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, tinggibadan, Beratbadan.

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 object-object 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.

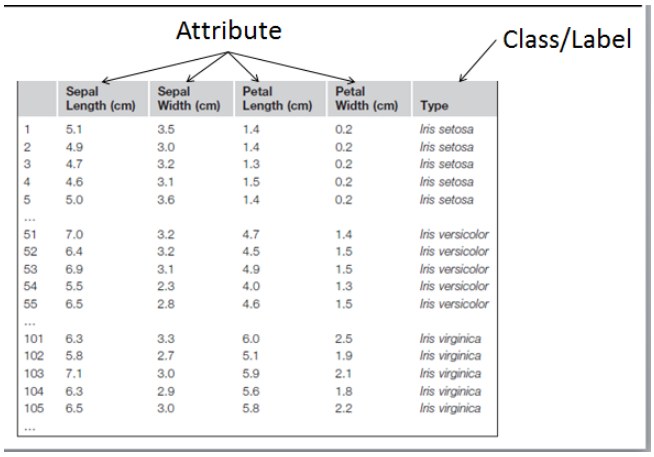
Atribut memiliki jenis nilai yang beragam,misalnya berat badan mempunyai nilai dengan jenis numerik (kuantitatif) sehingga bisa dibandingkan satu dengan yang lainnya. Sedangkan warna kulit tidak bisa dibandingkan karena karena jenis nilainya bersifat kualitatif. Pada umumnya tipe atribut terdiri dari 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,no-ktp,jenis kelamin
	Ordinal	Nilai atribut berupa nama yang memiliki arti informasi terurut(<,<=,>=,>)	Kondisi barang misl : (Istimewa, baik, sedang, cukup,buruk) Tingkatkelulusan,nilai huruf
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) .



	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
...					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris versicolor*
53	6.9	3.1	4.9	1.5	*Iris versicolor*
54	5.5	2.3	4.0	1.3	*Iris versicolor*
55	6.5	2.8	4.6	1.5	*Iris versicolor*
...					
101	6.3	3.3	6.0	2.5	*Iris virginica*
102	5.8	2.7	5.1	1.9	*Iris virginica*
103	7.1	3.0	5.9	2.1	*Iris virginica*
104	6.3	2.9	5.6	1.8	*Iris virginica*
105	6.5	3.0	5.8	2.2	*Iris virginica*
...					

Gambar atribut dan class

Jenis Dataset terdiri dari private dan public.

Private Dataset: data set dapat diambil dari organisasi yang dijadikan obyek penelitian seperti misalnya Bank, Rumah Sakit, Industri, Pabrik, Perusahaan Jasa.

Public Dataset: data set dapat diambil dari repositori pubik yang disepakati oleh para peneliti data mining,misalnya:UCI Repository (*http://www.ics.uci.edu/~mlearn/MLRepository.html*), ACM KDD Cup (*http://www.sigkdd.org/kddcup/*)

Normalisasi data

Skala pengukuran untuk variabel yang berbeda cenderung bervariasi, sehingga analisis dengan pengukuran yang mentah dapat mengarah ke variabel dengan nilai absolut yang lebih tinggi. Membuat semua jenis unit variabel yang berbeda ke dalam urutan yang sama besarnya sehingga bisa menghilangkan potensi pengukuran outlier yang dapat mengakibatkan salah mengartikan temuan/pola dan hal ini mengurangi keakuratan dari kesimpulan yang dihasilkan. Dua metode yang umum digunakan untuk me-*rescaling* data adalah normalisasi dan standardisasi. Untuk melakukan normalisasi data dilakukan dengan menggunakan penskalaan Min-Max; dengan rumus seperti yang diberikan di bawah ini, hal ini akan menskalakan semua nilai numerik dalam rentang 0 hingga 1.

$$X_{normalized} = \frac{(X - X_{min})}{(X_{max} - X_{min})}$$

Catatan: outlier yang ekstrim harus dihapus sebelum menerapkan teknik di atas karena dapat memiringkan nilai normal dalam data ke interval kecil.

Teknik standardisasi akan mengubah variabel menjadi rata-rata nol dan standar deviasi menjadi bernilai satu. Formula untuk standardisasi diberikan di bawah ini dan hasilnya dikenal sebagai z-scores:

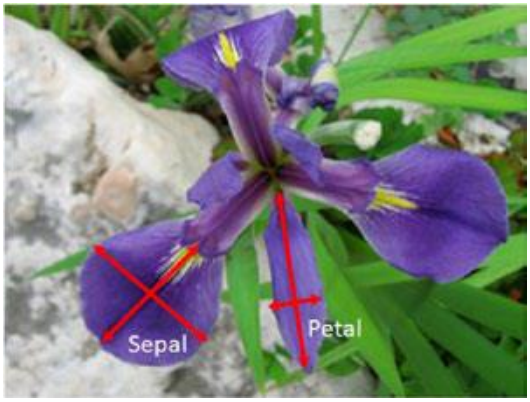
$$Z = \frac{(X - \mu)}{\sigma}$$

Dimana **μ** adalah mean dan **σ** adalah standar deviasi. Standardisasi sering menjadi metode yang disukai untuk berbagai analisis karena memberi tahu posisi di mana setiap titik data berada dalam distribusinya dan indikasi kasar keberadaan outlier.

