Data Pre-processing : Data yang kita dapatkan biasanya belum dalam keadaan yang siap diolah oleh algortima DM

Data preprocessing adalah proses mengubah data mentah ke dalam bentuk yang lebih mudah dipahami.

Dalam *machine learning, data preprocessing* berperan memastikan bahwa *big data* sudah diformat dan informasi didalamnya dapat dipahami oleh algoritma perusahaan sehingga bisa mengeluarkan hasil yang lebih akurat.

Manfaat:

Memperlancar proses data mining

Membuat data lebih mudah untuk dibaca

Mengurangi beban representasi dalam data

Mengurangi durasi data mining secara signifikan

Mempermudah proses analisis data dalam machine learning

Tahapan:

Data Cleaning: data mentah akan dibersihkan melalui beberapa proses seperti mengisi nilai yang hilang, menghaluskan *noisy* data, dan menyelesaikan inkonsestensi yang ditemukan.

Data integration adalah tahap yang menggabungkan data dari berbagai sumber menjadi satu kesatuan data (*dataset*).

Data Transformation*:* data akan dinormalisasi dan digeneralisasi.

Data Reduction*:* pengurangan jumlah data

How to handle missing data:

Abaikan tupel: biasanya dilakukan ketika label kelas hilang (dengan asumsi tugas adalah klasifikasi — bukan
efektif dalam kasus-kasus tertentu)

• Isi nilai yang hilang secara manual: membosankan + tidak layak?

• Gunakan konstanta global untuk mengisi nilai yang hilang: misalnya, "tidak diketahui", kelas baru?!

• Gunakan atribut mean untuk mengisi nilai yang hilang

• Gunakan mean atribut untuk semua sampel dari kelas yang sama untuk mengisi nilai yang hilang:

• Gunakan nilai yang paling mungkin untuk mengisi nilai yang hilang

Exploratory data analys(EDA): proses untuk mengenali dan mencari pola pada dataset yang digunakan, proses ini dapat mendeteksi kesalahan sebelum nantinya akan dilakukan proses selanjutnya.

membantu menentukan model yang lebih sesuai dengan data

Apa saja yang dilakukan EDA:

Jumlah data per kelas. Seimbang tidak? -> seimbangkan

• Distribusi data di setiap kolom -> bar chart. Seperti di Kaggle (tab Data).

• Min, max, mean per kolom-> df.describe()

• Korelasi antar atribut. Korelasi masing-masing atribut terhadap kelas. -> Scatter plot / pair-plot. df.corr() atau semacamnya: correlation matrix, correlation heatmap.

• Distribusi data (bar chart) setiap kolom terhadap kelas. Seperti di Weka.

• Scatter plot dg warna titik sesuai kelas. 2D atau 3D. Untuk melihat "batas" visual antara dua kelas.

Cara meningkatkan akurasi:

1. **Parameter turning**: Pemberian nilai parameter yang berbeda akan mempengaruhi performa (akurasi) model.

Beberapa metode DM memiliki parameter yang harus ditentukan nilainya.

• KNN memiliki parameter K.

• K-Means memiliki parameter K.

• DBSCAN memiliki parameter eps dan minPts.

• C4.5 memiliki parameter minNumObj, etc

• XGBOOST memiliki parameter max_depth, n_estimators, etc.

•NB tidak memiliki parameter.

2.Grid search CV:

Grid = matrix kombinasi setting parameter • Search = mencari setting parameter terbaik • CV = cross validation

Overfitting, Underfitting:

Overfitting = model terlalu nge-fit (“ketat”) terhadap data. Terlalu detail / spesifik. Terlalu “kaku”.

Underfitting = model tidak begitu mengikuti “bentuk” data -> “bodoh” / ngaco.

Kenapa masih terjadi eror:

ada faktor lain yang mempengaruhi fenomena ini dan belum termasuk dalam model kita.

Karena data training kita terbatas, hanya sampel, belum cukup merepresentasikan segala macam kemungkinan yang ada

ARM:

Rumus Mencari Support

$$Support = \frac{Jumlah\ transaksi\ A}{Jumlah\ transaksi}$$

Rumus support metode association rule

Rumus Mencari Confidence

$$Confidence = \frac{Jumlah\ trans\ A + B}{Jumlah\ Trans\ A}$$

Rumus mencari confidence metode association rule

No	Barang
1	Gula, kopi
2	Permen, Marie, Gula
3	Sabun, Permen, Sikat Gigi
4	Permen, Snack, Marie
5	Air Mineral, Permen, Roti
6	Roti, Gula, Kopi, Kacang Kulit\
7	Roti, Permen
8	Pasta Gigi, Sikat Gigi
9	Roti, Permen, Snack
10	Snack, Permen, Marie
11	Susu, Kopi, Gula
12	Susu, Marie, Roti
13	Pasta Gigi, Sikat Gigi, Kopi
14	Permen, Snack
15	Kacang Kulit, Kopi, Gula
16	Gula Kopi, Snack
17	Snack, Susu, Gula
18	Kopi, Gula, Roti
19	Pasta Gigi, Gula, Susu
20	Snack, Kacang Kulit

menentukan nilai support dengan menghitung jumlah transaksi yang mengandung item A dibagi dengan jumlah transaksi yang ada, maka didapat
Contoh:
Gula, kopi 5 transaksi jadi 5/20 =0,25 atau 25%
Gula,permen 1 transaksi jadi 1/20=0,05 atau 5%

