|  |  |
| --- | --- |
|  | **UNIVERSITAS TEKNOLOGI YOGYAKARTA**  **FAKULTAS SAINS DAN TEKNOLOGI**  **PROGRAM STUDI SAINS DATA PROGRAM SARJANA**  TUGAS BESAR  MATA KULIAH MANAJEMEN DAN INFRASTRUKTUR DATA |
| NAMA | Lathif Ramadhan |
| NIM | 5231811022 |
| KELAS | 23A |
|  | |
| Tugas besar dikumpulkan pada minggu ke-14 tanggal 31 Desember 2024 | |
| SOAL | |
| 1. Buat atau cari data set yang sesuai dengan keahlian atau tema/topik penelitian anda. 2. Buat deskripsi yang menjelaskan tentang data set tersebut 3. Carilah kemungkinan-kemungkinan metode atau cara pengolahan yang dapat diterapkan ke data set tersebut. 4. Setelah anda mendapatkan metode dan cara pengolahan yang paling optimal, lakukan pengolahan dataset tersebut, jelaskan langkah-langkah nya beserta gambar kalau ada, visualisasikan hasil pengolahan data set anda. 5. Ceritakan maksud dari hasil pengolahan dataset anda sehingga mampu memberikan pengetahuan yang bermanfaat dari data yang diolah. | |
|  | |
|  | |
|  | |
|  | |

**Tentang Dataset (Soal No. 1 & 2)**

Judul Dataset : **Bank Marketing Campaign**

Link Dataset : <https://www.kaggle.com/code/janiobachmann/bank-marketing-campaign-opening-a-term-deposit>

**Deskripsi Dataset Bank Marketing Campaign**

A screenshot of a computer

Description automatically generated

Dataset ini berfokus pada kampanye pemasaran yang dilakukan oleh sebuah bank untuk memprediksi apakah seorang pelanggan akan membuka deposito berjangka (term deposit) berdasarkan sejumlah variabel yang relevan. Kampanye ini melibatkan berbagai data demografis dan interaksi dengan pelanggan yang mencakup detail seperti usia, pekerjaan, status pernikahan, pendidikan, apakah pelanggan memiliki pinjaman rumah atau pribadi, serta jumlah kontak yang dilakukan selama kampanye.

**Fitur Utama:**

1. **Data Klien Bank**:
   * **Usia (age)**: Usia klien (numerik).
   * **Pekerjaan (job)**: Jenis pekerjaan klien, seperti 'admin.', 'blue-collar', 'entrepreneur', dll. (kategorikal).
   * **Status Pernikahan (marital)**: Status pernikahan klien ('divorced', 'married', 'single', dll.) (kategorikal).
   * **Pendidikan (education)**: Tingkat pendidikan klien ('primary', 'secondary', 'tertiary', dll.) (kategorikal).
   * **Default**: Apakah klien memiliki kredit macet ('no', 'yes', 'unknown') (kategorikal).
   * **Pinjaman Rumah (housing)**: Apakah klien memiliki pinjaman rumah ('no', 'yes', 'unknown') (kategorikal).
   * **Pinjaman Pribadi (loan)**: Apakah klien memiliki pinjaman pribadi ('no', 'yes', 'unknown') (kategorikal).
   * **Saldo (balance)**: Saldo akun klien (numerik).
2. **Terkait Kontak Terakhir dalam Kampanye**:
   * **Jenis Kontak (contact)**: Jenis komunikasi yang digunakan (misalnya, 'cellular', 'telephone') (kategorikal).
   * **Bulan (month)**: Bulan kontak terakhir dilakukan (kategorikal).
   * **Hari (day)**: Hari dalam minggu kontak terakhir dilakukan (kategorikal).
   * **Durasi (duration)**: Durasi kontak terakhir (numerik, detik).
3. **Fitur Lainnya**:
   * **Kampanye (campaign)**: Jumlah kontak yang dilakukan selama kampanye ini untuk klien tersebut (numerik).
   * **Pdays**: Jumlah hari sejak klien terakhir dihubungi dalam kampanye sebelumnya (numerik; 999 berarti klien belum pernah dihubungi sebelumnya).
   * **Sebelumnya (previous)**: Jumlah kontak yang dilakukan dalam kampanye sebelumnya (numerik).
   * **Hasil Kampanye Sebelumnya (poutcome)**: Hasil kampanye pemasaran sebelumnya ('failure', 'nonexistent', 'success') (kategorikal).

