Akmal Ikhsan (1301184059, IF-42-08)
Link Channel : https://youtu.be/Kftpsyagn7g

Laporan Tugas Besar Machine Learning

A. 4 **Formulasi Masalah** yang diterapkan berupa :

1. Articulate Your Problem Clearly
   Program ini akan mengatasi masalah untuk seseorang yang ingin memprediksi apakah besok akan turun salju atau tidak.

2. Identify Your Data Sources
   Subjek yang dituju adalah mereka yang ingin mengetahui kondisi cuaca di esok hari.

3. Identify Potential Learning Problems
   - Untuk Clustering, solusi yang ditampilkan berupa mengelompokkan beberapa data sesuai dengan Labelnya
   - Untuk Classification, solusi yang ditampilkan yaitu memprediksi apakah akan turun salju hari ini atau besok atau tidak keduanya

4. Think About Potential Bias and Ethics
   Fitur-fitur yang digunakan untuk melengkapi permasalahan yang ada sudah sesuai dengan label yang di ditampilkan. Fitur yang digunakan adalah SuhuMin, SuhuMax.

B. **Data Exploration** adalah langkah awal dalam analisis data, di mana pengguna mengeksplorasi set data besar dengan cara yang tidak terstruktur untuk mengungkap pola awal, karakteristik, dan tempat menarik.
   1. Terdapat 18182 baris data dan 23 Fitur, pada dataset yang digunakan.
   2. Jenis data Object, Terdapat 6 Fitur yang digunakan pada dataset ini yaitu KodeLokasi, ArahAnginTerkencang, ArahAngin9am, ArahAngin3pm, BersaljuHariIni, dan BersaljuBesok.
   3. Jenis data Float, ada 17 Fitur yang digunakan pada dataset ini, yaitu : SuhuMin, SuhuMax, Hujan, Penguapan, KecepatanAnginTerkencang, SinarMatahari, KecepatanAngin9am, KecepatanAngin3pm, Kelembaban9am, Kelembaban3pm, Tekanan9am, Tekanan3pm, Awan9am, Awan3pm, Suhu9am, Suhu3pm.

