Penambangan Data Genap 2023/2024

Assignment 06

Outlier Detection

Outlier detection adalah salah satu fungsi pada penambangan data untuk melihat nilai yang sifatnya jauh dari distribusi pada umumnya sebaran data. Deteksi outlier seringkali digunakan untuk melihat anomali pada data. Dalam mendeteksi outlier, terdapat berbagai metode, beberapa diantaranya adalah LOF (Local Outlier Factor), DBSCAN, dan Isolation Forest yang akan disimulasikan dalam tugas ini.

Untuk menjawab pertanyaan ini, silakan gunakan template Python yang telah disediakan. Untuk nomor 2 sampai 4, diperkenankan untuk menggunakan library scikit-learn dan dataset yang digunakan adalah dataset marketing_campaign yang dapat diunduh dari Scele. Dataset ini dapat diasumsikan bersih, sehingga tidak perlu melakukan imputasi atau transformasi sendiri, kecuali yang diminta oleh soal. Berikut adalah deskripsi data dari dataset marketing_campaign:

Sumber data: Kaggle (dengan modifikasi)

ID - nomor pelanggan
Year_birth - tahun lahir pelanggan
AcceptedCmp1 - nilai 1 jika pelanggan menerima penawaran pada penawaran pertama, nilai 0 kalau menolak
AcceptedCmp2 - nilai 1 jika pelanggan menerima penawaran pada penawaran kedua, nilai 0 kalau menolak
Response (target) - nilai 1 jika pelanggan menerima penawaran pada penawaran pada penawaran terakhir, nilai 0 kalau menolak
Complain - nilai 1 apabila pelanggan memiliki komplain dalam 2 tahun terakhir, nilai 0 jika tidak
DtCustomer - tanggal pendaftaran pelanggan dengan perusahaan
Education - tingkat pendidikan pelanggan
Marital - status pernikahan pelanggan
Kidhome - banyaknya anak-anak di rumah pelanggan
Teenhome - banyaknya remaja di rumah pelanggan
Income - pendapatan tahunan pelanggan
MntMeatProducts - banyaknya pengeluaran di produk daging dalam 2 tahun terakhir
MntFruits - banyaknya pengeluaran di produk buah dalam 2 tahun terakhir
NumStorePurchases - banyaknya pembelian di toko
NumWebPurchases - banyaknya pembelian melalui web
NumWebVisitsMonth - banyaknya kunjungan web pada bulan terakhir
Recency - jumlah atau selang hari dari pembelian terakhir
Response - nilai 1 apabila pelanggan menerima penawaran pada penawaran terakhir, nilai 0 kalau menolak

Pertanyaan:

1. Menggunakan Python, diketahui data sebagai berikut [1, 2, 5, 1, 3, 100, 3, 5, 40, 45, 279, 130, 71]. Tentukan nilai:
   a. Q1
   b. Median
   c. Q3
   d. Inter-quartile range (IQR)
   e. Angka outlier dari data yang diberikan

2. Lakukan proses reduksi dimensi dataset marketing campaign dengan menggunakan PCA hingga menjadi 2 dimensi dengan kolom yang telah dipilih. Data hasil reduksi ini akan digunakan untuk mengerjakan nomor 3 dan 4 juga. Kemudian, dengan menggunakan algoritma LOF (Local Outlier Factor) tentukan banyaknya customer yang outlier dan ID mereka. Kemudian buatlah scatter plot yang memvisualisasikan data outlier dan normal.

3. Dengan menggunakan algoritma DBSCAN tentukan banyaknya customer yang outlier dan ID mereka. Kemudian buatlah scatter plot yang memvisualisasikan data outlier dan normal.

4. Dengan menggunakan algoritma Isolation Forest tentukan banyaknya customer yang outlier dan ID mereka. Kemudian buatlah scatter plot yang memvisualisasikan data outlier dan normal.

5. Dari algoritma yang telah disimulasikan pada nomor 2 sampai 4, manakah ID pelanggan yang dianggap outlier oleh ketiga algoritma?

Dilarang keras menyontek. Plagiarisme tidak ditoleransi dan akan dikenai penalti atau nilai akhir E.

Pengurangan nilai akibat keterlambatan pengumpulan tugas akan ditentukan berdasarkan jumlah menit keterlambatan Anda dalam mengumpulkan. Misalnya, apabila terlambat 1 menit, nilai akhir akan dikurangi 1 poin, apabila terlambat 10 menit, nilai akhir akan dikurangi 10 poin, dan seterusnya.

Bobot penilaian:
Soal 1: 15 (masing-masing 3)
Soal 2: 25
Soal 3: 25
Soal 4: 25
Soal 5: 10

Pengumpulan tugas:
Kumpulkan berkas dengan format penamaan seperti berikut:

Assignment6_[NPM]_[NamaLengkap].ipynb

Contoh:
Assignment6_1906438834_TimothyOrvinEdwardo.ipynb