Laporan Tugas Besar MK Pembelajaran Mesin

CSH3L3 Pembelajaran Mesin Genap 2019/2020

Ariq Musyaffa Ramadhani 1301174354 IF-41-02

Formulasi Masalah

Diberikan dataset used_car yang kemudian akan dilakukan clustering dan classification. Clustering akan dilakukan dengan menggunakan K-means, sedangkan classification akan menggunakan K-Nearest Neighbors dan Naïve Bayes yang sebelumnya dataset akan dilakukan proses preparation.

Clustering

Data Preparation/Pre-Processing

1. Memeriksa jumlah data setiap fitur
   Dilakukan pemeriksaan terhadap fitur-fitur pada dataset. Pada dataset used_car terdapat fitur yang memiliki jumlah baris 0, yaitu fitur ‘county’ maka fitur tersebut dihapus.

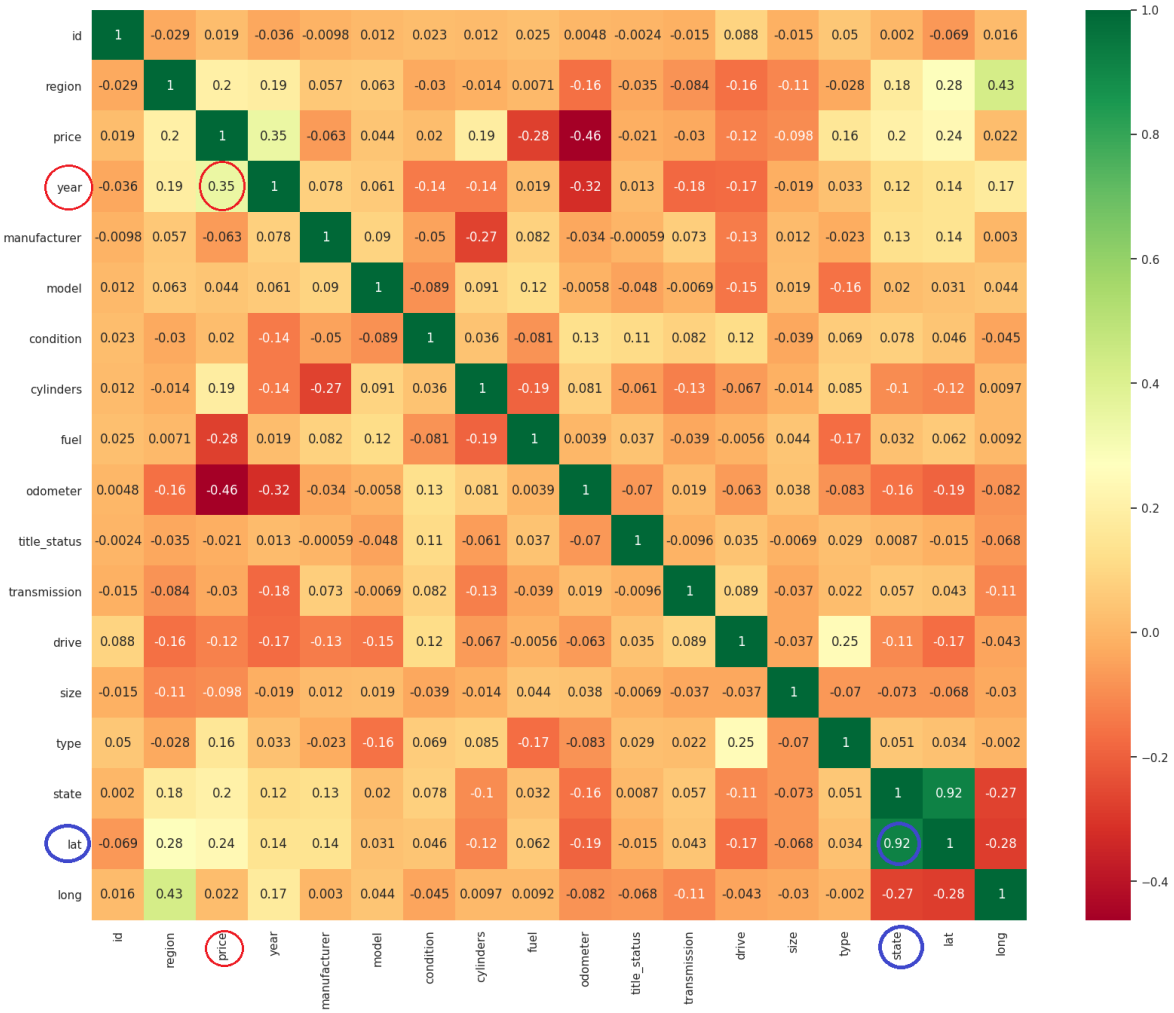
```
[347] data = pd.read_csv("drive/My Drive/malin/used_cars.csv")
      data
      data.describe()
```