**Variabel Target:**

* **y**: Apakah klien membuka deposito berjangka ('yes', 'no').

**Tujuan Analisis:**

* Memprediksi apakah seorang pelanggan akan berlangganan deposito berjangka atau tidak berdasarkan variabel-variabel yang ada.
* Mengidentifikasi faktor-faktor yang paling berpengaruh terhadap keputusan pelanggan untuk berlangganan deposito berjangka.
* Memberikan insight kepada bank untuk mengoptimalkan strategi pemasaran mereka di masa mendatang.

**Ringkasan**

Dataset ini berasal dari UCI Machine Learning Repository dan berfokus pada analisis keberhasilan kampanye pemasaran bank dengan tujuan utama untuk memprediksi apakah pelanggan akan membuka deposito berdasarkan serangkaian variabel. Fitur-fitur dalam dataset mencakup demografi pelanggan, rincian kontak kampanye, serta hasil dari kampanye sebelumnya. Target output adalah keputusan pelanggan untuk membuka atau tidak membuka deposito berjangka. Data ini sangat berguna untuk aplikasi dalam pemasaran yang lebih terarah dan segmentasi pelanggan.

**Kemungkinan Metode atau Cara Pengolahan untuk Dataset Bank Marketing (Soal No. 3)**

1. **Pra-Pengolahan Data**: Sebelum melakukan analisis lebih lanjut, penting untuk membersihkan dataset terlebih dahulu. Hal ini termasuk menangani data yang hilang (misalnya, mengganti nilai yang hilang dengan rata-rata atau median) dan mengkonversi variabel kategorikal menjadi numerik menggunakan teknik seperti one-hot encoding.
2. **Eksplorasi Data (EDA)**: Dengan tujuan untuk memahami lebih dalam tentang dataset, kita bisa menggunakan teknik visualisasi seperti histogram, boxplot, atau scatter plot untuk melihat distribusi usia, saldo, dan durasi kontak. Ini akan membantu menemukan pola-pola penting yang dapat memengaruhi keputusan pelanggan dalam membuka deposito.
3. **Feature Engineering**:

* **Encoding Kategorikal**: Menggunakan **one-hot encoding** untuk variabel seperti job, marital, dan education.
* **Pembuatan Fitur Interaksi**: Menggabungkan variabel seperti age dan job untuk melihat pola dalam kelompok tertentu.
* **Binning**: Mengelompokkan variabel numerik seperti age dan balance dalam interval untuk memudahkan analisis.
* **Fitur Baru**: Membuat fitur baru seperti total jumlah interaksi (campaign + previous).

1. **Analisis Korelasi**: Menggunakan korelasi untuk melihat hubungan antar variabel yang ada, misalnya, apakah ada hubungan yang kuat antara durasi kontak dengan keputusan untuk membuka deposito. Ini penting karena variabel seperti durasi sering kali sangat berkaitan dengan hasil kampanye.
2. **Segmentasi Pelanggan**: Teknik seperti klastering (misalnya K-Means) dapat digunakan untuk mengelompokkan pelanggan berdasarkan karakteristik yang serupa, seperti usia, pekerjaan, dan status pernikahan. Dengan segmentasi ini, bank dapat menargetkan pelanggan dengan lebih tepat, meningkatkan efektivitas kampanye.
3. **Model Prediksi**: Setelah pemahaman yang cukup terhadap data, kita bisa mencoba membangun model prediksi untuk memprediksi kemungkinan pelanggan membuka deposito. Beberapa model yang bisa diterapkan termasuk:
   * **Logistic Regression**: Untuk memprediksi variabel biner (ya/tidak) seperti apakah pelanggan membuka deposito.
   * **Decision Trees**: Untuk membuat keputusan berdasarkan serangkaian pertanyaan terkait variabel input.
   * **Random Forests**: Menggunakan banyak pohon keputusan untuk meningkatkan akurasi model.
4. **Evaluasi Model**: Setelah model dibangun, kita perlu mengevaluasi kinerjanya menggunakan metrik seperti akurasi, precision, recall, dan confusion matrix. Ini membantu dalam menilai seberapa baik model dalam memprediksi keputusan pelanggan yang tepat.

Dengan pendekatan yang sistematis ini, kita dapat memperoleh wawasan yang berguna untuk meningkatkan kampanye pemasaran dan pengambilan keputusan yang lebih baik di masa mendatang.