|  |  |  |
| --- | --- | --- |
| **Nama:**  **(Isi Nama Anda)**  **NIM:**  **(Isi NIM Anda)** | C:\Users\RPL-SI 02\Pictures\288px-Trisakti_Logo.svg.png | **MODUL 3**  **Nama Dosen:**  **( )** |
| **Hari/Tanggal:**  **Hari, Tanggal Bulan Tahun** | **Nama MK** | **Nama Asisten Labratorium: 1. ( )** |

## Teori Singkat

Encoding

Encoding adalah proses mengubah data dari satu format ke format lain yang lebih sesuai untuk penyimpanan atau transmisi. Dalam konteks komputer, encoding sering digunakan untuk mengonversi data, seperti teks, gambar, atau suara, menjadi bentuk digital yang dapat diproses oleh komputer.

Model machine learning dan deep learning memerlukan semua variabel input dan output dalam bentuk numerik. Ini berarti bahwa jika data Anda mengandung data kategorikal, Anda harus mengodekannya menjadi angka sebelum Anda dapat menyesuaikan dan mengevaluasi model. Ada beberapa teknik untuk mengodekan variabel kategorikal. Mari kita bahas ini:

* 1. One-Hot Encoding
  2. Label Encoding
  3. Target Encoding
  4. Target Guided Ordinal Encoding
  5. Custom Encoding

## Alat dan Bahan

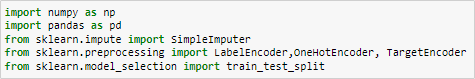
Hardware : Laptop/PC

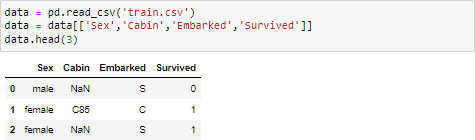
Software : Spyder (Anaconda Python)

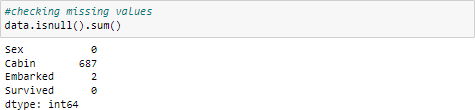
## Elemen Kompetensi

1. Latihan pertama

# Data Preparation

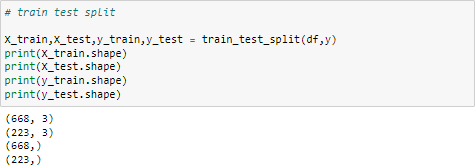
Import package

Read file

Memeriksa missing value pada data

Membuat df baru dan mengisi missing value

Memisahkan data train dan test

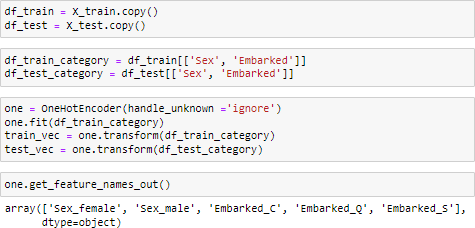


1. Latihan kedua

**Onehot Encoding**

Menggunakan sklearn Onehot Encoder

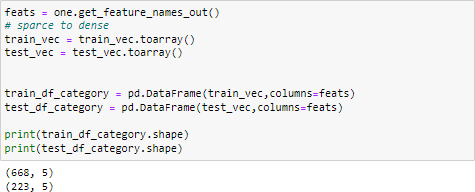
Langkah pertama adalah kita akan meng-copy dari train test split ke dataframe baru, sehingga hasil dari data pre-processing tidak akan mengganggu data orisinal



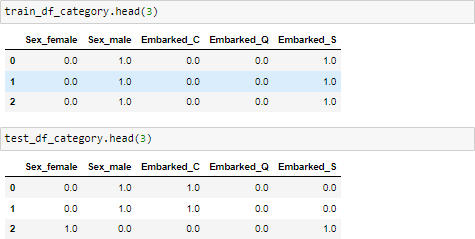
## Catatan:

