Laporan Analisis dan Pemodelan Classification

Nama : Satya Athaya Daniswara

NIM : 1103213152

1. Pendahuluan

Dataset yang digunakan bertujuan untuk memprediksi apakah seorang pelanggan akan berlangganan deposito berjangka (`y`). Dataset ini terdiri dari berbagai fitur numerik dan kategorikal yang mencakup informasi demografis, keuangan, dan interaksi pemasaran. Proses analisis dan pemodelan dibagi menjadi beberapa tahap: eksplorasi data, visualisasi, preprocessing, pelatihan model, dan evaluasi.

2. Eksplorasi Data

- Inisialisasi File:Dataset dimuat menggunakan `pandas` untuk mempersiapkan data dalam format tabular.

- Distribusi Variabel Target (`y`):

- Hasil: Mayoritas pelanggan tidak berlangganan deposito berjangka (`no`), menunjukkan ketidakseimbangan data yang signifikan.

- Implikasi: Perlu pendekatan khusus untuk menangani ketidakseimbangan, seperti penyesuaian evaluasi atau teknik sampling.

- Statistik Deskriptif: Distribusi numerik menunjukkan bahwa beberapa fitur, seperti `balance` dan `duration`, memiliki rentang nilai yang sangat lebar.

3. Visualisasi Data

- Distribusi Umur Berdasarkan Status Berlangganan:

- Pelanggan yang berlangganan cenderung berusia 30-60 tahun.

- Boxplot Saldo Berdasarkan Status Berlangganan:

- Pelanggan yang berlangganan memiliki saldo rata-rata lebih tinggi.

- Scatter Plot Umur vs Saldo:

- Pelanggan dengan saldo tinggi memiliki kemungkinan lebih besar untuk berlangganan, tetapi tidak ada hubungan linier dengan umur.

- Distribusi Durasi Panggilan:

- Durasi panggilan lebih panjang terkait dengan peningkatan kemungkinan berlangganan.

- Pairplot Variabel Numerik:

- Beberapa hubungan fitur menunjukkan pengelompokan pelanggan yang berlangganan (`yes`), terutama pada fitur `duration`.

4. Preprocessing Data

- Transformasi Data:

- Fitur numerik diisi dengan median dan dinormalisasi menggunakan `StandardScaler`.

- Fitur kategorikal diisi dengan modus dan di-encode menggunakan `OneHotEncoder`.

- Pembagian Data:

- Dataset dibagi menjadi 80% data latih dan 20% data uji menggunakan stratifikasi untuk mempertahankan distribusi kelas.

5. Pemodelan

Tiga model klasifikasi diterapkan dengan pipeline preprocessing:

1. Logistic Regression:

  - Parameter terbaik: `C=10`.

  - Akurasi: 90%.

  - Logistic Regression efektif menangkap pola linear antara fitur dan target.

2.Decision Tree:

  - Parameter terbaik: `max_depth=5`, `min_samples_split=2`.

  - Akurasi: 90%.

  - Memberikan interpretabilitas tinggi, meskipun sedikit lebih rawan overfitting.

3.k-Nearest Neighbors (k-NN):

  - Parameter terbaik: `n_neighbors=7`, `weights='uniform'`.

  - Akurasi: 90%.

  - Performa stabil, tetapi lebih sensitif terhadap ukuran dataset dan outlier.

6. Evaluasi

- Laporan Klasifikasi:

- Precision dan recall untuk kelas minoritas (`yes`) cukup rendah (~35%), menunjukkan bahwa model cenderung bias terhadap kelas mayoritas.

- Confusion Matrix:

  - Model memiliki tingkat salah prediksi yang lebih tinggi untuk pelanggan yang berlangganan dibandingkan yang tidak.

- Fitur Penting:

  - `duration` adalah fitur yang paling signifikan dalam menentukan prediksi.

7. Insights Utama

- Fitur Penting:

- `duration`, `balance`, dan `age` adalah fitur utama yang memengaruhi prediksi.

- Ketidakseimbangan Data:

- Ketidakseimbangan kelas mengharuskan pendekatan evaluasi khusus, seperti penggunaan metrik seperti `F1-score` daripada akurasi murni.

-Strategi Pemasaran:

  - Interaksi lebih lama dengan pelanggan meningkatkan peluang keberhasilan pemasaran.

8. Kesimpulan

- Logistic Regression adalah model terbaik dalam studi ini dengan parameter regulasi optimal karena Performanya kompetitif dengan model lain (akurasi 90%), Kemampuan regulasi yang mencegah overfitting., Kesederhanaan dan interpretabilitas yang tinggi, Kemampuan untuk menangkap hubungan linier dalam data

- Ketidakseimbangan data tetap menjadi tantangan utama, yang dapat diperbaiki dengan teknik sampling atau algoritma berbasis resampling seperti SMOTE.

- Analisis memberikan wawasan yang dapat digunakan untuk merancang strategi pemasaran yang lebih efektif.