**Data Mining**

**Dataset**

Data adalah fakta dan angka (dapat juga disebut sebagai data mentah) yang berhubungan dengan konteks suatu permasalahan, Data terdiri dari dua aspek yaitu Object dan atribut, contoh object manusia, pohon, binatang, contoh atribut misalnya umur, tinggi badan, berat badan. Dataset merupakan kumpulan objek data. Dataset memiliki nama lain record,point,vector,pattern,event,observasi,case atau data. Object data digambarkan dengan

menggunakan sejumlah atribut yang menangkap karakteristik dari object data tersebut. Atribut disebut juga sebagai karakteristik,variabel,field,fitur,atau dimensi.

**Atribut**

Atribut merupakan sifat dari suatu object data yang nilainya bisa bermacam-macam diantara objectobject data yang diamati, misalnya tinggi badan abdul bisa berbeda dengan tinggi badan asep, berat badan abdul bisa berbeda dari waktu ke waktu. Warna kulit bisa memiliki nilai

[kuninglangsat,hitam,putih,sawomatang], dan nilai dari tinggi badan bisa berupa angka numerik misalnya 165,170,180.

Atribut memiliki jenis nilai yang beragam,misalnya berat badan mempunyai nilai dengan jenis

numerik (kuantitatif) sehingga bisa dibandingkan satu dengan yang lainnya. Sedangkan warna kulit tidak bisa dibandingkan karena karena jenis nilainya bersifat kualitatif. Pada umumnya tipe atribut terdiri dari kualitatif(diskrit) dan kuantitatif(numerik). Sifat penting yang dimiliki atribut diantaranya distinctness(=,<>),order(<.<=,>=,>),addition(+,-),multiplication (\* , /).

**Tabel tipe atribut**

|  |  |  |  |
| --- | --- | --- | --- |
| Tipe Atribut | | Keterangan | Contoh |
| Diskrit(kualitatif) | Nominal | Nilai atribut berupa nama yang membedakan dengan nilai lainnya(=,<>). | Nim,kodepos,noktp,jenis kelamin |
| Ordinal | Nilai atribut berupa nama yang memiliki arti informasi terurut(<,<=,>=,> ) | Kondisi barang misal : (Istimewa, baik, sedang, cukup,buruk) Tingkat |
| Numerik(kuantitatif) | Interval | Perbedaan dari dua nilai atribut memiliki makna berarti (+,-) data diukur pada skala interval | Tanggal,temperatur, rating IQ |
| Rasio | Perbedaan dan rasio dari dua nilai atribut memiliki perbedaan berarti (\*,/) | Tinggi,panjang,umur |

Secara umum, atribut/variabel dapat diukur pada empat skala yang berbeda. Mean, median, dan modus adalah cara untuk memahami central tendency, yaitu titik tengah dari distribusi data. Standar deviasi, varian, dan range adalah ukuran dispersi yang paling umum digunakan untuk memahami penyebaran data.

Atribut merupakan faktor atau parameter yang menyebabkan class/label/target terjadi, sementara itu suatu Class adalah Atribut yang akan dijadikan target dan class ini sering disebut sebagai label(AtributTarget) .

**Jumlah Peserta Pelatihan Manajerial Berdasarkan Jenis Kelamin**

Open data Repository (<https://opendata.jatimprov.go.id/frontend/dataset/1358/detail_dataset>)

|  |  |  |  |
| --- | --- | --- | --- |
| **periode\_update** | **nama\_diklat** | **laki\_laki** | **perempuan** |
| Semester II 2020 | PKN II | 50 | 10 |
| Semester II 2020 | PKA | 184 | 66 |
| Semester II 2020 | PKP | 148 | 103 |
| Semester II 2020 | LATSAR | 925 | 2218 |
| Semester II 2021 | PKN II | 102 | 13 |
| Semester II 2021 | PKA | 485 | 160 |
| Semester II 2021 | PKP | 249 | 208 |
| Semester II 2021 | LATSAR | 3665 | 7281 |
| Semester I 2022 | PKN II | 171 | 19 |
| Semester I 2022 | PKA | 1104 | 670 |
| Semester I 2022 | PKP | 212 | 126 |
| Semester I 2022 | LATSAR | 1290 | 3897 |

<https://opendata.jatimprov.go.id/frontend/dataset/1369/detail_dataset>

<https://opendata.jatimprov.go.id/frontend/dataset/1358/detail_dataset>