Rakamin Academy

Homework
Data Pre-Processing

Final Project

Estimasi Waktu Pengerjaan

3 - 5 jam

Jumlah Soal

2 Soal

Total Point

100 poin

Teknis Pengerjaan

1. Pekerjaan dilakukan secara **berkelompok, sesuai kelompok Final Project**

2. Masing-masing anggota kelompok tetap perlu submit ke LMS (jadi bukan perwakilan)

3. File yang perlu dikumpulkan:
   - File **jupyter notebook** (.ipynb) yang berisi source code.
   - File **laporan homework** (.pdf) yang berisi rangkuman dari apa saja yang telah dilakukan.

4. Upload hasil pengerjaanmu melalui LMS.
   - Masukkan semua file ke dalam **1 file** dengan format **ZIP**.
   - Nama File:

     `Preprocessing - <Nama Kelompok>.zip`

1. Data Cleansing (50 poin)

Lakukan pembersihan data, sesuai yang diajarkan di kelas, seperti:

A. Handle missing values

Sebelum melakukan handle missing values, pertama-tama kami memisahkan data berdasarkan kaetgori yaitu data numerikal dan kategorikal. Setelah itu kami baru melakukan pemeriksaan apakah ada missing value dalam data atau tidak. Ditemukan beberapa missing value antara lain :

Kolom	Missing Value
Age	226
TypeOfContact	25
DurationOfPitch	251
NumberOfFollowUps	45
PreferredPropertyStar	26
NumberOfTrips	140
NumberOfChildrenVisiting	66
MonthlyIncome	233

1. Data Cleansing (50 poin)

Dari beberapa data yang memiliki value kosong, maka kami memutuskan untuk mengisi data tersebut sesuai dengan karakteristik dari value tersebut. Kami memilih untuk mempertahankan data (tidak menghapus) dengan pertimbangan bahwa dengan melakukan pengisian data maka dapat menghindari kehilangan informasi berharga, mempertahankan akurasi data, dan menghindari bias data.

Terdapat beberapa kolom yang menggunakan median dikarenakan terdapat distribusi data yang tidak normal seperti Age, DurationOfPitch, NumberOfFollowups. Sedangkan untuk data yang bersifat kategorikal dapat diisi dengan nilai terbanyak/mode.

Setelah melakukan pengisian data kami memastikan kembali dengan memeriksa apakah masih ada yang tergolong missing value. Setelah semua data sudah terisi maka kami akan memasuki tahap handling duplikasi data.

1. Data Cleansing (50 poin)

B. Handle duplicated data

Pada tahap ini kami memeriksa duplikasi semua data dan hasilnya adalah terdapat 141 data duplikasi. Namun jika tidak menyertakan kolom CustomerID maka tidak ada duplikasi data yang terjadi. Sehingga kami memutuskan untuk menghapus baris duplikasi yang terdapat saat kami mencamtumkan kolom CustomerID dengan pertimbangan bahwa hal tersebut dapat membantu untuk meningkatkan keaslian, akurasi, dan keandalan analisis data selanjutnya.

```
[ ] df.duplicated(subset=df.columns.difference(['CustomerID'])).sum()
```

```
141
```

```
[ ] df.drop_duplicates(subset=df.columns.difference(['CustomerID']), inplace=True)
```

```
[ ] df.duplicated(subset=df.columns.difference(['CustomerID'])).sum()
```

```
0
```

1. Data Cleansing (50 poin)

C. Handle outliers

Pada tahap ini kami melakukan visualisasi dengan boxplot untuk melihat adanya outlier pada data dengan metode perhitungan IQR. Data yang berada di luar batas 1.5 kali IQR dari Q1 dan Q3 dianggap outlier dan dihapus. Terdapat 3 kolom data yang memiliki outlier diantaranya adalah MonthlyIncome, DurationOfPitch, NumberOfTrips. Lalu kami melakukan drop data outlier seperti dibawah ini