No	Jenis Barang	Jml	Sup
1.	Gula, Kopi	5	25 %
2.	Gula,Permen	1	5 %
3.	Gula, Susu	4	20 %
4.	Permen. Marie	3	5 %
5.	Permen, Sikat gigi	1	5 %
6.	Permen, Kacang Kulit	1	5 %
7.	Kopi, Susu	1	5 %
8.	Marie, Susu	1	5 %
9.	Sikat gigi, pasta gigi	2	10 %
10.	Pasta gigi, susu	1	5 %

Setelah mendapat nilai support maka langkah selanjutnya **adalah mencari nilai confidence** dengan mencari banyaknya kemunculan item a dan b pada nota

Contoh
Gula kopi
Gula+kopi ada 5 sementara gula saja ada 8 jadi 5/8=0,62 atau 62%

no	barang	conf
1.	Gula, Kopi	5/8 = 62 %
2.	Gula, Susu	4/8 = 50 %
3.	Sikat gigi, Pasta Gigi	2/3 = 67 %

Selanjutnya nilai suppoer dikali dengan nilai confidence
Gula kopi= 25% x 62% = 15%

no	barang	final
1.	Gula, Kopi	15 %
2.	Gula, Susu	10 %
3.	Sikat gigi, Pasta gigi	6 %

Kesimpulan: Jika user memasukkan nilai minimal adalah 10% maka kombinasi yang memenuhi adalah
kombinasi gula, kopi, dan gula susu. Atau
hasil perhitungan diatas dapat diartikan 15% orang yang membeli gula juga akan membeli kopi, 10% orang membeli gula juga akan membeli susu.

Istilah:
sum of squared-error:jumlah dari kesalahan kuadrat
unsupervised learning: algoritma *machine learning* yang bertujuan untuk menarik kesimpulan dari kumpulan data input tanpa label respon.
Euclidean distance : jarak linier (garis lurus) antara dua titik pada bidang Euclidean atau koordinat Kartesian.
minPts: Jumlah minimum poin (ambang batas) dikelompokkan bersama-sama agar suatu wilayah dianggap padat.
eps (ε): Ukuran jarak yang akan digunakan untuk menemukan titik-titik di lingkungan titik mana pun.
Core: titik yang masuk ke eps dan minpts
Border: masuk ke dalam eps tapi tidak minpts
Noise: tidak masuk eps dan minpts
Itemset: himpunan dari item-item yang muncul bersama-sama
Support dari suatu itemset X (supp(X)) adalah rasio dari jumlah transaksi dimana itemset muncul dengan total jumlah transaksi
MAE (**Mean Absolute Error**) adalah rata-rata selisih mutlak nilai sebenarnya (aktual) dengan nilai prediksi (peramalan).
MAPE (*Mean Absolute Percentage Error*) adalah alat statistik yang digunakan untuk mengukur keakuratan suatu model statistik dalam melakukan prediksi atau peramalan.
RMSE menghitung rata-rata dari selisih kuadrat antara nilai prediksi dan nilai aktual kemudian diambil akar kuadratnya.
Confusion Matrix adalah **pengukuran performa untuk masalah klasifikasi machine learning dimana keluaran dapat berupa dua kelas atau lebih**.
Nilai **True Negative** (TN) merupakan jumlah data negatif yang terdeteksi dengan benar, sedangkan **False Positive** (FP) merupakan **data negatif namun terdeteksi sebagai data positif**. Sementara itu, **True Positive** (TP) merupakan data positif yang terdeteksi benar.
Recall atau sensitivity: **menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi**.
F1 Score F1 Score merupakan **perbandingan rata-rata presisi dan recall yang dibobotkan**
Silhouette Coefficient **digunakan untuk mengevaluasi seberapa baik suatu objek dikelompokkan ke dalam sebuah klaster**.
Cross-validation adalah **teknik untuk menguji kinerja model pada data yang tidak terlihat sebelumnya, mencegah overfitting, dan memilih model terbaik**.
Maximal_depth : Parameter ini digunakan untuk **membatasi kedalaman pohon keputusan**.