	id	price	year	odometer	county	
	)0e+04	2.000100e+04	19989.000000	1.761200e+04	0.0	1
	)9e+09	7.664058e+04	2009.830657	9.916435e+04	NaN	
	20e+06	8.335762e+06	7.913613	7.963487e+04	NaN	
	)7e+09	0.000000e+00	1917.000000	0.000000e+00	NaN	
	14e+09	3.970000e+03	2007.000000	5.013300e+04	NaN	

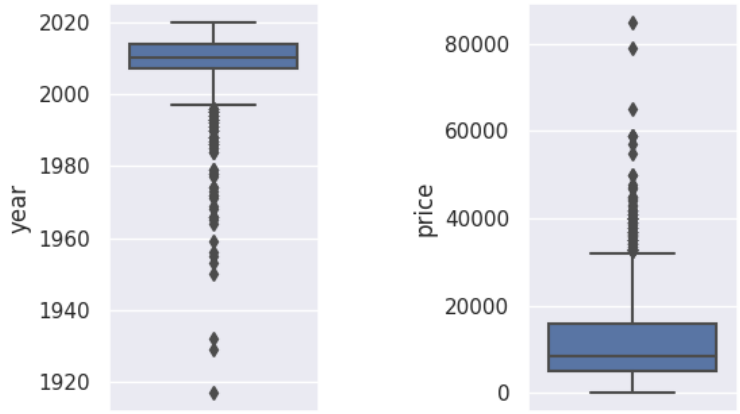
2. Missing value handling
   Setelah itu dilakukan penanganan missing value dengan cara menghapus baris yang memiliki nilai null/NaN. Metode ini dipilih karena jumlah data yang mencapai 20.000 sehingga akan lebih efektif jika baris yang memiliki nilai null dihapus. Setelah proses ini jumlah baris data berkurang menjadi 2814.

3. Memilih fitur-fitur kandidat
   Memilih fitur-fitur yang akan diproses selanjutnya. Fitur-fitur yang akan dihapus adalah url, region_url, vin, paint_color, image_url, description. Fitur tersebut dihapus karena dirasa tidak terlalu penting dalam proses clustering dan classification.
4. Data Encoding
   Pada dataset used_car terdapat beberapa fitur yang bertipe string, dilakukan encoding untuk merubah data menjadi bertipe numerik.

