Nama Kelompok : Pegasus
Dataset : Banking - Marketings Target
Stage : 2

1. **Data Cleansing**
    a. **Handle Missing Values:**
        i. Missing Values treatment tidak dilakukan karena dataset telah bersih dari missing value.
        ii. Telah mereplace nilai ‘unknown’ pada kolom job, education, dan contact menjadi nilai modus pada masing-masing kolom

        Untuk kolom balance terdapat dua opsi yang dapat dilakukan
        iii. Merubah nilai minus pada kolom balance menjadi nilai nol
        iv. Tidak merubah nilai minus pada kolom balance menjadi nilai nol

    b. **Handle Duplicated Data:**
        i. Terdapat 4163 baris data duplikat berdasarkan kolom age, job, marital, education, default, balance, balance, housing, dan loan yang telah dilakukan duplicate treatment(drop)

    c. **Handle Outliers**
        I. Handling outlier menggunakan teknik Z-Score dikarenakan hanya ingin mengikis data secara sedikit, tidak berlebih dikarenakan presentasinya yang hanya 0.3% (karena masih ingin meng-keep data yang kira2 berharga). Juga saat mencoba untuk menggunakan teknik IQR dan mengalami kendala karena banyak data yang hilang, khususnya di data kategorikal. Untuk ke depannya, alangkah lebih baik jika setiap data dilakukan treatment yang berbeda berdasarkan cara menghapus outliernya (data yang berdistribusi mendekati normal menggunakan Z-score, sedangkan data yang banyak terdapat outlier bisa menggunakan IQR)

    d. **Feature Transformations**
        i. Untuk kolom Day dan Age sementara dipilih menggunakan Normalisasi, sedangkan lainnya menggunakan Standarisasi karena yang lain terlihat di boxplot banyak sekali terdapat outlier.
        ii. Berdasarkan boxplot:
            - Kolom Age masih memiliki sedikit outlier
            - Kolom Day Tidak memiliki outlier (sejak awal) ->karena nilai maksimalnya sudah pasti
            - Kolom Balance, Duration, Pdays, Campaign dan Previous memiliki banyak outlier (pdays & previous = extreme)
        iii. Berdasarkan histogram:
            - Kebanyakan grafik masih bersifat skewed kecuali kolom Day dan age yang mendekati normal

Untuk tahapan transformasi ini sebenarnya bisa menyesuaikan dengan model Machine Learning yang akan dipilih nanti, tidak semua harus ditransformasi/harus menggunakan teknik tertentu (pemilihan teknik dapat bervariasi menyesuaikan syarat model) . Nantinya mungkin tahap ini akan diterka lagi.

e. Feature Encoding

Encoding dilakukan dengan menggunakan dua cara, yaitu:

i. Label Encoding
   Teknik Encoding yang mengganti data kategorikal dalam satu kolom menjadi numerik berurut, tanpa membuat kolom baru.
   Pada kasus dataset ini, diterapkan pada kolom default, housing, loan, dan kolom label (y).

ii. One Hot Encoding
   Teknik Encoding yang digunakan untuk data nominal, yaitu data yang tidak bisa dibandingkan nilainya, seperti Agama, Pekerjaan, dll.
   Pada kasus ini, teknik One Hot Encoding diterapkan pada kolom job, marital, education, contact, mount, dan poutcome.

f. Imbalance Dataset

Imbalance dataset dilakukan untuk mengatasi masalah timpangnya jumlah value yang sedikit dan value yang dominan pada klasifikasi. Jika hal ini terjadi, maka akan memungkinkan model untuk tidak dapat mengenali data berdasarkan klasifikasinya, model hanya akan menganggap value dominan. Adapun cara yang dilakukan untuk mengatasi imbalance dataset pada kasus ini, dilakukan dengan:

i. Teknik under sampling
   Teknik sampling dengan cara mengurangi jumlah data dominan.

ii. Teknik Over Sampling
   Teknik sampling dengan cara meningkatkan jumlah data minor.

iii. Teknik SMOTE
   Teknik sampling dengan mensintesis sample baru dari data minor untuk menyeimbangkan data.

Setelah menerapkan tiga metode imbalance ke dalam dataset maka selanjutnya bisa dianalisa metode apa yang menghasilkan performa paling baik untuk model.

2. Feature Engineering

a. Feature Selection

Hasil dari analisa diatas dipilih beberapa feature yang kemungkinan paling bermanfaat:
Feature reduction: categorical: job, marital, education (kurang ada relasi dengan target), default (mirip dengan loan), month, poutcome numerical: nums = pdays (ada nilai -1) dan lebih dari value -1 ada lebih dari 75%. .
Jadi fitur yang akan diproses ke stage selanjutnya adalah:

i. Optional feature 1: housing, loan, contact, y , age, balance, duration, campaign, previous, day.

ii. Optional feature 2: housing, loan, contact, y , age, duration, campaign

iii. Optional feature 3: *kemungkinan akan ada perubahan fitur yg dipilih/diuji lagi di next stage.

b. Feature Extraction

Tidak ada fitur yang perlu diekstrak, karena setiap fitur sudah cukup fungsional dengan data aslinya, mungkin beberapa perlu di encoding.

c. New Feature Idea

Ide fitur tambahan: mempunyai anak/tanggungan berapa, gaji, gagal bayar, status pegawai tetap, saran terhadap fitur 'job' untuk diisi dengan : PNS/Pengusaha/Karyawan/Frelencer (agar segmentasinya agar tidak terlalu lebih banyak).