

## **Modul 2 praktikum datamining**

### **Eksplorasi data menggunakan matplotlib**

---

#### **1. Jelaskan makna atribut dataset iris. Jelaskan pula makna kelas dataset iris**

##### **Dataset**

Dataset merupakan kumpulan set dari data data yang dikumpulkan untuk tujuan utama agar dapat diproses oleh computer yang mana dapat diakses, diubah, dan di 'manipulasi'-kan

##### **Data Training**

Data yang biasanya digunakan untuk melatih sebuah algoritma. Biasanya digunakan untuk pelatihan AI / Machine Learning

##### **Data Testing**

Berbeda dengan Data Training, Data Testing digunakan untuk mengetahui lebih jauh tentang performa algoritma yang telah dibangun sebelumnya, dengan menggunakan data yang 'baru' alias tidak dikenal oleh algoritmanya. Diharapkan algoritma tersebut dapat beradaptasi

#### **1. Jelaskan proses penyusunan dataset**

Karena Dataset ini digunakan oleh pemrosesan computer, maka perlu sebuah 'format' agar memang dapat dikenali oleh computer. Bagaimana agar dikenali oleh computer?

- Tentunya kita dapat memanfaatkan table, agar lebih terorganisir
- Nama Kolom harus termuat dalam 1 baris
- Menggunakan tipe data yang spesifik per kolom

Kita juga dapat melakukan proses penyusunan datases melalui :

- Penentuan atribut (property objek) dari suatu data
- Mengisikan data dari atribut yang sudah ditentukan
- Menentukan pola dari atribut yang sudah dimiliki
- Dari data dan pola yang sudah ada kita menentukan kelas yang dihasilkan berdasarkan nilai atributnya.

**2. Jelaskan makna attribut dataset iris. Jelaskan pula makna kelas dataset iris**

- Attribut pada Dataset Iris menunjukkan seberapa Panjang / lebar dari petal dan sepal iris dengan variasi kelas kelasnya.
- Sedangkan Kelas pada dataset iris dimaksudkan untuk mengkategorikan 'variasi' dari iris itu sendiri

**6. Jelaskan bagaimana karakter pengelompokan setiap kelas pada scatterplot**

Berdasarkan Kelas :

- a. Setosa, Berada paling bawah yang mana memiliki Panjang / lebar paling kecil, terutama pada sepal length nya, kebanyakan mendekati 0 serta bergerombol dengan jarak yang dekat. Pada SW nya ada beberapa data yang berjarak 'jauh'
- b. Versicolor, Cenderung berdempetan dengan Kelas Virginica. Cenderung berada di tengah. Rata rata nilainya berdekatan
- c. Virginica, dapat dilihat dari scatternya, bahwa jenis ini paling Panjang / besar, baik dari segi petal dan sepal nya. Bisa dilihat kalau jenis ini yang paling atas, disetiap pasangan attribut
- d. Unknown, Cenderung mendekati kepada ke-3 variasi yang lainnya, meskipun tersebar secara keseluruhan mendekati variasi" lainnya

**7. Perhatikan data yang berjenis unknown. Jelaskan bagaimana cara meramal jenis data unknown**

Kelas ini memiliki nilai yang mana selalu muncul pada jarak yang dekat di ke-semua kelas, terkadang antar value / nilai, memiliki jarak yang 'jauh'. Yang bisa saya ambil dari kelas ini. Data nya cenderung akan berada pada didekat data kelas lain yang berdempetan / berjarak dekat

**9. Jelaskan bagaimana karakter pengelompokan setiap kelas pada scatterplot**

- A. Setosa, Penyebarannya Cenderung di bawah
- B. Versicolor, Penyebarannya cenderung diantara bawah dan/atau tengah
- C. Virginica, Penyebarannya cenderung mengarah keatas, meskipun ada banyak pula yang ada di pertengahan

D. Unknown, Penyebarannya dapat dibilang tidak menentu. Penyebarannya setiap attribute cukup bervariasi. Yang bisa diambil adalah, pada setiap attribute, kelas unknown ini cenderung mendekat ke tiga variasi lainnya