Exploratory Data Analysis (EDA).

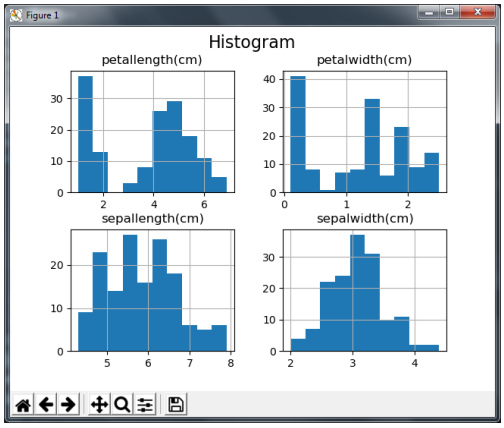
EDA adalah analisis dalam memahami data dengan menggunakan teknik peringkasan dan visualisasi.Pada tingkat lebih tinggi , EDA dapat dilakukan dalam dua cara, yaitu, analisis univariat dan analisis multivariat.

Mari kita pelajari contoh dataset untuk mempelajari penggunannya. Dataset iris adalah salah satu dataset terkenal yang digunakan secara luas dalam literatur pengenalan pola. Disediakan di UC Irvine Machine Learning Repository. Dataset ini berisi panjang kelopak, lebar kelopak, panjang sepal, dan pengukuran lebar sepal untuk tiga jenis bunga iris, yaitu, setosa,versicolor, dan virginica.



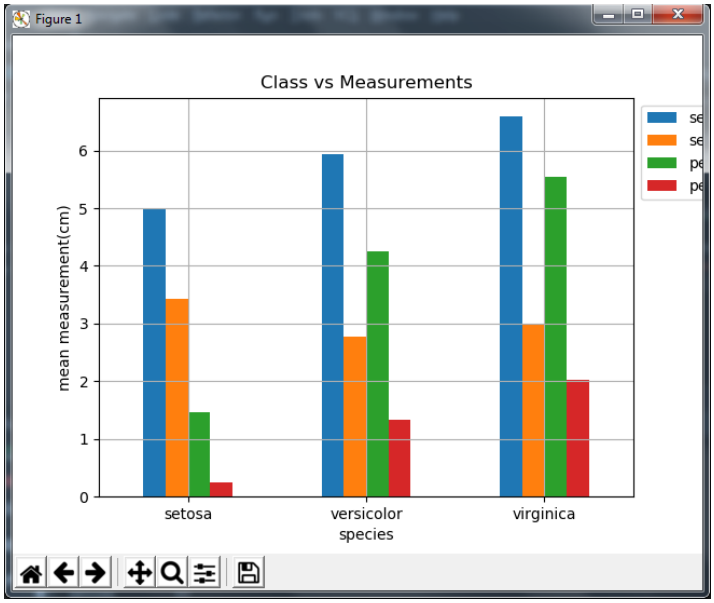
Analysis Univariate

Variabel individu dianalisis secara terpisah untuk memiliki pemahaman yang lebih baik tentang variabel tsb. Panda menyediakan fungsi uraian untuk membuat statistik ringkasan dalam format tabel untuk semua variabel. Statistik ini sangat berguna untuk jenis variabel numerik untuk memahami masalah kualitas seperti missng value dan keberadaan outlier.



Analisis multivariat

Analisis multivariate merupakan analisis terhadap hubungan antar variabel-variabel yang ada.



