Bab 7
Korelasi, Pembersihan, dan Normalisasi Data

Muhammad Imam Dinata, S.Kom.,M.T

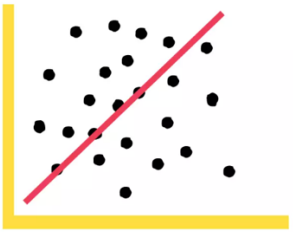
a) Korelasi

Korelasi adalah ukuran statistik untuk mengetahui sejauh mana hubungan antara dua variabel atau lebih. Korelasi mencerminkan kekuatan maupun hubungan antara variabel-variabel. Dalam statistika, korelasi digunakan untuk mendeskripsikan hubungan sederhana antara variabel tanpa mempertanya-kan sebab-akibatnya. Untuk merangkum hubungan variabel dalam satu angka, korelasi menggunakan metode yang disebut koefisien korelasi. Koefisien korelasi biasanya dilambangkan dengan simbol r dan bernilai antara -1 hingga +1.

Nilai atau Hasil Korelasi

- Nilai positif

Nilai positif terjadi ketika nilai korelasi lebih besar dari 0. Ini menandakan kedua variabel yang dibandingkan memiliki hubungan positif sempurna. Ketika satu variabel bergerak lebih tinggi atau lebih rendah, variabel lain bergerak ke arah yang sama dan besaran nilainya juga sama.

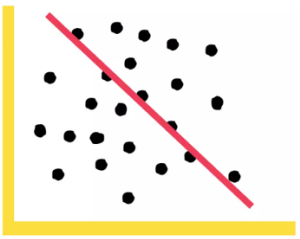


Nilai positif

PEMBELAJARAN DARING KOLABORATIF
KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET DAN TEKNOLOGI.
DIREKTORAT JENDERAL PENDIDIKAN TINGGI, RISET DAN TEKNOLOGI.
DIREKTORAT PEMBELAJARAN DAN KEMAHASISWAAN

- Nilai negatif

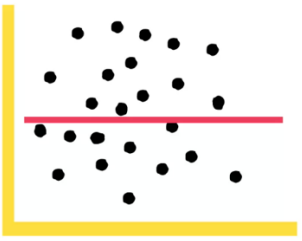
Berkebalikan dengan nilai positif, nilai negatif terjadi saat korelasi kurang dari 0. Nilai ini mengindikasikan kedua variabel bergerak berlawanan arah. Jika satu variabel meningkat, variabel lain justru menurun dengan nilai yang sama.



Nilai negatif

- Nilai nol (tidak ada korelasi)

Nilai nol mengartikan tidak adanya korelasi antar variabel. Masing-masing variabel bergerak sendiri-sendiri dengan sifat berbeda**.**

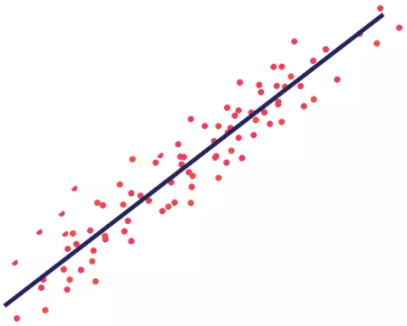


Nilai nol

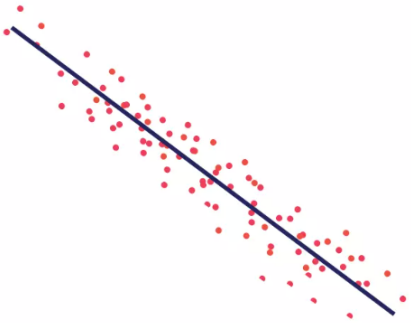
Ada beberapa metode yang biasa digunakan untuk melakukan korelasi pada sebuah data diantaranya, yaitu:

a. **Korelasi Pearson**

Korelasi Pearson adalah pengukuran yang paling umum digunakan. Korelasi jenis ini bertujuan mengukur hubungan linier antara variabel X dan variabel Y. Kedua variabel meningkat dan menurun secara bersamaan secara konstan dan dapat dimodelkan dengan garis lurus seperti gambar di bawah.



