Nama: Muhamad Wildan

NIM: 231011403558

Kelas: 05 TPLE 004

Mata Kuliah: Machine Learning

**1. Deskripsi Dataset**

Dataset yang digunakan berasal dari **Bank Marketing Dataset** (UCI Repository / Kaggle). Dataset ini berisi data hasil kampanye pemasaran melalui panggilan telepon yang dilakukan oleh sebuah bank di Portugal. Tujuan utama dari data ini adalah untuk memprediksi apakah seorang nasabah akan **berlangganan produk deposito berjangka** (y = yes atau no).

Dataset terdiri dari **41.188 baris** dan **21 kolom**, yang terbagi menjadi variabel numerik dan kategorikal.
Beberapa fitur penting di antaranya:

- age: usia nasabah
- job: jenis pekerjaan
- marital: status pernikahan
- education: tingkat pendidikan
- balance: saldo rata-rata tahunan
- housing: apakah memiliki pinjaman rumah
- loan: apakah memiliki pinjaman pribadi
- contact: jenis kontak komunikasi (telepon, seluler, dsb)
- duration: lama panggilan terakhir (detik)
- campaign: jumlah kontak selama kampanye
- pdays: jumlah hari sejak terakhir dihubungi
- previous: jumlah kontak sebelumnya
- y: **target**, berisi nilai “yes” jika nasabah berlangganan deposito, “no” jika tidak.

Dataset ini bersifat **tidak seimbang (imbalanced)** karena sebagian besar nasabah tidak berlangganan produk bank, sehingga evaluasi model perlu memperhatikan **precision, recall, dan F1-score**, bukan hanya akurasi.

**2. Model yang Digunakan**

Dua algoritma klasifikasi digunakan dalam penelitian ini:

**a. Logistic Regression**

Merupakan model linear sederhana yang digunakan untuk memprediksi probabilitas dari suatu kelas biner. Logistic Regression cocok digunakan pada dataset besar dan memberikan interpretasi yang mudah terhadap pengaruh tiap variabel.

**b. Decision Tree Classifier**

Model berbasis pohon keputusan yang membagi data ke dalam cabang-cabang berdasarkan fitur yang paling berpengaruh terhadap target. Decision Tree dapat menangani data non-linear dan fitur kategorikal dengan baik, namun cenderung overfitting jika tidak dilakukan pruning.

Sebelum pelatihan, dilakukan preprocessing berupa:

- **Encoding variabel kategorikal** menggunakan OneHotEncoder.
- **Normalisasi fitur numerik** menggunakan StandardScaler.
- **Pembagian dataset** menjadi data latih dan data uji dengan proporsi 80:20.

---

**3. Hasil Evaluasi dan Pembahasan**

Model dievaluasi menggunakan **Confusion Matrix**, serta metrik **Accuracy, Precision, Recall, dan F1-Score**.

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Logistic Regression | 0.90 | 0.75 | 0.48 | 0.58 |
| Decision Tree | 0.88 | 0.62 | 0.60 | 0.61 |

**Analisis:**

- **Logistic Regression** memiliki akurasi lebih tinggi karena cenderung memprediksi mayoritas kelas (“no”).
  Namun, **recall-nya lebih rendah**, artinya model kurang baik dalam mendeteksi nasabah yang benar-benar tertarik berlangganan.
- **Decision Tree** sedikit lebih rendah akurasinya, tapi lebih seimbang antara precision dan recall — artinya lebih baik dalam mendeteksi pelanggan potensial.
- Berdasarkan **F1-Score**, Decision Tree menunjukkan performa yang lebih baik secara keseluruhan.
- ROC Curve menunjukkan area di bawah kurva (**AUC**) sebesar ±0.79 untuk Logistic Regression dan ±0.82 untuk Decision Tree.

**Kesimpulan:**

Model **Decision Tree** lebih direkomendasikan untuk kasus ini karena mampu menangkap lebih banyak calon pelanggan potensial tanpa terlalu banyak kesalahan klasifikasi, meskipun Logistic Regression memberikan akurasi yang sedikit lebih tinggi.