1. handle\_unknown = 'ignore' sangat penting. Jika kategori yang tidak dikenal ditemukan selama transformasi, kolom yang dihasilkan untuk one-hot encoding dari fitur ini akan semuanya nol. Dalam kasus sebelumnya, kita menyelaraskan dan mengisi kategori yang tidak dikenal dari data uji dengan nol. Di sini, ini akan ditangani secara otomatis.
2. Hal penting lainnya adalah bahwa secara default, one hot encoder akan mengembalikan vektor jarang (sparse). Ini akan berguna untuk perhitungan yang lebih cepat saat menangani data berdimensi tinggi. Kita juga bisa mematikannya dengan

mengatur sparse = False jika kita ingin mengonversinya menjadi padat (dense) jika kita ingin melihat data frame.

Convert data yang sudah di proses to array lalu tampilkan

Menampilkan tiga data teratas



Catatan:

Contoh untuk mendapatkan vektor non-sparse (padat) secara langsung.

Tidak terdapat perbedaan pada hasil output, hanya cara penyimpanan nya yang berbeda, sehingga lebih hemat memori. Ini sangat disarankan pada kasus categorical yang sangat variatif. Contoh pada kasus warna baju T-Shirt, yang terdapat pilihan warna bisa lebih dari 10.

Contoh ilustrasi:

["merah", "biru", "hijau", "merah"] [[1, 0, 0], # merah

[0, 0, 1], # biru

[0, 1, 0], # hijau

[1, 0, 0]] # merah

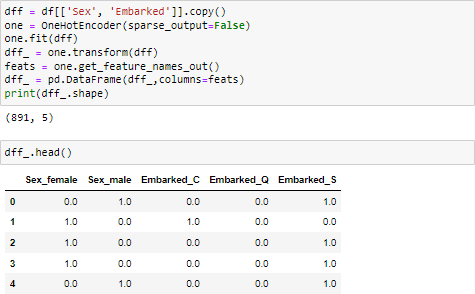
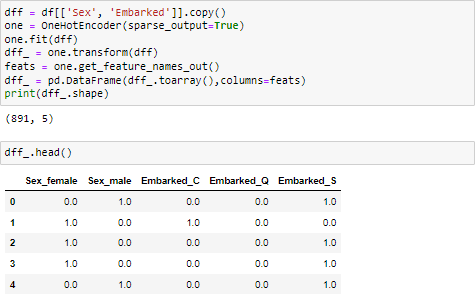
Cara ubah ke sparse adalah mengubah hasil encoding ke format ---> **(index\_data, posisi\_angka\_1\_dalam\_array)**

(0, 0) 1

(1, 2) 1

(2, 1) 1

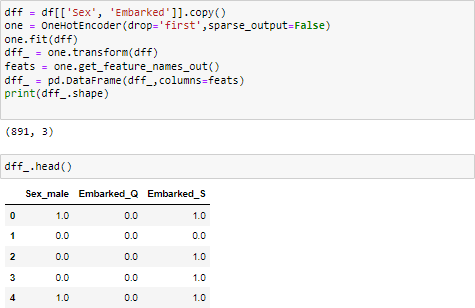
(3, 0) 1



Catatan:

Menghapus fitur one-hot pertama dari setiap kolom. Ini akan sangat membantu terutama dalam kasus model linear di mana multikolinearitas dapat mempengaruhi model.

**INGAT!!!** ---> Penanganan dummy variable pada one hot encoding : (n-1) kolom Dengan menambahkan drop='first', maka kolom pertama dari hasil encoding dari setiap variabel categorical akan dihapus

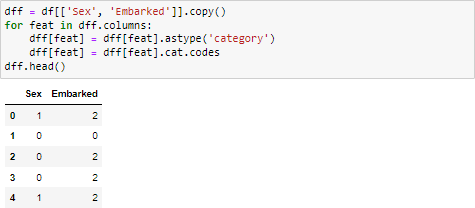


1. Latihan Ketiga

# Label Encoding

Catatan: Dalam label encoding, setiap label diubah menjadi nilai integer. Kita akan membuat variabel yang berisi kategori yang mewakili kualifikasi pendidikan seseorang.