Hubungan linier positif yang kuat



Hubungan linier negatif yang kuat

Semakin jauh nilainya dari 0, semakin kuat hubungannya. Nilai 1 menandakan hubungan linier positif, yaitu ketika satu variabel bertambah dan variabel lainnya juga bertambah. Sementara -1 mewakili hubungan linier negatif, yaitu ketika satu variabel menurun sementara variabel lain meningkat.

b. Korelasi Spearman

Korelasi Spearman adalah korelasi yang digunakan untuk mengukur hubungan monotonic (tidak selalu linear) antara dua variabel ordinal atau interval. Metode ini digunakan untuk mengidentifikasi apakah ada hubungan searah atau sebalik antara dua variabel. Metode ini sering digunakan ketika data tidak memenuhi asumsi untuk korelasi Pearson (korelasi linear) atau ketika kita ingin menilai hubungan antara peringkat atau urutan.

Korelasi Spearman menghasilkan nilai antara -1 dan 1. Nilai 1 menunjukkan hubungan sempurna searah, sementara -1 menunjukkan hubungan sempurna berlawanan arah. Nilai 0 menunjukkan tidak adanya hubungan monotonic antara variabel-variabel tersebut.

Keuntungan dari korelasi Spearman adalah bahwa metode ini lebih kuat daripada korelasi Pearson dalam mendeteksi hubungan yang mungkin tidak linear antara variabel. Hal ini juga dapat digunakan dengan baik pada data ordinal atau data yang memiliki asumsi non-normal.

Model yang meningkat secara monotonik

Model yang menurun secara monotonik

c. Korelasi Kendall

Korelasi Kendal merupakan korelasi metode statistik yang mirip dengan Korelasi Spearman. Metode ini digunakan untuk mengukur hubungan monotonic (tidak selalu linear) antara dua variabel yang memiliki data dalam **bentuk ordinal atau peringkat**. Korelasi Kendall lebih fokus pada peringkat atau urutan data daripada nilai sebenarnya dari variabel-variabel tersebut. Hasil korelasi Kendall juga berkisar antara -1 dan 1. Seperti Spearman, Kendall juga melihat hubungan monotonik antara variabel. Hal yang membedakan dengan Spearman terletak pada pengolahan nilai rank yang diperoleh dari variabel X dan Y.

Contoh korelasi Kendall :

tingkat pendidikan yang dapat dikelompokkan menjadi “pendidikan menengah”, “S1”, “S2”, dll.

b) Pembersihan Data

Pembersihan data adalah proses mempersiapkan dan membersihkan data mentah agar dapat diolah dan dianalisis dengan benar pada suatu data di dalam sebuah dataset. Tujuannya adalah untuk menghilangkan kesalahan, kebingungan, atau noise dalam dataset. Pada saat menggabungkan beberapa data sources sekaligus, ada kemungkinan data terduplikasi atau bahkan salah label. Situasi seperti ini juga memerlukan data cleaning agar tidak muncul masalah yang lebih rumit dan akan menyebabkan perubahan hasil pada dataset tersebut.

Mengapa diperlukan pembersihan dataset? Karena untuk menghindari data berkualitas buruk akan memberikan hasil dan algoritma yang tidak bisa dijamin kebenaranya meski proses analisisnya benar. Berikut adalah beberapa alasan mengapa data cleaning harus dilakukan:

c) Menghilangkan kesalahan dan inkonsistensi yang muncul saat beberapa data sources dikumpulkan dalam satu dataset.

d) Meningkatkan efisiensi kerja karena proses ini akan memudahkan Anda dan tim pengolah data untuk menemukan apa yang dibutuhkan dari data.

e) Tingkat error yang lebih rendah juga akan mendatangkan kepuasan pelanggan dan mengurangi beban kerja tim.

f) Membantu memetakan beberapa fungsi data yang berbeda. Proses ini juga akan membuat Anda lebih mengenal kegunaan data dan mempelajari asalnya.

Ada beberapa cara dalam melakukan pembersihan data, Agar pembersihan data dapat dilakukan secara menyeluruh, diantaranya yaitu:

1. Mendeteksi Error

   Langkah awal yang harus dilakukan adalah memantau notifikasi error atau corrupt. Ada baiknya Anda mencatat titik yang paling sering terjadi error.

2. Hapus Duplikat Data atau Data yang tidak diperlukan

   Perbaikan dan penghapusan berlaku untuk duplikat dan data yang dirasa tidak perlu. Untuk mencegah terjadinya duplikasi data, pada saat menggabungkan beberapa data sources.

3. Perbaiki Kesalahan Struktur

   penamaan yang aneh, typo, atau penggunaan simbol aneh saat sedang melakukan transfer data, Bisa jadi ada kesalahan struktur pada dataset. Hal ini akan menyebabkab timbulnya inkonsistensi pada data

4. Filter Outlier yang tidak diinginkan

   Pada dataset, terkadang muncul data yang sekilas tampak tidak sesuai atau terpaut jauh dengan data lain. Inilah yang disebut dengan outlier atau pencilan. Penyaringan outlier memang bisa membantu performa data yang sedang dikerjakan. Meski begitu, perlu diingat bahwa kemunculan outlier bukan berarti pengolahan dataset anda salah. Justru sebaliknya, adanya outlier bisa menjadi indikator untuk menentukan validitas data.

