Kelompok: Kelompok 5 - Anaconda
Stage: 2
Mentor: Fiqry R
Pukul/ Tanggal: 12 Juni 2022, 14.00 WIB

Pembagian tugas di stage 2:

1. Duplikasi, handling outlier & feature scaling: Irfan - tampilin data-data duplikasi.
2. Feature encoding: Ramado, Bima - Tambah lagi encodernya dan untuk month cari lagi variasi yang bisa dipake quarter
3. Handle imbalance: Irfan, Denindra
4. Feature selection: Denindra - mencoba menggunakan pandas profiling
5. Feature extraction: Irfan
6. Additional features: Ni Putu Tasya
7. Slide presentasi: semuanya

Poin Pembahasan:

- Data Preprocessing
- Feature Engineering

Tindak Lanjut:

1. **Pastikan data duplicated sudah di reset index**
2. **Tampilkan di slide contoh duplicated data nya**
3. **Revisi Feature imbalance dan Feature encoding apabila perlu**
4. **Slide 5 tampilkan kode-nya.**
5. **Untuk feature selection, coba dulu pandas profiling correlation, lalu pertimbangkan lagi mana feature yang mau dibuang.**
6. **Pertimbangkan untuk buang Operating system, traffic type, browser, region dari awal, dari pada salah interpretasi atau asumsi.**

Kelompok: Kelompok 5 - Anaconda
Stage: 2
Mentor: Fiqry R
Pukul/ Tanggal: 12 Juni 2022, 14.00 WIB

Hasil Diskusi:

1. Untuk data duplikasi, tampilkan lima data yang duplikasi menggunakan attribute head().
2. Slide mengenai split dataset di-hide dan tambahkan informasi mengenai splitting saat tahap handling imbalance.
3. Gunakan pipeline di data training agar model mengikuti penyebaran data yang kita miliki preprocessing dulu baru resampling. kalo resampling dulu nanti datanya jadi tidak murni, tidak sesuai dengan kenyataan data yang ada
4. Data feature numerik semuanya right-skewed, sehingga sebenarnya gak perlu melakukan transformasi, scaling, atau normalisasi jika tidak menggunakan algoritma linear atau yang berbasis jarak. Hal ini juga bisa mempengaruhi interpretasi model.
5. Beberapa hal yg membuat kita gak perlu buang outlier:
   - dataset adalah sebenar-benarnya kondisi yang ada,
   - data yang sangat mahal (misal: data kedokteran – terkait dengan nyawa dll)
   - ada kondisi dimana kita bisa memilih model yang bisa melakukan training tanpa ada risiko ketika ada outliers. Contohnya menggunakan tree-based / ensemble model.
6. Bagian feature encoding: idealnya ketahui dulu tipe data nya itu nominal atau ordinal. Pertimbangan :
   - banyaknya unique value (missal bentuknya 0 dan 1 → pilih one-hot encoding)
   - ordinal encoding→ feature harus dianggap sebagai angka yang punya order.
7. Sebenarnya gak perlu analisis korelasi antara x dan y untuk menentukan feature mana yang harus dibuang. Pembuktiannya itu saat di interpretasi model.
8. Metrics utama ada yang mengukur kualitas model dalam memprediksi data, ada juga yang mengukur bagaimana model bisa bekerja sebagaimana mestinya (contoh: ROC AUC score)