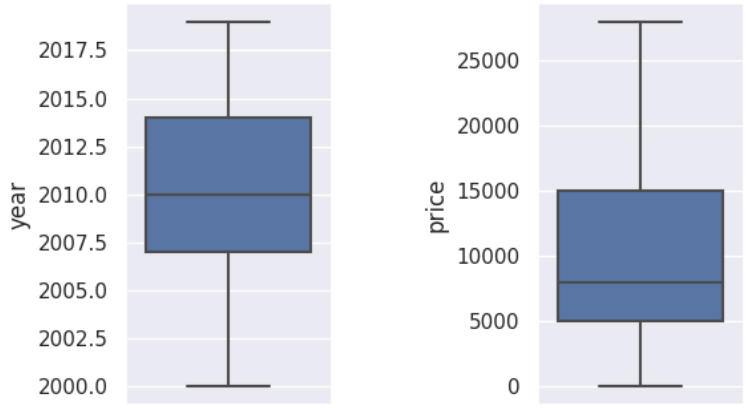
5. Pemeriksaan Korelasi antar fitur dan pemilihan fitur
   Digunakan HeatMap yang terdapat pada library seaborn pada python untuk memetakan nilai korelasi antar fitur.



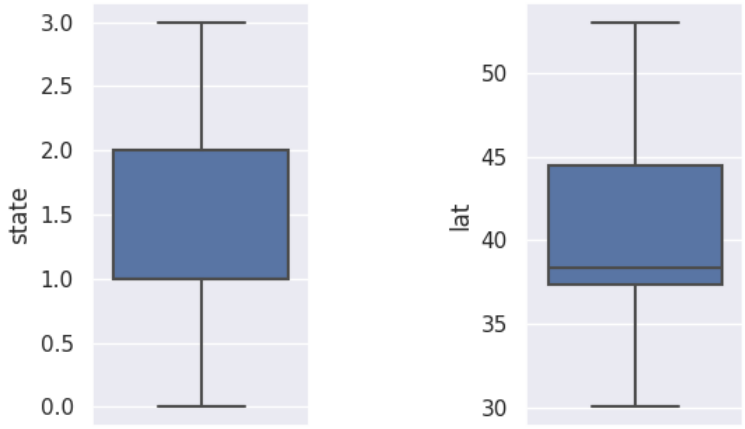
   Fitur yang digunakan pada skenario 1 adalah year dan price dengan nilai 0.35 sedangkan pada skenario 2 adalah state dan lat dengan nilai 0.92.
6. Outlier
   Pada skenario 1 terdapat outlier pada kedua fitur.



Maka selanjutnya akan dilakukan penanganan dengan cara menghapus data-data yang berada pada titik outlier.



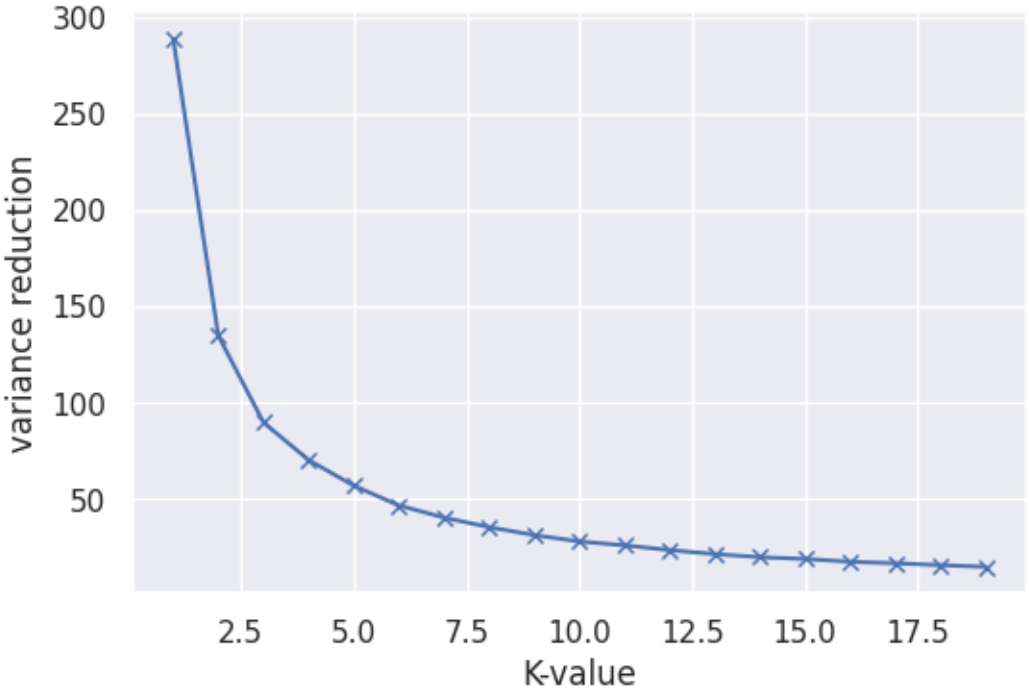
Sedangkan pada skenario 2 tidak terdapat outlier sehingga tidak dilakukan handling.