5. Menangani Data yang hilang

   - Cara pertama, masukkan nilai yang hilang berdasarkan observasi lain. Cara ini sangat riskan karena mengandalkan asumsi, yang mana bisa mengancam integritas data.

- Cara kedua, buang observasi dengan nilai yang hilang. Namun, langkah ini bisa membuat Anda kehilangan informasi penting.
- Cara terakhir, mengubah bagaimana cara data digunakan agar nilai yang kosong dapat dinavigasikan dengan efektif.

6. Validasi dan Melakukan QA

   Melakukan validasi dan QA (quality assurance). harus bisa memastikan bahwa data bisa diterima dan memang masuk akal. Selain itu, data juga harus sesuai dengan aturan yang ada.

c) **Normalisasi Data**

Normalisasi dalam konteks big data mengacu pada proses mengatur dan menyelaraskan data agar dapat digunakan dengan lebih efisien dalam analisis dan pemrosesan. Tujuan normalisasi adalah untuk menghilangkan ketidakseimbangan, anomali, atau inkonsistensi dalam data yang dapat mengganggu analisis dan pemahaman data. ormalisasi juga dapat membantu dalam mengintegrasikan data dari berbagai sumber yang mungkin memiliki format dan struktur yang berbeda. Ada beberapa pengolahan yang dilakukan pada proses normalisasi data, yaitu

a) Format Data yang Konsisten: Data dari berbagai sumber mungkin memiliki format yang berbeda. Normalisasi mencakup penyelarasan format data sehingga semuanya konsisten, yang mempermudah pengolahan dan analisis data.
b) Pembersihan Data: Normalisasi melibatkan pembersihan data dari noise, nilai yang hilang, dan outlier. Ini membantu memastikan bahwa data yang digunakan dalam analisis adalah data yang valid.
c) Pengurangan Redundansi: Data redundan, yaitu data yang diulang dalam berbagai bentuk atau lokasi, dapat memakan ruang penyimpanan yang berharga. Normalisasi dapat membantu mengurangi redundansi dan menghemat penyimpanan.
d) Skala yang Seragam: Terkadang, data yang memiliki skala yang berbeda perlu dinormalisasi sehingga memiliki skala yang seragam. Hal ini berguna ketika Anda ingin membandingkan atau menggabungkan data dari berbagai sumber.
e) Integrasi Data: Big data sering berasal dari berbagai sumber. Normalisasi memungkinkan untuk mengintegrasikan data ini ke dalam satu sumber data tunggal yang terstruktur.
f) Pengelolaan Metadata: Normalisasi juga dapat mencakup manajemen metadata, yaitu informasi

tambahan tentang data seperti penjelasan, sumber, dan atribut lainnya.

g) Optimasi Kueri dan Pemrosesan: Normalisasi dapat memungkinkan kueri dan pemrosesan data yang lebih efisien, terutama dalam konteks pengolahan data besar. Data yang sudah dinormalisasi akan lebih mudah diakses dan dianalisis.

Perlu untuk diingat bahwa normalisasi pada big data harus dipertimbangkan dengan hati-hati, dan pendekatan yang digunakan akan bervariasi tergantung pada tipe data, tujuan analisis, dan infrastruktur yang digunakan dalam pengelolaan big data. Teknik normalisasi yang tepat akan membantu memaksimalkan nilai informasi yang dapat diperoleh dari big data dan mengurangi hambatan dalam analisisnya.

KESIMPULAN

Pada proses Korelasi, pembersihan, dan normalisasi data adalah tahapan penting dalam analisis data yang diperlukan karena masing-masing memberikan manfaat khusus dalam memproses dan memahami data pada sebuah big data. Sehingga, ketiga langkah ini saling melengkapi dan mendukung pengolahan dan analisis data yang baik. Korelasi membantu pada saat memahami hubungan antara variabel, pembersihan data memastikan data yang dianalisis adalah data yang berkualitas, dan normalisasi memudahkan pemrosesan data dalam skala besar. Dalam kombinasi, mereka memungkinkan kita untuk mendapatkan wawasan yang lebih baik dari data, membuat keputusan yang lebih informasi, dan meningkatkan kualitas analisis data pada sebuah big data