Matriks Korelasi

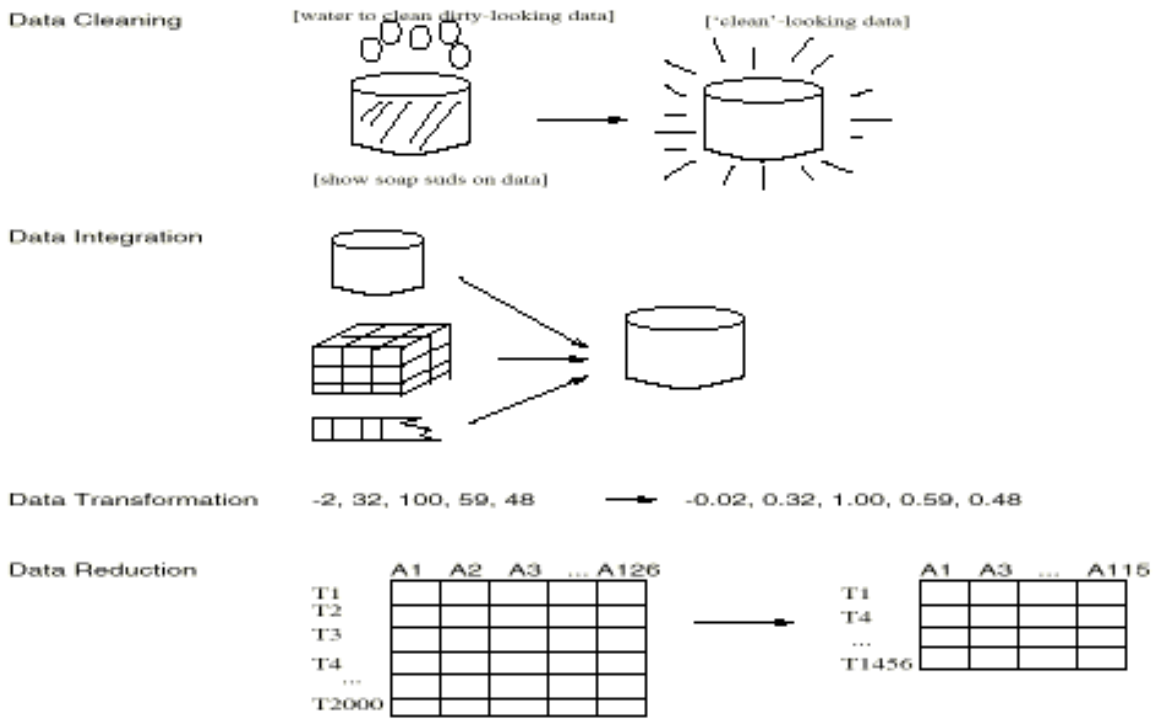
Koefisien korelasi digunakan di dalam statistik untuk mengukur seberapa kuat hubungan antara dua variabel. Fungsi korelasi menggunakan koefisien korelasi Pearson, yang menghasilkan angka antara -1 hingga 1. Hubungan negatif yang kuat ditunjukkan oleh koefisien yang mendekati angka -1 dan korelasi positif yang kuat ditunjukkan oleh koefisien yang mendekati angka 1.

Pemrosesan Awal Data

Dataset yang akan diproses menggunakan metode datamining seringkali harus melalui pekerjaan awal yang secara keseluruhan terpisah dari metode dalam datamining. Masalah-masalah seperti jumlah populasi data yang terlalu besar, terdapat data menyimpan (anomali), dimensi yang terlalu tinggi,terdapat fitur yang tidak memiliki kontribusi besar,dan lain-lain menjadi pemicu munculnya pemrosesan awal yang harus dilakukan sebelum akhirnya data dilepaskan untuk diproses dalam datamining.

Dalam datamining data dikatakan kotor apabila data tersebut terdapat noise(data error,outlier),tidak lengkap (nilai-nilai atribut kurang, atribut terpenting tidak disertakan, atau hanya memuat data agregasi),data tidak konsisten.

Tugas utama data pre-processing diantaranya adalah Pembersihan data,integrasi data,transformasi data,reduksi data.



Pembersihan data

Mengisi nilai-nilai yang hilang, menghaluskan noisy data, mengenali atau menghilangkan outlier, dan menghilangkan ke-tidak-konsistenan data.

Integrasi Data

Integrasi data banyak database, tabel, atau record.

Transformasi data

Transformasi data dengan melakukan Normalisasi dan agregasi

Reduksi data

Mendapatkan representasi data yang direduksi dalam volume tetapi menghasilkan hasil analitikal yang sama atau mirip dengan data sebelum reduksi.

Beberapa proses pengolahan data awal.

Agregasi

Penggabungan obyek ke dalam sebuah obyek tunggal Sum,average,min,max

Cabang	IDTX	Tanggal	Total
Gresik	2012102	30-01-2013	250,000
Gresik	2012103	30-01-2013	300,000
Surabaya	2012201	30-01-2013	500,000
Surabaya	2012202	30-01-2013	450,000
Surabaya	2012203	31-01-2013	350,000

Cabang	Tanggal	Total
Gresik	30-01-2013	550000
Surabaya	30-01-2013	950000
Surabaya	31-01-2013	350000

Sampling

Sampling adalah pemilihan bagian obyek data yang akan dianalisis.Sample yang diambil seharusnya representatif (mewakili seluruh data),sample disebut resprentatif jika mempunyai sifat yang sama dengan seluruh data yang biasa diukur dengan menghitung rata-rata/mean.

