LAPORAN TUGAS BESAR

“UNSUPERVISED LEARNING”

Disusun Untuk Memenuhi Tugas Matakuliah Pembelajaran Mesin

DISUSUN OLEH:

Dhimas Hafid Kurniawan

(1301184054)

PROGRAM STUDI INFORMATIKA

FAKULTAS INFORMATIKA

TELKOM UNIVERSITY

2020/2021

1. **Pokok Masalah**

   Membuat sistem unsupervised yang dapat memprediksi apakah besok akan turun salju atau tidak berdasarkan temperature, hujan, sinar matahari dan awan pada jam 3 sore dengan dataset salju_train.csv.

2. **Spesifikasi Program**

   Program berupa sistem yang mencakup:

   - Perancangan program menggunakan bahasa python dengan compiler google colab (colab.research.google.com/)
   - Library yang digunakan
     - Panda untuk CRUD data file
     - Numpy untuk CRUD operasi matematika
     - Csv untuk CRUD file format csv
     - Matplotlib untuk CRUD graphic 2D
     - Seaborn untuk CRUD graphic 2D berbasis marplotlib
     - Io untuk upload file
     - Random untuk generate angka random

3. **Strategi Penyelesaian Masalah**

   - Data Understanding

     Dataset yang didapatkan adalah salju.csv, dataset ini adalah data source dengan pengelompokan data sesuai nama kolom yaitu id, tanggal, kode lokasi, suhu min, suhu max, hujan, penguapan, sinar matahari, arah angin terkencang, kecepatan angin terkencang, arah angin 9am, arah angin 3pm, kelembaban 9am, kelembaban 3pm, tekanan 9am, tekanan 3pm, awan 9am, awan 3pm, suhu 9am, suhu 3pm, bersalju hari ini dan bersalju besok dengan mayoritas data bertipe float64.

   - Data Exploration

     Teknik mencari data yang berguna untuk dianalisis dengan cara yaitu penghitungan record data berupa 109096 record data, mencari missing values dan mencari duplicated data yang mengubah data menjadi 42411 record data mengidentifikasi data dengan mengelompokan tipe data tiap kolom pada dataset dengan tipe data jenis float64 sebanyak 16 kolom dan object sebanyak 7 kolom.

   - Data Cleansing

     Mengatasi permasalahan yang telah ditemukan pada data exploration contohnya membuat fungsi untuk mengatasi missing value, duplicate record, useless coloumns serta fungsi untuk melakukan penghapusan kolom serta nilai dari kolom dataset tersebut yang tidak diperlukan pada proses unsupervised clustering yaitu id, tanggal, kode lokasi, suhu min, suhu max dan kolom selain kolom yang digunakan untuk clustering prediksi cuaca bersalju.

- Feature Engineering
  Melakukan transforming data (scaling) dari bentuk dua dimensi (array, tabel) ke grafik diagram atau pie chart (particular range) yang range data sangat penting dan sensitive untuk algoritma yang menggunakan distance seperti pada metode K-Means dan euclidians distance.
- Data Modeling
  Menggunakan metode K-Means dengan langkah-langkah sebagai berikut:
  - Menentukan centroid awal
  - Memasukkan setiap objek ke cluster paling dekat
  - Update tiap centroid dengan menentukan titik tenah cluster
  - Lakukan berulang hingga tidak ada titik perubahan cluster

  Rumus yang digunakan sebagai berikut:

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

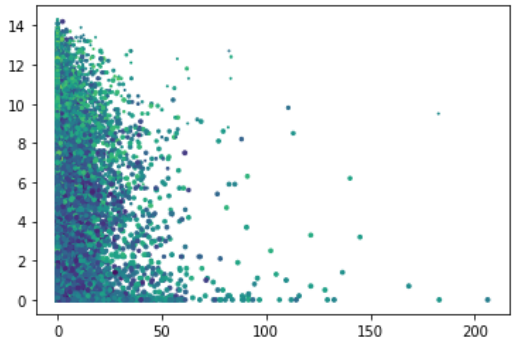
  Diketahui d adalah jarak, p dan q sebagai atribut serta i adalah jumlah iterasi untuk setiap data yang tersedia dalam dataset. p dan q adalah dua poin pada euclidian sebanyak n. p_i dan q_i adalah euclidian vector (initial point)Adapun Dataset yang digunakan dibedakan menjadi beberapa atribut, yaitu temperature dalam celcius, ujan dalam mm, sinar matahari dan cloud 3pm dalam rasio