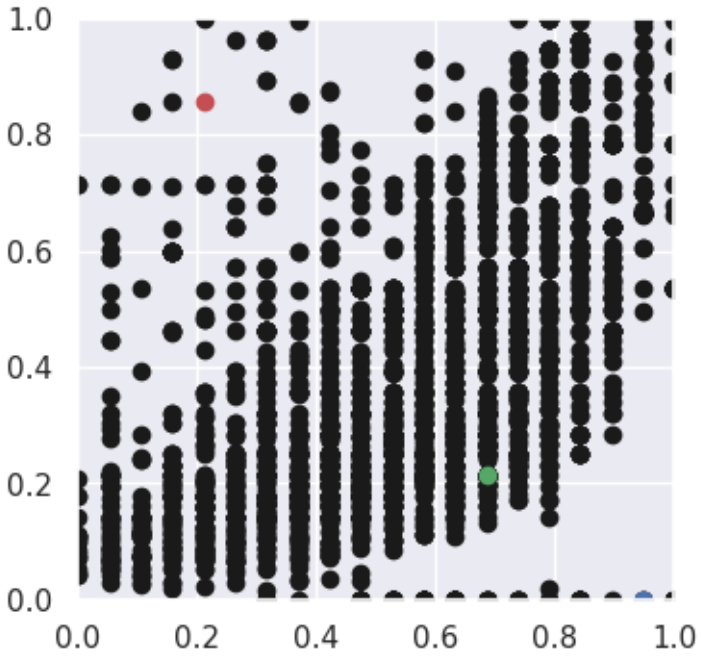
7. Scalling
   Range data pada kedua fitur terbilang cukup jauh, maka dari itu dilakukan scalling agar kedua fitur memiliki range data yang sama, yaitu antara 0 sampai 1. Metode yang dipilih adalah Min Max Scalling.

Clustering Skenario 1

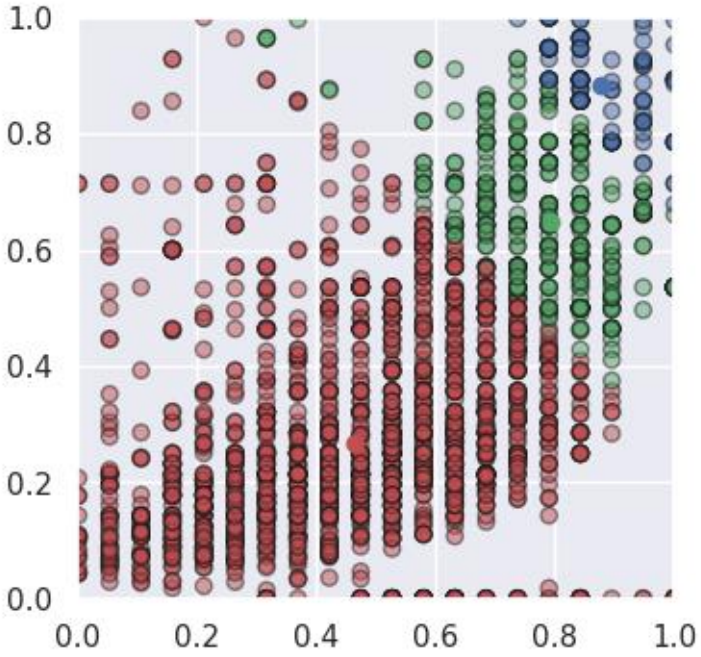
Clustering akan menggunakan metode K-Means. K-Means digunakan karena cukup mudah digunakan dan dapat menangani dataset yang cukup besar. Nilai K sangat berpengaruh pada metode ini sehingga diperlukan elbow method untuk membantu penentuan nilai K.