Penggunaan sample yang baik tidak menjamin bahwa hasil pemrosesan datamining pada data sample sama bagusnya dengan pemrosesan pada seluruh data asli (populasi). Pendekatan sampling menggunakan Simple random sampling tanpa pengembalian dan dengan pengembalian.

Binerisasi

Transformasi data dari tipe numerik ataupun diskret menjadi tipe biner. Contoh penggunaanya adalah pada algoritma asosiasi yang membutuhkan data dengan atribut bertipe biner.

Pada proses binerisasi jumlah atribut yang dibutuhkan (N) dapat ditentukan dengan menggunakan rumus N=log2(M), dimana M adalah jumlah class kategori.

Contoh: {rusak,jelek,sedang,bagus,sempurna}

Jumlah class (M)=5

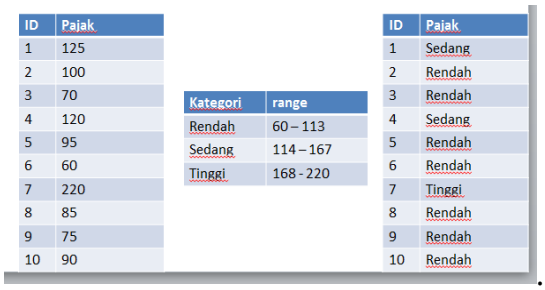
maka N=log2(5) = 3, sehingga hasil binerisasi berupa 3 atribut x1,x2,x3

Class	Nilai integer	x1	x2	X3
Rusak	0	0	0	0
Jelek	1	0	0	1
Sedang	2	0	1	0
Bagus	3	0	1	1
Sempurna	4	1	0	0

Variabel yang dihasilkan oleh proses binerisasi adalah x1,x2,dan x3.

Diskretisasi

Diskretisasi merupakan transformasi data dari tipe nnumerik menjadi tipe diskrit



ID	Pajak
1	125
2	100
3	70
4	120
5	95
6	60
7	220
8	85
9	75
10	90

Kategori	range
Rendah	60 – 113
Sedang	114 – 167
Tinggi	168 - 220

ID	Pajak
1	Sedang
2	Rendah
3	Rendah
4	Sedang
5	Rendah
6	Rendah
7	Tinggi
8	Rendah
9	Rendah
10	Rendah

.

Pengurangan Dimensi

Pengurangan dimensi atau pengurangan atribut diperlukan guna mengurangi jumlah waktu dan memory yg dibutuhkan, membuat data lebih mudah divisualisasi,dan membantu mengurangi fitur-fitur yang tidak relevan/mengurangi gangguan.Teknik yang digunakan dalam pengurangan dimensi Principal Component Analysis (PCA) dan Singular Value Decomposition(SVD).

Pemilihan Fitur (Feature Subset Selection)

Pemilihan Fitur merupakan proses pencarian terhadap semua kemungkinan subset fitur yang digunakan untuk datamining.Hal ini juga diperlukan untuk menghilangkan fitur yang redundan, misl: harga_jual,pajak,discount

Pekerjaan lainnya dalam pemilihan fitur adalah menghilangkan fitur-fitur yang tidak mengandung informasi yang berguna untuk pekerjaan datamining

Misl: tinggi badan mhs pada pekerjaan prediksi kelulusan mhs , tidak relevan.

Teknik-teknik yang digunakan dalam pemilihan fitur diantaranya : Brute-force yaitu pada proses data mining dilakukan dengan mencoba semua fitur.Filtering yaitu memilih/menyaring fitur sebelum proses datamining dilakukan secara manual.Wrapper menggunakan algoritma datamining utk memilih sub-set fitur yang paling baik.

Pembuatan Fitur

Proses membuat fitur baru yang dapat menangkap informasi penting dalam sebuah himpunan fitur yang lebih efisien daripada fitur-fitur yang ada.Metode Pembuatan Fitur:Ekstraksi Fitur,pemetaan menggunakan transformasi fourier/wavelet,konstruksi fitur dengan menggabungkan fitur-fitur yang ada.

Transformasi Fitur

Merupakan proses yang memetakan keseluruhan himpunan nilai dari fitur-fitur yang diberikan ke suatu subset nilai pengganti sedemikian sehingga nilai yang lama dapat dikenali dengan satu dari nilai-nilai yang baru tersebut.

Metode dalam transformasi fitur:Standarisasi (median , standar deviasi).

Normalization, dimana data sebuah atribut diskalakan ke dalam rentang (kecil) yang ditentukanŠ (Metode: Min-max Normalization, z-score Normalization,Normalization by Decimal Scaling).