Laporan Tugas Besar Tahap 1
Pembelajaran Mesin
Task Clustering

1301184103

FIRLISA ANGGRAENI

IF-42-11

Program Studi Sarjana Informatika

Fakultas Informatika

Universitas Telkom

Bandung

2021

1. Formulasi Masalah

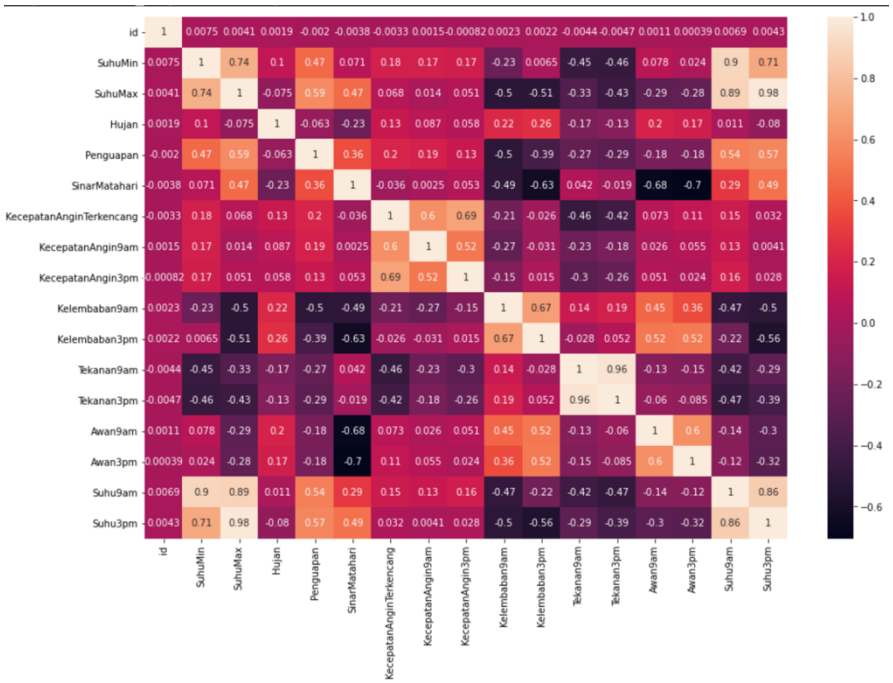
Clustering adalah salah satu metode penganalisaan data yang digunakan untuk mengelompokkan data berdasarkan karakteristik, apabila data memiliki karakteristik yang sama maka akan berada di satu kelompok yang sama dan apabila data memiliki karakteristik yang berbeda maka akan dimasukkan ke kelompok yang berbeda. Terdapat banyak metode yang dapat digunakan untuk clustering, salah satunya adalah K-Means. K-Means akan mengelompokkan data kepada centroid yang paling terdekat dari data tersebut.

Pada tugas ini, dataset yang digunakan adalah dataset salju yang memiliki 23 fitur. Fitur yang dipilih adalah SuhuMax dan Kelembaban9am. Berdasarkan nilai korelasi yang didapatkan, SuhuMax dan Kelembaban9am saling berbanding terbalik. Sehingga permasalahan yang akan diselesaikan pada tugas ini adalah pengelompokkan data SuhuMax terhadap Kelembaban9am yang memiliki karakteristik yang sama dengan jumlah cluster sebanyak k.

2. Eksplorasi dan Persiapan Data

2.1 Eksplorasi Data

Untuk mengetahui korelasi antar fitur digunakan digunakan heatmap dengan hasil sebagai berikut.



Semakin cerah warna kotak pada heatmap tersebut artinya semakin tinggi pula nilai korelasinya. Contoh fitur yang memiliki korelasi tinggi adalah SuhuMax dan Suhu3pm yaitu 0.98. Berdasarkan heatmap di atas nilai korelasi untuk fitur SuhuMax dan Kelembaban9am adalah -0.5 yang artinya kedua fitur tersebut salingberbanding terbalik, ketika nilai SuhuMax tinggi maka kelembaban9am rendah.