## Menggunakan pendekatan kode kategori

Pendekatan ini memerlukan kolom kategori untuk memiliki tipe data 'category'. Secara default, kolom non-numerik adalah tipe 'object'. Jadi, Anda mungkin perlu mengubah tipe menjadi 'category' sebelum menggunakan pendekatan ini.

## using Sklearn Label Encoder

1. Latihan Keempat

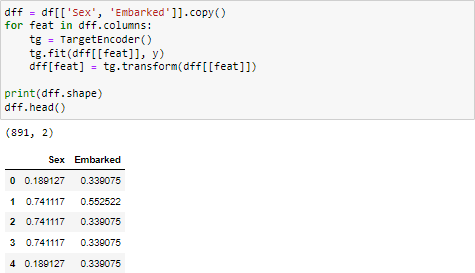
# Target Encoding

Berikut adalah teknik encoding yang merupakan teknik untuk mengubah nilai kategorial dengan nilai rata-rata variabel target yang berkaitan dengan variabel kategorial itu sendiri.

Target Encoding berbeda dengan teknik encoding lainnya, yakni memerlukan nilai

y atau target.

Oleh sebab itu, encoding ini hanya cocok pada kasus Supervised Learning



1. Latihan Kelima

# Target Guided Ordinal Encoding

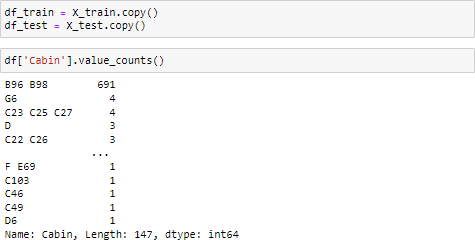
Sama dengann Target Encoding, hanya terdapat tambahan, yakni sistem ranking (ordinal encoding). Dimana dari hasil rata-rata, diurutkan dengan nilai tertinggi, kemudian dituliskan rankingnya



1. Latihan Keenam

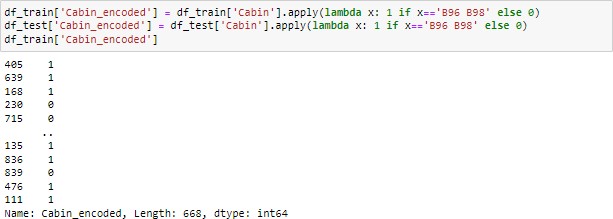
# Custom Encoding

Di sini, kita hanya mengubah kategori menjadi angka yang kita inginkan. Tentu kita harus memiliki rules atau aturan yang tetap untuk melakukan ini. Selain itu, juga ada basis alasan yang kuat akan penggunaan rules tersebut.



Dalam kasus Cabin, kita dapat melihat bahwa ada 147 nilai unik. Namun, dari nilai-nilai ini, cabin 'B96 B98' muncul sebanyak 691 kali, sementara yang lainnya muncul hanya dalam jumlah yang sangat sedikit. Jadi, misalkan kita ingin mengkodekan 'B96 B98' sebagai 1 dan sisanya sebagai 0. Jenis encoding ini disebut sebagai custom encoding.

Mari kita lihat bagaimana kita dapat melakukannya.



## Kesimpulan

a. Kita dapat mengetahui… (Tolong Isi lebih dari dua baris!)

## Cek List ()

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Elemen Kompetensi** | **Penyelesaian** | |
| **Selesai** | **Tidak Selesai** |
| **1.** | Latihan Pertama |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **2.** | Latihan Kedua |  |  |

1. **Formulir Umpan Balik**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Elemen Kompetensi** | **Waktu Pengerjaan** | **Kriteria** |
| **1.** | Latihan Pertama | … Menit | … |
| **2.** | Latihan Kedua | … Menit | … |

Keterangan:

* 1. Menarik
  2. Baik
  3. Cukup
  4. Kurang

Penanggung Jawab Praktikum Kepala Lab. Praktikum

( ) ( )