Pada grafik diatas didapat bahwa pengurangan variansi yang signifikan terjadi pada saat K=3, maka nilai K adalah 3. Setelah itu menentukan centroid awal. Penentuan centroid awal adalah dengan angka random. Angka yang dirandom adalah angka-angka yang terdapat pada masing-masing data pada fitur.

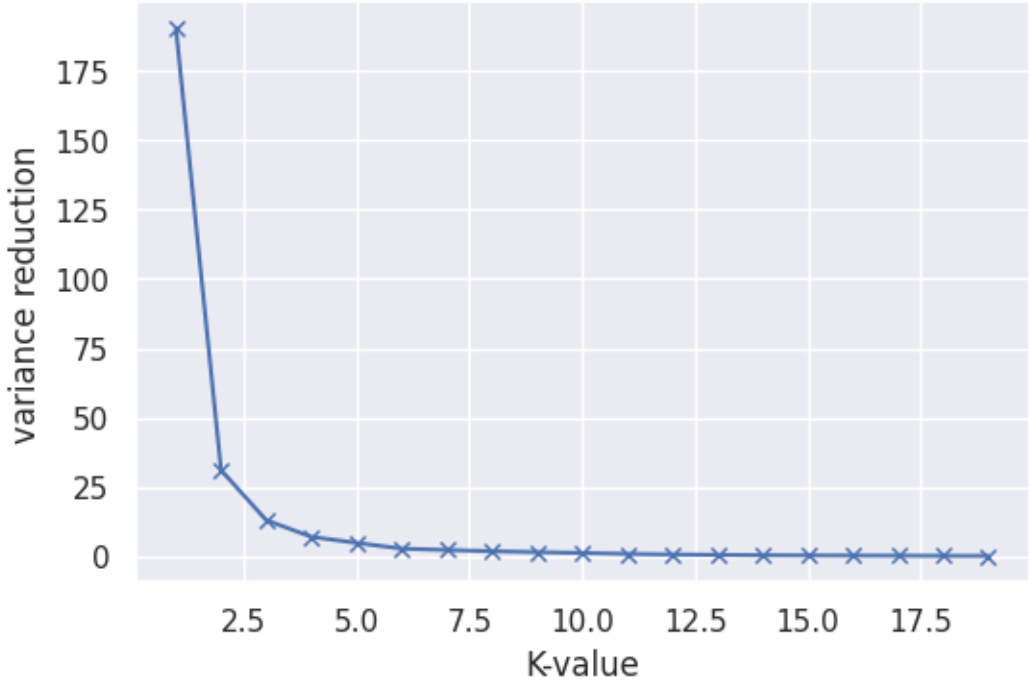


Langkah selanjutnya adalah menghitung jarak setiap data ke masing-masing centroid dengan Euclidean distance kemudian memperbarui posisi centroid. Langkah tersebut terus dilakukan hingga posisi centroid tidak berubah dari posisi awal.