Teknik yang digunakan yaitu **Distribution, Heatmap, Boxplot**

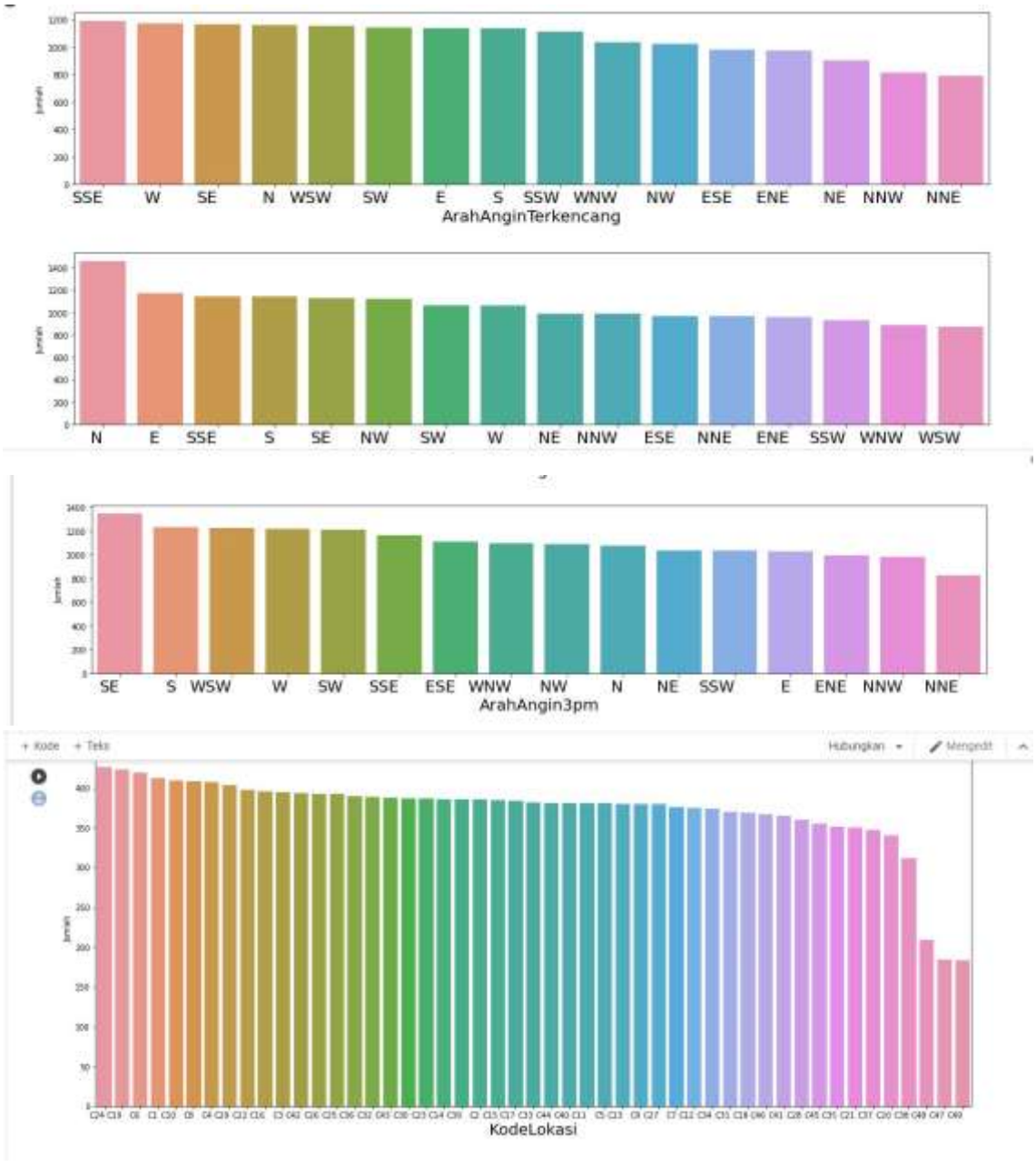
Sebelum mengeksplorasi data, tahap yang harus dilakukan yaitu mengkategorikan semua fitur menjadi 3 bagian kategori :

● Data berjenis object dikategorikan sebagai col_ArahAngin.

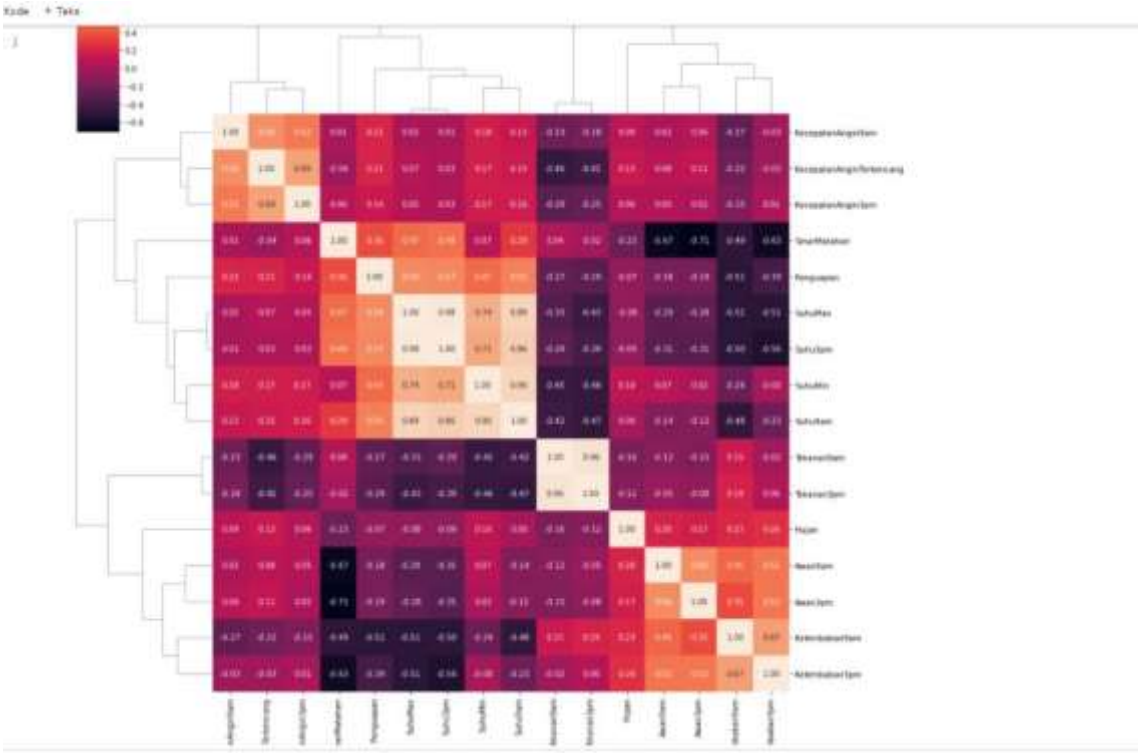
● Data berjenis Float sebagai col_number.

