Discretization: Trik Penting dalam Analitik dan Pemodelan

Discretization seringkali membantu dalam proses eksplorasi dan visualisasi. Tanpa *discretization*, *data scientist* bisa sulit melihat pola-pola umum pada data akibat banyaknya *noise*.

PENULIS

BAGUS SARTONO
Dosen di Departemen Statistika IPB dan Wakil Ketua FORSTAT

YUNANTO PUTRANTO
Data Scientist di DataLabs Analytics

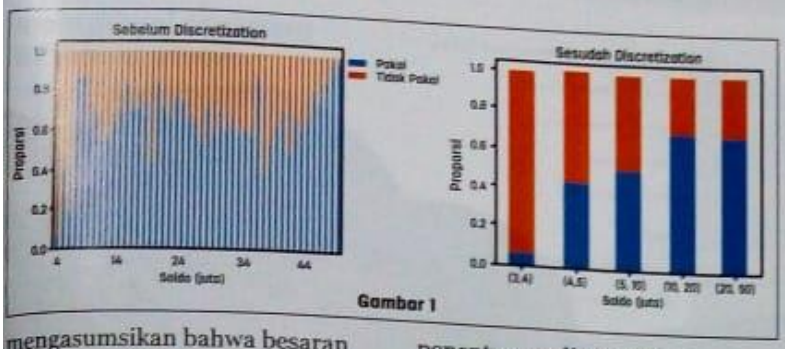
MEMPERSIAPKAN data merupakan tahapan yang berkontribusi besar dalam pembuatan *predictive modeling*. Salah satu prosesnya adalah penyiapan data variabel prediktor, yang banyak juga dikenal sebagai proses *feature engineering*. Termasuk di dalamnya adalah teknik-teknik reduksi banyaknya variabel, transformasi, penanganan data hilang, serta *discretization*. Fokus dari tulisan ini adalah tentang *discretization*.

Secara umum, *discretization* (disering disebut dengan *diskretisasi* oleh orang Indonesia) bekerja dengan mengubah variabel numerik (kontinu) menjadi variabel baru yang nilainya berupa selang-selang nilai asal yang tidak tumpang tindih. Meskipun tidak ada ketentuan khusus, umumnya nilai selang yang dibentuk sebanyak empat sampai sepu buah. Karena prosesnya seper membagi-bagi nilai data ke dalam selang-selang nilai, ad yang menamakannya sebagai proses *binning*. Sementara itu dalam pembicaraan tentang sk pengukuran di statistika, pros ini mengubah variabel numeri menjadi variabel kategorik, sehingga disebut juga proses kategorisasi.

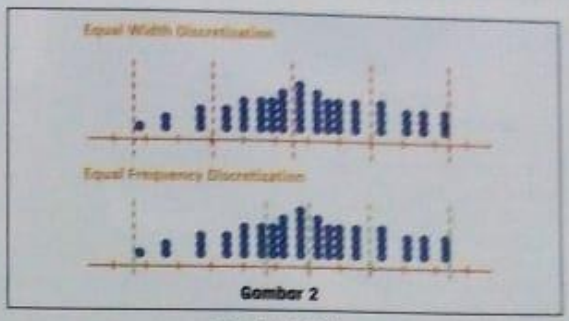
Apa keuntungannya?

Terdapat beberapa hal yang menjadi alasan mengapa *discretization* ini perlu dikerjakan pada proses analitik dan pemodelan. Pertama, *discretization* seringkali membantu dalam proses eksplorasi dan visualisasi. Tanpa ada *discretization*, terlihat banyak *noise* yang menganggu sehingga pola-pola umum pada data menjadi tidak begitu terlihat. Sebagai ilustrasi, grafik pertama pada **Gambar 1** menunjukkan proporsi pengguna *internet banking* untuk berbagai nilai saldo nasabah dibulatkan pada nilai jutaan terdekat. Sementara grafik kedua adalah proporsi pengguna setelah proses *discretization*. Bisa disebutkan bahwa kecenderungan penggunaan *internet banking* lebih besar pada nasabah dengan saldo besar jadi lebih mudah terlihat pada grafik kedua.