2.2 Persiapan Data

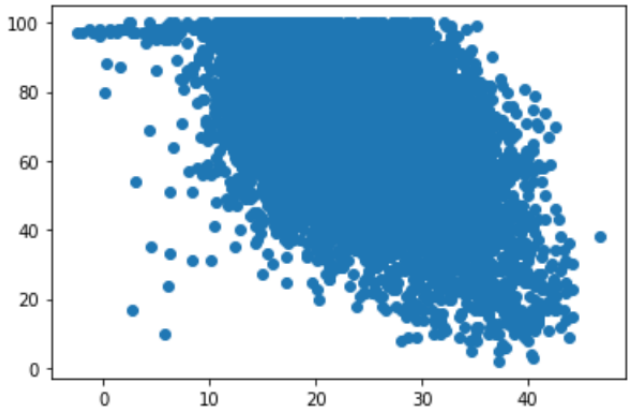
Pada tugas ini data yang diambil adalah sebanyak 10000 data. Selanjutnya data yang memiliki nilai null akan di drop sehingga dari 10000 data yang diambil hanya 9778 yang akan digunakan.

9750	9970	27.3	90.0
9751	9971	13.6	79.0
9752	9972	20.0	81.0
9753	9973	24.9	59.0
9754	9974	23.1	51.0
9755	9975	27.0	26.0
9756	9976	23.9	34.0
9757	9977	21.8	56.0
9758	9978	30.9	56.0
9759	9979	10.9	96.0
9760	9980	32.8	75.0
9761	9981	24.8	95.0
9762	9982	9.3	92.0
9763	9983	14.3	81.0
9764	9984	24.5	53.0
9765	9985	30.7	64.0
9766	9986	39.9	8.0
9767	9987	16.9	39.0
9768	9988	34.6	24.0
9769	9989	25.6	50.0
9770	9991	31.5	22.0
9771	9992	30.3	80.0
9772	9993	13.9	99.0
9773	9994	17.0	59.0
9774	9995	27.0	66.0
9775	9996	18.6	78.0
9776	9997	21.3	71.0
9777	9998	12.3	83.0

```
#Drop nilai null

df_cluster.dropna(inplace=True)

missing_data = df_cluster.isnull()
for column in missing_data.columns.values.tolist():
  print(column)
  print (missing_data[column].value_counts())
  print("")


df_cluster.info()
```

```
SuhuMax
False    9778
Name: SuhuMax, dtype: int64

Kelembaban9am
False    9778
Name: Kelembaban9am, dtype: int64
```

Berikut ini adalah plot data yang akan digunakan.



3. Permodelan

Pada tugas ini, K-Means akan digunakan untuk memodelkan data. K-Means memiliki tahapan proses sebagai berikut:

1. Mendefinisikan jumlah cluster yang akan dibentuk
2. Memilih centroid awal secara acak
3. Menghitung jarak setiap data pada semua centroid

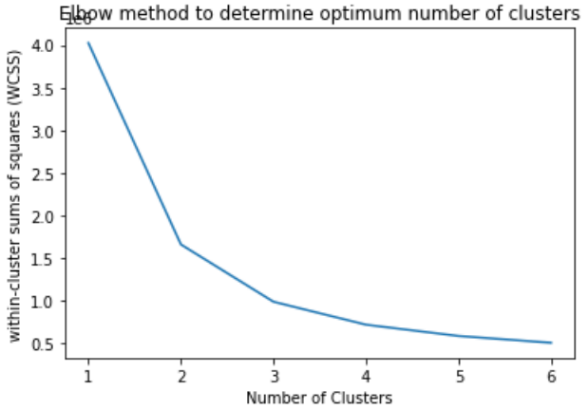
4. Mengelompokkan data pada centroid yang memiliki jarak paling minimum diantara jarak ke centroid lainnya.
5. Menentukan centroid baru dengan menghitung rata-rata data agar centroid berada di tengah-tengah.
6. Lakukan Langkah ke 3-5 hingga mendapatkan centroid yang konvergen

4. Evaluasi

Elbow method merupakan metode evaluasi yang akan digunakan pada tugas ini. Elbow method digunakan untuk mendapatkan nilai k yang optimal. Berikut tahapan proses yang digunakan dalam elbow method.