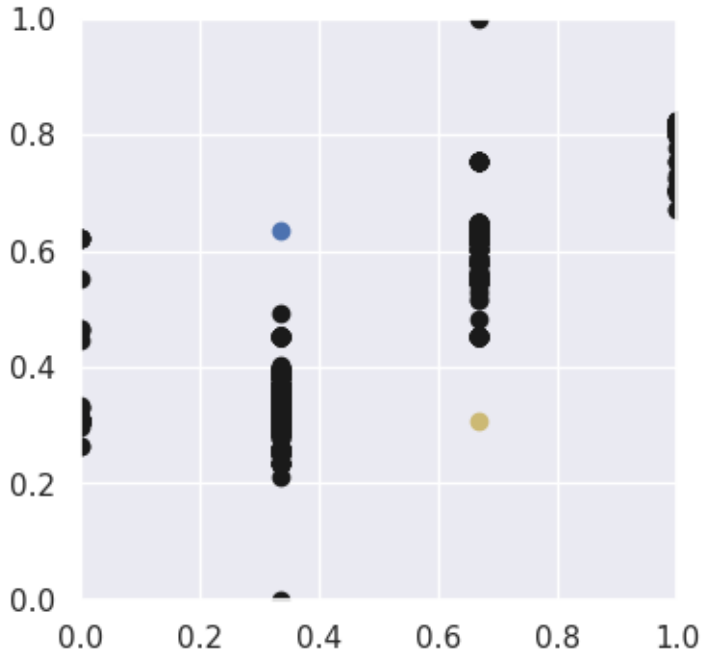


Clustering Skenario 2

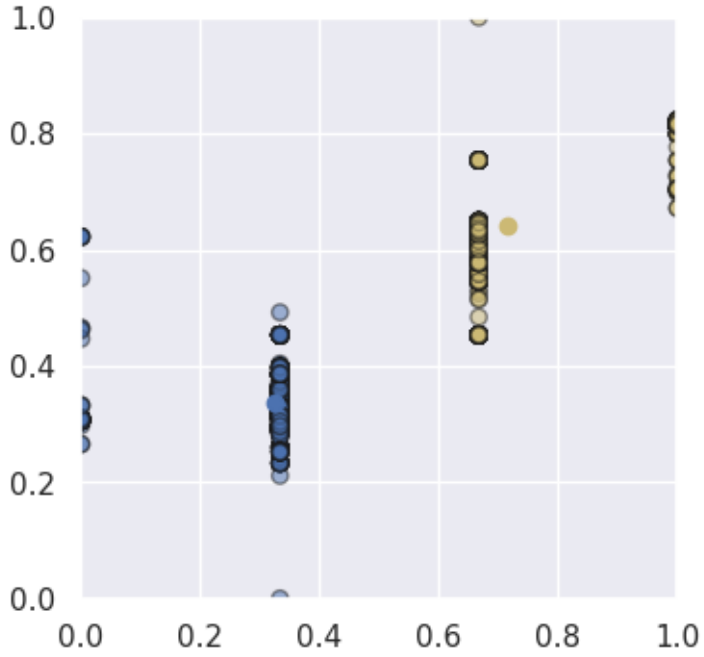
Clustering akan menggunakan metode K-Means. K-Means digunakan karena cukup mudah digunakan dan dapat menangani dataset yang cukup besar. Nilai K sangat berpengaruh pada metode ini sehingga diperlukan elbow method untuk membantu penentuan nilai K.



Pada grafik diatas didapat bahwa pengurangan variansi yang signifikan terjadi pada saat K=2, maka nilai K adalah 2. Setelah itu menentukan centroid awal. Penentuan centroid awal adalah dengan angka random. Angka yang dirandom adalah angka-angka yang terdapat pada masing-masing data pada fitur.



Langkah selanjutnya adalah menghitung jarak setiap data ke masing-masing centroid dengan Euclidean distance kemudian memperbarui posisi centroid. Langkah tersebut terus dilakukan hingga posisi centroid tidak berubah dari posisi awal.



