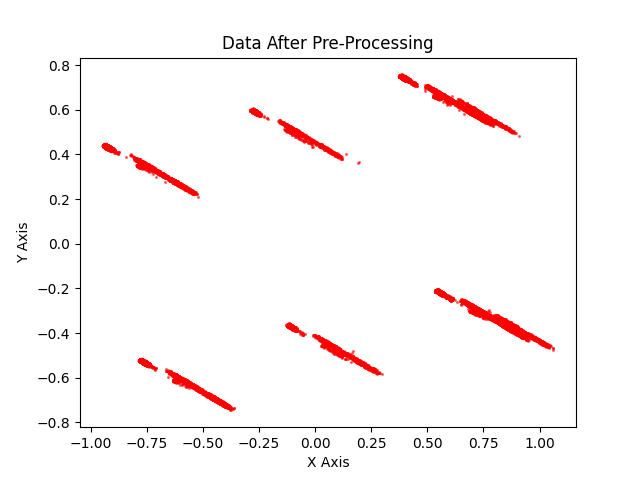
Dapat dilihat jika fitur SIM dan Premi memiliki standar deviasi yang rendah, maka dari itu dilakukan drop fitur untuk keduanya.

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.

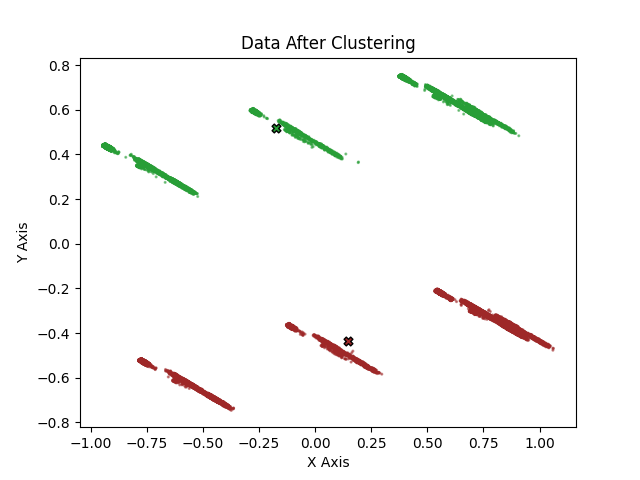
1. Dimensionality Reduction

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 juga centroid ini dihitung menggunakan rumus euclidian. 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