● Fitur BersaljuHariIni, BersaljuBesok, dan tanggal karena nilainya tidak akan digunakan, maka fitur tersebut di drop.

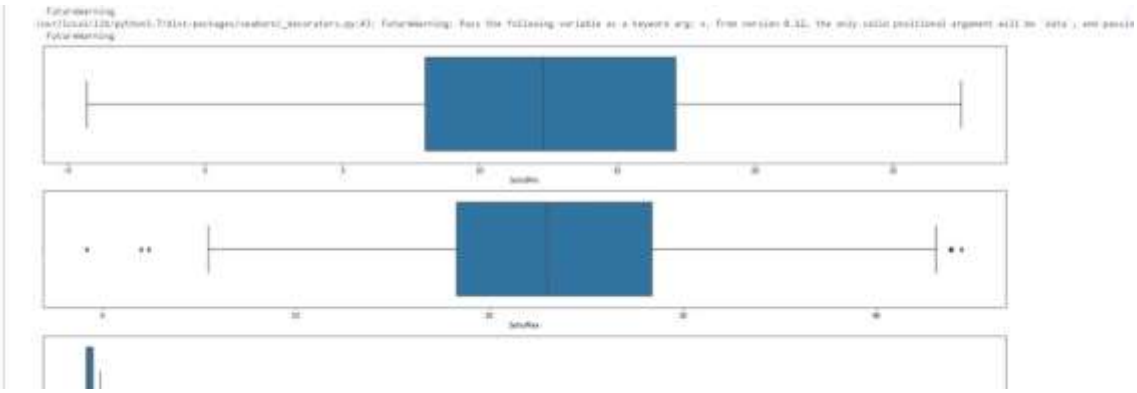
Untuk **distribution** merupakan teknik yang digunakan untuk menampilkan data yang berada pada fitur dengan jenis data berupa **Col_ArahAngin/Categorical.** ini adalah salah satu contoh dari fitur ArahAngin.



Untuk **Heatmap** merupakan teknik yang digunakan untuk melihat nilai korelasi antar fitur dengan fitur lainnya.



Untuk **Boxplot** merupakan teknik yang digunakan untuk menampilkan data outlier yang berada pada fitur dengan jenis data berupa **Col_number/Numeric,** dimana jenis datanya **Float.**



Data Cleansing merupakan proses mengidentifikasi bagian data yang salah, tidak lengkap, tidak akurat, tidak relevan atau hilang dan kemudian memodifikasi, mengganti atau menghapusnya sesuai dengan kebutuhan.

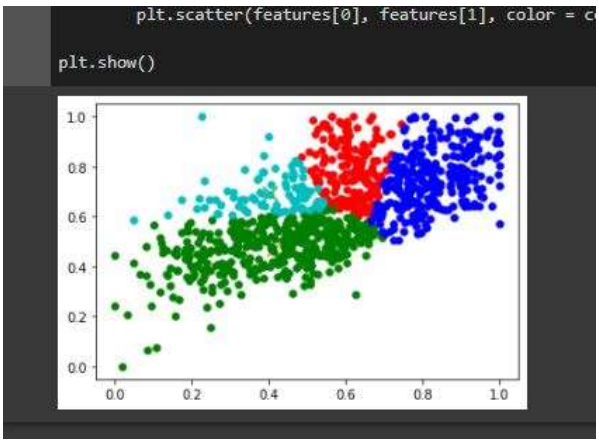
Pada tahap sebelumnya, program ini sudah membaca bahwa fitur mana sajakah yang memiliki missing value. Pada program ini cara yang dilakukan untuk membersihkan data tersebut yaitu dengan cara **mereplace** data tersebut sesuai dengan jenis datanya.