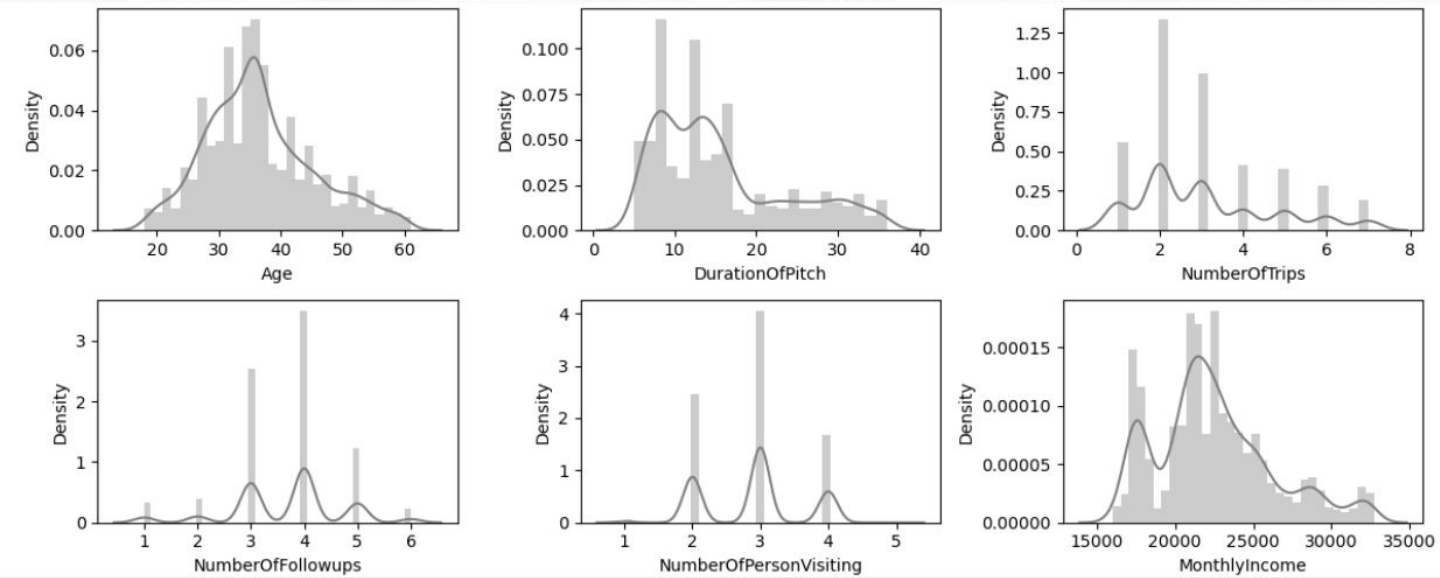
```
[ ]  df = clear_outlier(df, 'MonthlyIncome')
     df = clear_outlier(df, 'DurationOfPitch')
     df = clear_outlier(df, 'NumberOfTrips')
```

```
Number of rows dropped: 366
Number of rows dropped: 2
Number of rows dropped: 102
```

1. Data Cleansing (50 poin)

D. Feature Transformation

Kami melakukan visualisasi dengan sns.displot dengan hasil dibawah ini. Menurut hasil dari visualisasi tersebut data tergolong normal sehingga kami melakukan normalisasi data.



1. Data Cleansing (50 poin)

E. Feature Encoding
Pada tahap ini kami mendapati pada kolom marital status terdapat beberapa kategori seperti divorce, maried, single dan unmaried. Kami mengganti value dengan status Unmarried menjadi Single agar distribusi data menjadi lebih sederhana serta menghinari kerancuan data. Kemudian Pada columns *['TypeofContact', 'Occupation', 'Gender', 'MaritalStatus']* digunakan One-Hot Encoding dikarenakan data bersifat nominal Sedangkan pada columns *['ProductPitched', 'Designation']* digunakan Label Encoding karena data tersebut bersifat ordinal

1. Data Cleansing (50 poin)

F. Handle Class Imbalance
Pada handle class imbalance kami menggunakan metode smote yang dapat mempertahankan seluruh data mayoritas, sambil menghasilkan data tambahan untuk kelas minoritas, sehingga tidak ada informasi yang hilang.

SMOTE menghasilkan sampel yang lebih seimbang dan menyebar di ruang fitur, yang dapat membantu model mempelajari lebih banyak pola daripada sekadar fokus pada kelas mayoritas.

2. Feature Engineering (35 poin)

Cek feature yang ada sekarang, lalu lakukan:

A. Feature selection (membuang feature yang kurang relevan atau redundan)
Berdasarkan visualisasi sns.heatmap Terdapat dua feature antara 'ProductPitched' dan 'Designation' yang redundant, sehingga kami menghapus salah satu kolom tersebut.

B. Feature extraction (membuat feature baru dari feature yang sudah ada)
Kami memutuskan untuk Membuat klasifikasi umur untuk holiday package (<18 = young, 19-40 = adult, >40 = old)

C. Tuliskan minimal 4 feature tambahan (selain yang sudah tersedia di dataset) yang mungkin akan sangat membantu membuat performansi model semakin bagus (ini hanya ide saja, untuk menguji kreativitas teman-teman, tidak perlu benar-benar dicari datanya dan tidak perlu diimplementasikan)
-Menambahkan fitur HealthRate (mengukur tingkat kesehatan customer)
-Membuat klasifikasi kepuasan pelanggan
-Membuat klasifikasi berdasarkan pendapatan customer
-Membuat fitur easy to persuade (membandingkan antara numberOfPitch dengan ProdTaken, semakin kecil numberOfPitch dari pelanggan yang mengambil produk (ProdTaken = 1), maka semakin mudah dibujuk untuk membeli produk

3. Git (15 poin)

Upload project teman-teman di sebuah repository git. Berkolaborasilah di Git jika ada perubahan version dari waktu ke waktu.

A. Buat Repository Git
B. Upload file notebook atau file pengerjaan lainnya pada repository tersebut

Untuk file README, dapat merupakan summary dari proses data preproses yang telah dilakukan. Boleh menggunakan repositori yang sama atau membuat baru.

Link Github: https://github.com/reynaldi15/Rakamin-Analytic-Adventurers

Rakamin Academy

Selamat Mengerjakan!