Introduction

Categorical variable hanya mengambil sejumlah nilai yang terbatas.

* Pertimbangkan survei yang menanyakan seberapa sering Anda sarapan dan memberikan empat pilihan: "Tidak pernah", "Jarang", "Hampir setiap hari", atau "Setiap hari". Dalam hal ini, data bersifat kategoris, karena tanggapan termasuk dalam kumpulan kategori yang tetap.
* Jika orang menanggapi survei tentang merek mobil apa yang mereka miliki, tanggapan tersebut akan masuk ke dalam kategori seperti "Honda", "Toyota", dan "Ford". Dalam hal ini, datanya juga kategoris.

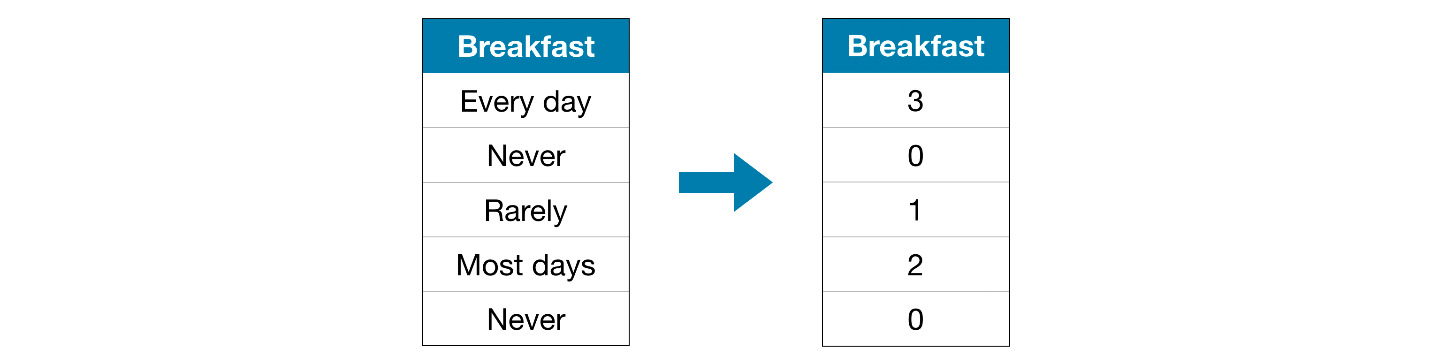
Anda akan mendapatkan kesalahan jika mencoba memasukkan variabel-variabel ini ke sebagian besar model pembelajaran mesin di Python tanpa memprosesnya terlebih dahulu. Dalam tutorial ini, kami akan membandingkan tiga pendekatan yang dapat Anda gunakan untuk menyiapkan data kategorikal Anda.

1.Drop categorical variable

Pendekatan termudah untuk menangani variabel kategoris adalah dengan menghapusnya dari dataset. Pendekatan ini hanya akan bekerja dengan baik jika kolom tidak berisi informasi yang berguna.

2.Ordinal encoding

Pengkodean ordinal memberikan setiap nilai unik ke bilangan bulat yang berbeda.

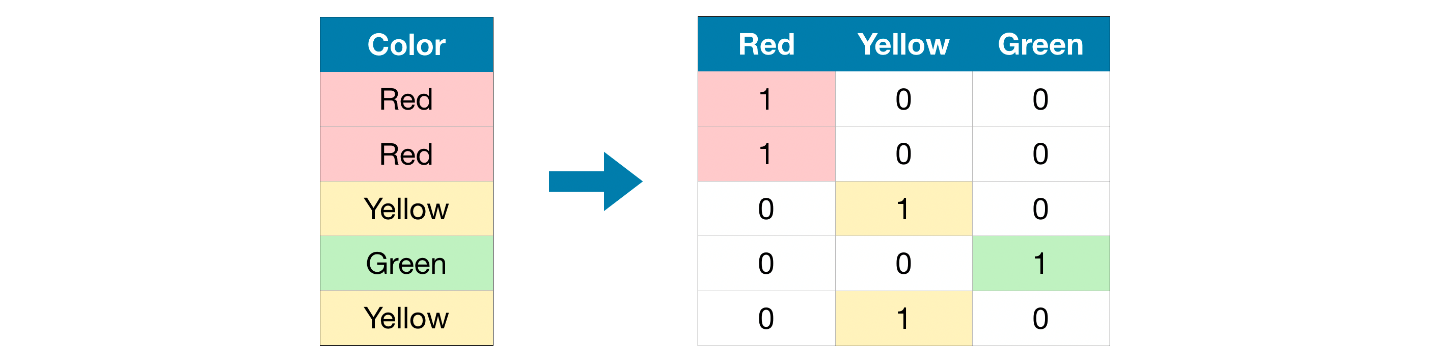


Pendekatan ini mengasumsikan urutan kategori: "Tidak pernah" (0) < "Jarang" (1) < "Kebanyakan hari" (2) < "Setiap hari" (3).

Asumsi ini masuk akal dalam contoh ini, karena ada peringkat tak terbantahkan untuk kategori. Tidak semua variabel kategori memiliki urutan nilai yang jelas, tetapi kami merujuk pada variabel yang melakukan sebagai variabel ordinal. Untuk model berbasis pohon (seperti pohon keputusan dan hutan acak), Anda dapat mengharapkan pengkodean ordinal bekerja dengan baik dengan variabel ordinal.

3.one-hot encoding

One-hot encoding membuat kolom baru yang menunjukkan ada (atau tidak adanya) setiap nilai yang mungkin dalam data asli. Untuk memahami ini, kita akan bekerja melalui sebuah contoh.

Dalam dataset asli, "Warna" adalah variabel kategori dengan tiga kategori: "Merah", "Kuning", dan "Hijau". Encoding one-hot yang sesuai berisi satu kolom untuk setiap nilai yang mungkin, dan satu baris untuk setiap baris dalam kumpulan data asli. Di mana pun nilai aslinya adalah "Merah", kami menempatkan 1 di kolom "Merah"; jika nilai aslinya adalah "Kuning", kami menempatkan 1 di kolom "Kuning", dan seterusnya.

Berbeda dengan pengkodean ordinal,one-hot encoding tidak mengasumsikan urutan kategori.

Dengan demikian, Anda dapat mengharapkan pendekatan ini bekerja dengan sangat baik jika tidak ada urutan yang jelas dalam data kategorikal (mis., "Merah" tidak lebih dan tidak kurang dari "Kuning"). Kami mengacu pada variabel kategoris tanpa peringkat intrinsik sebagai variabel nominal.

Encoding one-hot umumnya tidak bekerja dengan baik jika variabel kategori mengambil sejumlah besar nilai (yaitu, Anda biasanya tidak akan menggunakannya untuk variabel yang mengambil lebih dari 15 nilai yang berbeda).