1. Melakukan clustering dengan menggunakan K-Means dan menggunakan nilai K yang berbeda.
2. Menghitung nilai WCSS (*Within-Cluster Sums of Squares*) untuk setiap K
3. Plot nilai WCSS dengan nomor cluster K
4. Indikator dari nilai K atau jumlah cluster yang sesuai dapat dilihat ketika terdapat siku pada plot.

Berikut ini hasil evaluasi model K-Means dengan elbow method.



Dari gambar di atas dapat disimpulkan bahwa nilai k optimal dari model k-means yang digunakan pada dataset salju dengan banyak iterasi 100 adalah 2.

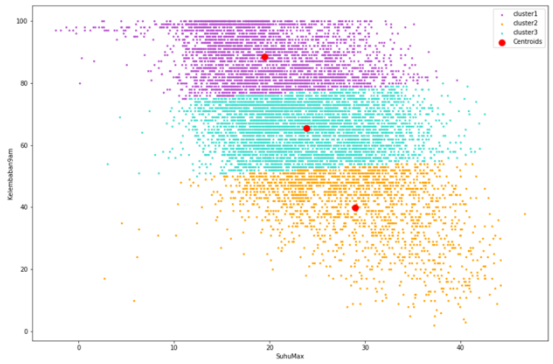
5. Eksperimen

Eksperimen yang dilakukan adalah dengan menggunakan dataset salju dengan jumlah data sebannyak 10000. Fitur yang digunakan pada eksperimen ini adalah SuhuMax dan Kelembaban9am. Nilai korelasi antara kedua fitur tersebut adalah -0,5 artinya kedua fitur tersebut berbanding terbalik, ketika SuhuMax tinggi maka Kelembaban9am rendah dan begitupun sebaliknya. Sebelum melakukan clustering data yang bernilai null di drop terlebih dahulu sehingga jumlah data yang dipakai hanya 9778.

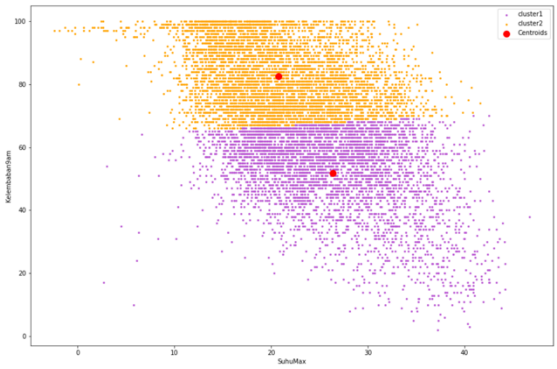
Nilai k yang digunakan pada eksperimen adalah 3 dan 2 dan iterasi sebanyak 100 kali. Nilai k=2 diambil berdasarkan hasil dari elbow method.

Berikut adalah hasil clusteringnya.

K = 3



K = 2



Berdasarkan hasil clustering di atas, terdapat perbedaan pada kedua hasil eksperimen tersebut.

1. Perbedaan posisi centroid
2. Pada eksperimen dengan k=3 data yang berada pada cluster yang satu dan lainnya tidak terlalu jauh berbeda sehingga sulit untuk melihat perbedaan data atau menganalisis maksud dari cluster tersebut sedangkan untuk k = 2, data dapat dibedakan lebih mudah yaitu, data dengan Kelembaban9am tinggi dan SuhuMax rendah dan Kelembaban rendah dengan SuhuMax tinggi.

6. Kesimpulan

Berdasarkan hasil evaluasi dari model K-Means yang dibangun, nilai k yang paling optimal adalah 2. Hal ini juga didukung dengan hasil eksperimen yang menunjukkan bahwa clustering dengan jumlah cluster = 2 terlihat lebih mudah dipahami serta memiliki perbedaan cluster yang satu dan yang lainnya terlihat lebih jelas.

Link presentasi : https://youtu.be/1mix0_ltjqc