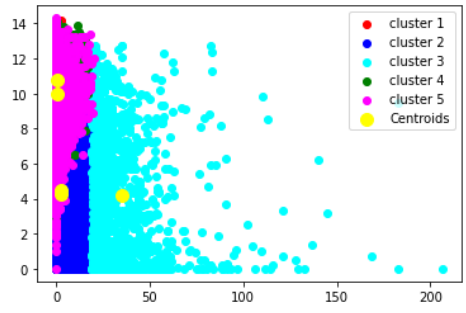
4. Analisis Hasil Eksperimen

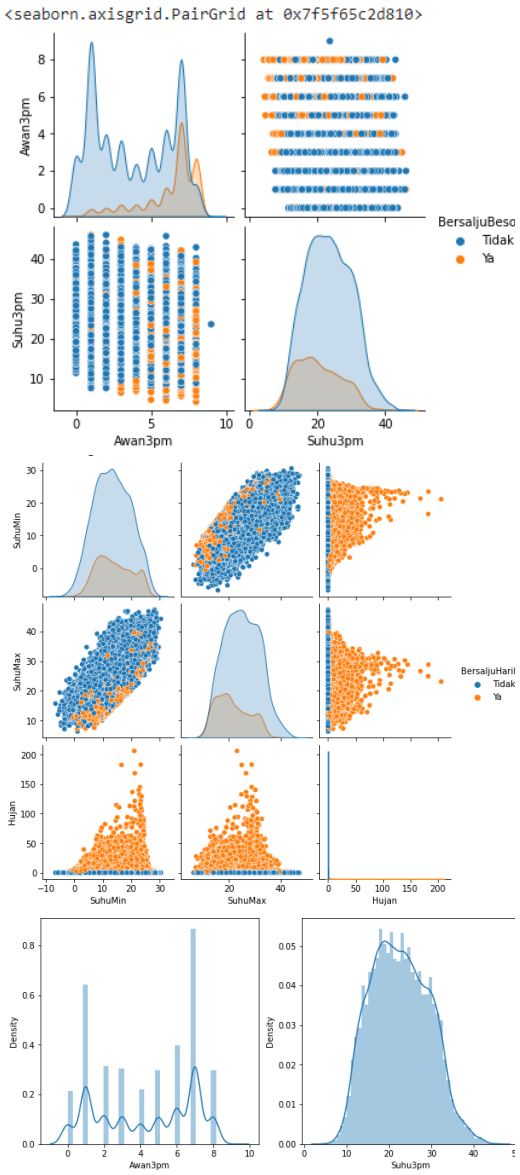
Nilai jumlah clustering pada program ini dibatasi menggunakan jumlah nilai input K yang diberikan pada prosedur tepat sebelum k_means. Setelah itu akan ditampilkan apakah terdapat perubahan centroid setelah clustering dan sebelum clustering untuk mengecek apakah fungsi yang diberikan telah bekerja. Centroid setelah clustering akan direpresentasikan dalam bentuk grafik dengan centroid serta cluster sebanyak input yang diberikan sebelumnya.

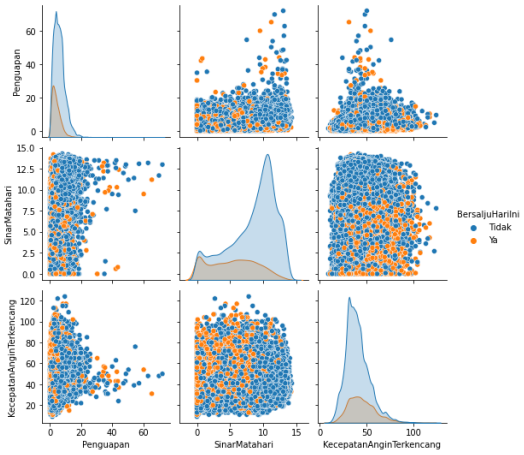
5. Output Program

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6 entries, 3 to 9
Data columns (total 23 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Tanggal                 6 non-null      object
 1   KodeLokasi              6 non-null      object
 2   SuhuMin                 6 non-null      float64
 3   SuhuMax                 6 non-null      float64
 4   Hujan                   6 non-null      float64
 5   Penguapan               6 non-null      float64
 6   SinarMatahari           6 non-null      float64
 7   ArahAnginTerkencang     6 non-null      object
 8   KecepatanAnginTerkencang 6 non-null     float64
 9   ArahAngin9am            6 non-null      object
 10  ArahAngin3pm            6 non-null      object
 11  KecepatanAngin9am       6 non-null      float64
 12  KecepatanAngin3pm       6 non-null      float64
 13  Kelembaban9am           6 non-null      float64
 14  Kelembaban3pm           6 non-null      float64
 15  Tekanan9am              6 non-null      float64
 16  Tekanan3pm              6 non-null      float64
 17  Awan9am                 6 non-null      float64
 18  Awan3pm                 6 non-null      float64
 19  Suhu9am                 6 non-null      float64
 20  Suhu3pm                 6 non-null      float64
 21  BersaljuHariIni         6 non-null      object
 22  BersaljuBesok           6 non-null      object
dtypes: float64(16), object(7)
memory usage: 1.1+ KB
```









6. Kesimpulan

- Dalam prediksi data cuaca bersalju menggunakan algoritma k-means berdasarkan temperature, hujan, sinar matahari dan awan sore hari memanfaatkan penghentian iterasi ketika pusat cluster ditemukan dan output dari algoritma k-means merupakan deret cluster dan satu centroid serta beberapa derajat keanggotaan untuk tiap-tiap data.
- Hasil analisa algoritma k-means dapat dikembangkan data prediksi cuaca bersalju berdasarkan parameter lainnya seperti sifat hujan atau sifat cuaca dari beberapa hari terakhir karena k-means dapat menentukan juga lokasi terbaik cluster berdasarkan proses iterasinya.