BINUS UNIVERSITY ONLINE

People
Innovation
Excellence

Personal Assignment 2

1. Konsep *Exploratory Data Analysis* (EDA)

   EDA merupakan sebuah tahapan awal dalam proses analisis data yang bertujuan untuk memahami data, dari karakteristik, struktur, pola, dan isi dari data sebelum dilakukan pemodelan atau pengambilan Keputusan. Teknik pemodelan yang digunakan ini merupakan teknik statistic dan visualisasi, untuk mendeteksi berbagai asumsi data, dari :

   Memahami struktur data, menangani *missing values*, mendeteksi *outliers*, mencari korelasi antar variable, dan juga menilai kualitas data.

   EDA penting dalam *data science workflow* karena membantu untuk menemukan pola, struktur tanpa asumsi awal yang kaku, serta menemukan error input, duplikasi dan *missing value*. Serta menjadi landasan validasi awal sebelum model prediktif dibuat.

2. Manfaat beberapa teknik visualisasi

   - Histogram, membantu untuk memahami bentuk distribusi data. Apakah normal, skewed (berat sebelah).
   - Boxplot, boxplot ini untuk melihat nilai median atau pusat dari data, lalu kuartil, dan penyebaran data serta nilai *outliers*.
   - Scatter plot, membantu untuk mengidentifikasi pola hubungan antra variable, misalnya secara linear, non-linear, atau bahkan tidak ada hubungannya. Lalu dengan scatter kita bisa melihat arah dan kekuatan korelasi data, apakah positif, negative atau tidak signifikan.
   - Heatmap, sebelumnya di scatter plot kita bisa melihat kekuatan korelasi data, di heatmap pun hampir sama, tapi disini tingkat korelasi antar variablenya digambarkan dengan gradasi warna. Selain itu juga untuk menemukan pola global dan variablenya yang lebih banyak.

3. Contoh penggunaan EDA di dunia industry

   Saya akan ambil contohnya dari Tokopedia, untuk analisis perilaku pelanggan. Tujuannya untuk mengetahui segmen pelanggan berdasarkan perilaku belanja, lalu deteksi anomaly atau penurunan aktivitas pelanggan, dan juga menentuka waktu promosi yang paling efektif.

Langkah awal yang dilakukan adalah memahami struktur data yang ada, diambil dari log order, seperti cust id, produk id yang dibeli, tanggan purchase, sampai ke payment type. Setelah itu lanjut untuk melakukan analisis deskriptif awal untuk menemukan :

- Jumlah pelanggan aktif
- Mean transaksi per pelanggan
- Produk terlaris

Setelah data di dapat, akan keluar output speerti contoh, 70% pelanggan hanya berbelanja 5x dalam 5 bulan terakhir

Selanjutnya melakukan visualisasi data menggunakan Histogram, boxplot, scatter plot , hingga heatmap. Tujuan visualisasi ini tentunya untuk mengetahui distribusi data belanja pelanggan, apakah low spender atau high spender, lalu untuk mendeteksi apakah ada variasi harga yang ekstrim, atau outlier. Serta melihat korelasi antar frekuensi pembelian, total pengeluaran, dan loyalitas pelanggan.