Classification

Pre-Processing

Tahap-tahap yang dilakukan hampir sama dengan pre-processing pada clustering. Namun berbeda pada saat pemilihan fitur-fitur yang digunakan. Langkah pertama adalah menentukan fitur yang dijadikan kelas(variable dependen). Pada skenario 1 dan skenario 2 akan menggunakan data yang sama, yang berbeda adalah pada saat proses klasifikasi dengan algoritma yang berbeda. Fitur yang dijadikan kelas adalah condition. Kemudian menetukan kandidat variabel independen dengan bantuan fungsi correlation yang sudah tersedia pada python .

```
year            0.139979
condition       1.000000
odometer        0.127944
title_status    0.113083
drive           0.115289
Name: condition, dtype: float64
```

Kemudian setiap kandidat variabel independen tidak boleh saling bergantung satu sama lain sehingga harus diperiksa masing-masing korelasi antar kandidat variabel independen.

	year	odometer	title_status	drive
year	1.000000	-0.322274	0.012871	-0.172332
odometer	-0.322274	1.000000	-0.069712	-0.063406
title_status	0.012871	-0.069712	1.000000	0.035379
drive	-0.172332	-0.063406	0.035379	1.000000

Berdasakan data diatas fitur odometer dihapus. Selanjutnya adalah memisahkan dataset menjadi data training dan data test.

Classification Skenario 1

Pada skenario 1 digunakan algoritma K-Nearest Neighboors(KNN). KNN adalah algoritma klasifikasi dengan cara mengklasifikasikan sebuah instance berdasarkan mayoritas karakteristik dari k-tetangga terdekat. (Sumber : https://medium.com/bee-solution-partners/cara-kerja-algoritma-k-nearest-neighbor-k-nn-389297de543e)

Nilai K yang akan digunakan adalah 5. Semua fungsi yang akan digunakan sudah tersedia pada library sklearn.neighbors. Berikut adalah hasil evaluasi dari algortima KNN.

```
Accuracy  :  0.5754884547069272
F1-Score  :  0.28097180647055353
Precision :  0.3343844657078093
Recall    :  0.2641197939842648
```

Classification Skenario 2

Pada skenario 2 digunakan algoritma Naïve Bayes. Naïve Bayes merupakan algoritma yang berdasarkan probabilitas dan statistika untuk memprediksi event yang akan datang. Naïve Bayes dapat dirumuskan sebagai berikut :

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

Dimana P(C|X) adalah probabilitas hipotesis berdasarkan kondisi, P(X|C) adalah probabilitas berdasarkan kondisi pada hipotesis, P(C) probabilitas hipotesis, dan P(X) adalah probabilitas dari kelas. (Sumber : https://informatikalogi.com/algoritma-naive-bayes/)

Semua fungsi yang akan digunakan sudah tersedia pada library sklearn.naive_bayes. Berikut adalah hasil evaluasi dari algoritma Naïve Bayes.

```
Accuracy  :  0.42451154529307283
F1-Score  :  0.2256773255496077
Precision :  0.2549905806645216
Recall    :  0.32234101474110755
```

Evaluasi

Akan digunakan beberapa perhitungan untuk mengevaluasi hasil klasifikasi, yaitu :

1. Akurasi : rasio dari data yang diklasifikasi dengan benar dibandingkan dengan seluruh data.
2. Recall : rasio dari data yang diklasifikasi positif dengan benar dibandingkan dengan seluruh data positif.
3. F1-Score : Menggunakan nilai Recall dan Precission sebagai acuan.
4. Precission : rasio dari data yang diklasifikasi positif dengan benar dibandingkan dengan seluruh data prediksi positif.

Kesimpulan

Pada percobaan yang telah dilakukan dapat disimpulkan bahwa :

- Pada skenario 1 diperlukan proses outlier handling sedangkan pada skenario 2 tidak perlu yang menyebabkan jumlah data pada skenario 2 lebih banyak.
- Tahap pre-processing dan pemilihan fitur yang digunakan sangat berpengaruh pada model yang dihasilkan pada saat percobaan clustering.
- Pemilihan algoritma sangat berpengaruh terhadap akurasi, pada percobaan kali ini akurasi tertinggi didapat pada saat menggunakan algoritma KNN.