Keuntungan lain dari dilakukannya *discretization* ini adalah mengakomodasi pola-pola yang tidak serupa dengan pola model parametrik. Sebut saja misalnya dalam pemodelan regresi logistik. Model ini



Gambar 1



Gambar 2

mengasumsikan bahwa besaran peluang terjadinya suatu kejadian cenderung bersifat monoton naik (atau turun) seiring dengan perubahan nilai dari variabel prediktor. Memaksakan penggunaan regresi logistik pada kasus data dengan pola peluang kejadian yang tidak seperti bentuk model tersebut akan menyebabkan dugaan model yang diperoleh memiliki ukuran pengepasan yang buruk. Akibatnya, didapatkan prediksi dengan akurasi rendah. *Discretization* variabel prediktor numerik dan melakukan pemodelan menggunakan variabel prediktor baru yang bersifat kategorik akan banyak membantu menghasilkan dugaan model yang lebih baik.

Tidak hanya dapat meningkatkan kebaikan dugaan model, *discretization* juga berguna dalam mengatasi keberadaan data pencilan (*outlier*) dan data hilang (*missing*). Data pencilan yang nilainya ekstrem akan dimasukkan ke dalam salah satu selang yang sama yaitu selang paling rendah atau paling tinggi bersama nilai-nilai lainnya sehingga tidak lagi ekstrem. Sementara data hilang, umumnya ditangani dengan membuat kategori/*bin* tersendiri.

Selain hal-hal di atas, *discretization* juga diperlukan dalam pemodelan untuk mempercepat proses komputasi. Misalnya saja pada tahapan penentuan *splitting point* pada pemodelan *classification tree*, prosesnya akan lama jika variabel prediktornya bersifat numerik kontinu dan algoritma pencariannya adalah *greedy search*. Waktu komputasi yang lama ini dikarenakan kandidat titik pemisahnya akan sangat banyak. Penggunaan variabel prediktor yang sudah mengalami *discretization* akan membuat proses identifikasi ini menjadi jauh lebih cepat, meskipun tentu saja ada pengorbanan pada beberapa detail tertentu yang umumnya masih bisa ditoleransi. Algoritma prediktif lain yang juga banyak terbantu dengan adanya *discretization* adalah Bayesian classifier.

Bagaimana melakukannya?

Berdasarkan cara pembentukan *bin*, metode *discretization* dikelompokkan menjadi dua yaitu metode *splitting* dan metode *merging*. Cara yang pertama adalah memandang semua nilai asal berada pada satu selang nilai yang sangat lebar kemudian dipisah-pisah menjadi selang-selang yang lebih kecil. Cara yang kedua bekerja sebaliknya dengan membuat selang-selang yang super kecil yang hanya memuat satu macam nilai dan kemudian melakukan penggabungan terhadap selang-selang yang bersebelahan. Sementara itu, jika dilihat dari aspek keterlibatan variabel lain (biasanya variabel target), metode *discretization* terbagi ke dalam pendekatan *supervised* dan *unsupervised method*.

Ada beberapa metode yang populer digunakan di banyak kesempatan oleh para *data scientist*. Dua di antaranya adalah *equal range discretization* dan *equal frequency discretization*. Metode *equal range discretization* membuat *bin* dalam bentuk selang-selang sama lebar yang memuat semua nilai pada data. Adapun metode *equal frequency discretization* membentuk selang-selang sedemikian rupa sehingga banyaknya amatan di setiap selang relatif sama. **Gambar 2** menyajikan ilustrasi perbedaan hasil kedua metode tersebut.

Kedua metode ini dapat disebut sebagai metode yang paling sederhana dan tergolong bersifat *unsupervised*. Meskipun sederhana, di berbagai terapan mampu memberikan hasil yang memuaskan. Pendekatan populer lain dari kelompok *supervised* antara lain adalah MDLP (Minimum Description Length Principle) dan ChiMerge dengan berbagai modifikasinya.

Penutup

Saat ini berbagai *software* komersial maupun *open source* sudah menyediakan berbagai fungsi, modul, atau prosedur melakukan *discretization*. Bahkan sebagian menyediakan proses otomatis yang sangat membantu pengguna yang tidak ingin terjebak pada kerumitan. Namun tentu saja *data scientist* yang andal akan berhati-hati dalam penggunaannya. Selamat mencoba.