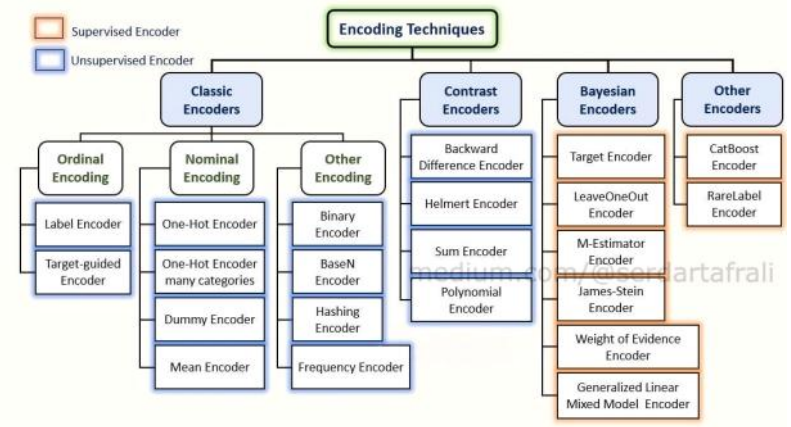
9. Ders - Missing Value - Part 1 - ENCODING

24 Eylül 2025 Çarşamba 22:33

Encoding, değişkenlerin temsil şekilleri ile ilgili değişiklikler yapmaktır. Yaygın kullanılan encoding yöntemlerine Label Encoding ve One Hot Encoding'i örnek verebiliriz. Gerekli görülmesi durumunda Rare Encoding gibi yöntemler de bu kapsamda kullanılabilir.



- **LABEL ENCODING**

- Bir kategorik değişkenin sınıfları labellardır. Label Encoding bu sınıfları modellerde daha kullanışlı ve modellerin anlayabileceği hale getirecek şekilde kodlama yöntemlerinden biridir.

- Eğer Label Encoding yapılacak olan bir kategorik değişkenin iki sınıfı var ise bu işleme özel olarak **Binary Encoding** adı verilir.

- Örnek :

SEX	IS_MALE
Male	1
Female	0
Male	1
Male	1
Female	0
Male	1
Female	0
Female	0

EDUCATION	LABEL_EDU
Pre-School	0
Secondary School	1
High School	2
Graduate	3
Master	4
PhD	5

(Label Encoding tekniğini nominal değişkenler özelinde kullanmak doğru bir yaklaşım olarak nitelendirilmeyebilir.)
((**Nominal Değişken** : Doğal bir sırası veya sıralaması olmayan kategorilere sahip bir değişken.))
-> Burada One-Hot Encoding başlığına bir geçiş yapıyoruz.

- **ONE-HOT ENCODING**

- One-hot encoding tekniğinde nominal değişkenin sınıfları değişkenlere dönüştürülür. Yani nominal değişken, her bir sınıftan ayrı bir sütun (değişken) oluşturulması yoluyla encode edilir.

TEAMS	LABEL_EDU
Detroit Pistons <3	0
Los Angeles Lakers	1
Miami Heat	2
Boston Celtics	3

Pistons	Lakers	Heat	Celtics
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

- One-hot encoding'in sınıflar adına oluşturduğu değişkenler **dummy** (kukla) değişkenlerdir.
- Dummy, eğer birbiri üzerinden oluşturulabilir haldeyse, bu durum ilgili değişkenler arasında yüksek bir korelasyona sebep olur. Bunun sonucunda ise ortaya bir ölçme problemi çıkmaktadır.

Bu sebeple Dummy değişken oluştururken **ilk sınıf drop edilirek** birbiri üzerinden oluşturulma durumu ortadan kaldırılmaya çalışılır.

- **RARE ENCODING**

- Nadir kategorileri gruplama yöntemidir.

- Düşük frekanstaki gözlemler için one-hot encode ile bir değer ataması yapılması uygun olmayacaktır.

Çünkü çok düşük frekanstaki her bir gözlem için yeni bir sütun oluşturulacak, ilgili sınıfa ait olmayan diğer gruplar için 0 değeri verilecektir.

Bu da çok sayıda 0'dan oluşan gereksiz sütunun oluşmasına neden olacaktır.

CAR_BRAND	CAR_BRAND_COUNT
Chevrolet	82
Audi	53
Rolls Royce	3
Renault	95
Bentley	2
Nissan	48
Bmw	32
Hyundai	64
Saab	9

CAR_BRAND	CAR_BRAND_COUNT
Chevrolet	82
Audi	53
Renault	95
Nissan	48
Bmw	32
Hyundai	64
Rare (Rolls Royce, Bentley, Saab)	14

- Rare Encoding kapsamında örnek bir kod çalışması :

```
#Nadir Marka Gruplama (RARE ENCODING & ETİKET KORUMA)
#rare encoding  - > diğer kategorisi ile gürültü yapılmasını engelliyoruz ender verilerin
min_shape = 0.02
brand_counts = df['Brand'].value_counts(normalize=True) #normalize burada oran olarak hesaplayacak
rare_brands = set(brand_counts < min_shape).index
df['_brand_compact '] = np.where(df['brand'].isin(rare_brands), 'OTHER', df['brand'])
```