Lalu tahap yang dilakukan selanjutnya yaitu menghapus baris dengan nilai yang sama/duplicate dan mengubah data outlier untuk memberikan hasil data yang lebih baik. Setelah Data Cleansing kita akan melakukan **Features Engineering** yang mana fitur ini merupakan proses dimana fitur akan dipilih atau diseleksi untuk digunakan.

Setelah fitur itu diseleksi kemudian semua fitur tersebut akan discalling. **Scalling** merupakan proses penskalaan atau penyamarataan data. Proses Scalling yang digunakan adalah Min-Max Normalization yaitu proses normalisasi yang mengambil nilai min dan max dari fitur yang digunakan.

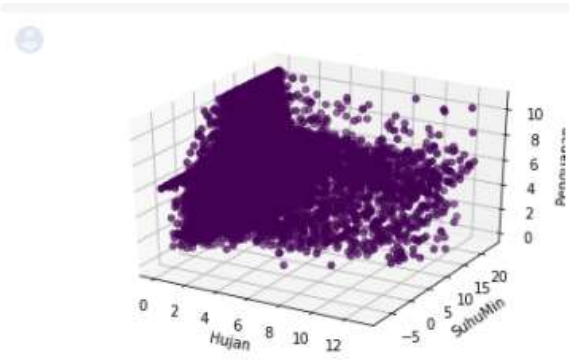
C. **Modelling** yang digunakan pada program ini adalah **K-Means** karena algoritma ini mudah dilakukan saat pengimplementasian dan dijalankan. Algoritma ini juga tergolong fleksibel dan menggunakan prinsip yang sederhana.

   Centroid yang di bentuk sesuai dengan nilai K, Cluster menggunakan K. gambar berikut adalah hasil output dari k = 4, sehingga K membagi data plotting menjadi banyak cluster.



D. **Eksperimen** disini ialah pemodelan 3 fitur dataframe, sehingga menghasilkan grafik 3 dimensi, karena terdapat kesalahan pada K-means dan Cluster maka output yang

dihasilkan tidak sesuai yang diharapkan. Gambar berikut adalah hasil visualisasi dari grafik, dengan fitur yang digunakan yaitu Suhumin, SuhuMax, dan Hujan.



Classification adalah proses memprediksi kelas atau kategori dari nilai yang diamati atau titik data yang diberikan. Pada program ini menggunakan algoritma **K-Nearest Neighbor (K-NN)** karena lebih efektif di data training yang besar dan dapat menghasilkan data yang lebih akurat. Tujuan dari algoritma ini untuk memprediksi sebuah data.

dan data tersebut akan di kategorikan kedalam kelompok apa saja. Sebagai contoh program ini memilik sebuah kelompok kelas. Dari fitur tersebut akan di pilih berdasarkan fitur suhu_min dan suhu_max. Sebelum melakukan pengelompokkan, fitur tersebut akan di diencode agar mudah untuk di kelompokkan. Setelah itu tahap yang dilakukan selanjutnya yaitu melakukan pembagian data yang akan dipecah menjadi data train dan data test. Kemudian hasil output akan menampilkan sebuah matrix yaitu Confusion Matrix yang isi nya merupakan nilai dari **True Positive and True Negative, False Positive and False Negative.**

Kesimpulan

Untuk Clustering algoritma yang digunakan adalah **K-Means** karena algoritma ini mudah dilakukan saat pengimplementasian dan dijalankan. Algoritma ini juga tergolong fleksibel dan menggunakan prinsip yang sederhana, dan Centroid mengelompokkan SuhuMin dan SuhuMax pada algoritma Clustering berdasarkan K-means/Cluster yang sesuai.

Untuk Classification algoritma yang digunakan adalah **K-Nearest Neighbor (K-NN)** karena lebih efektif di data training yang besar dan dapat menghasilkan data yang lebih akurat.