BURSA TEKNİK ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ
3. SINIF

21 MART 2024

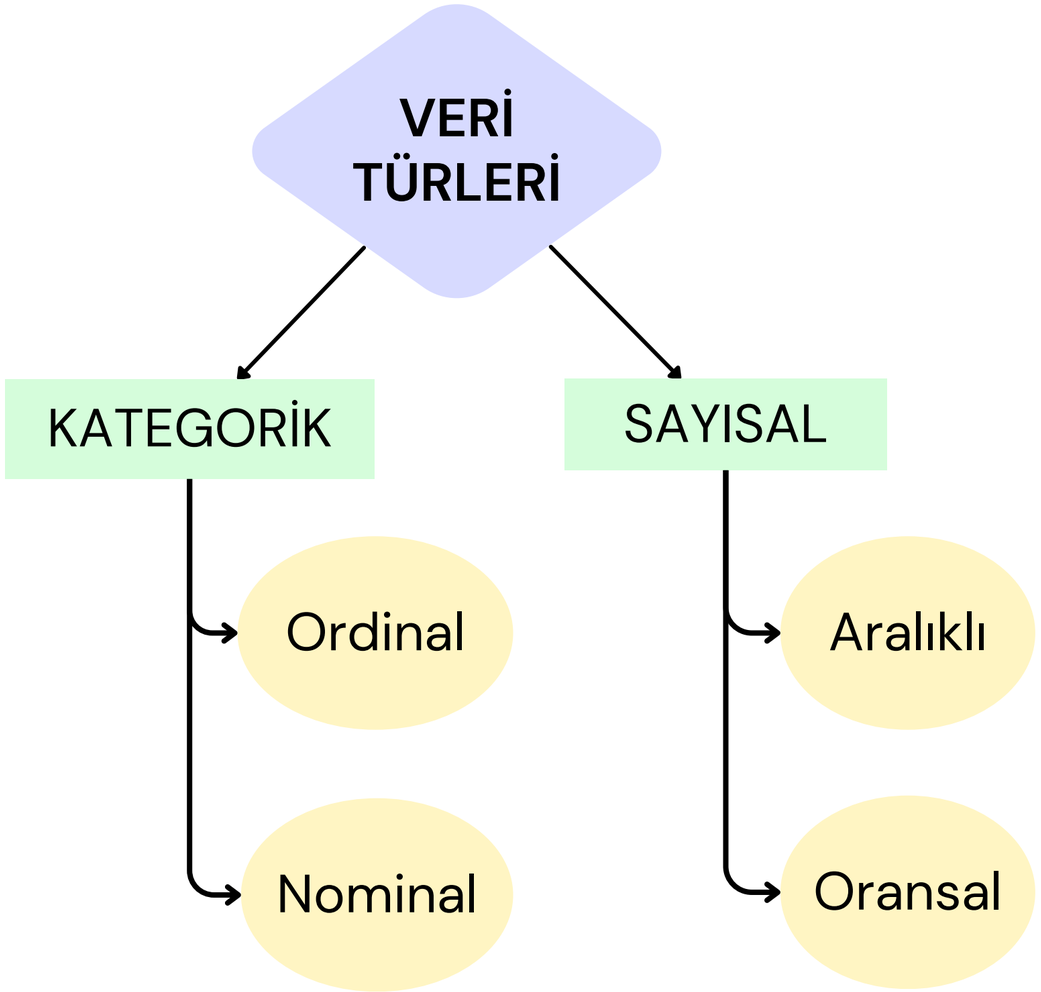
Kategorik Verilerin Ön İşlenmesinde Encoding İşlemi

Sümeyye Dural - 20360859065

İçindekiler

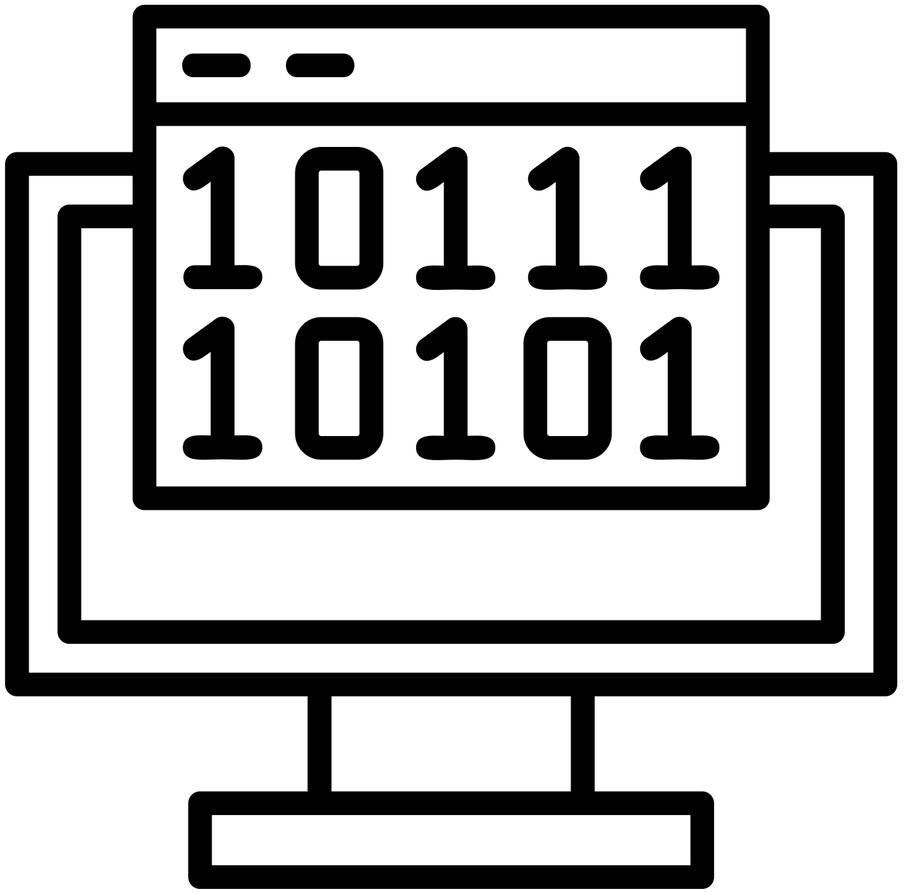
1) Kategorik Veri Nedir?

Ölçülemeyen, üzerinden sayısal işlem yapılamayan verilerdir. Nominal veriler birbirlerine üstünlüğü bulunmayan sınıflanmış verileri içerir. Ordinal veriler ise birbirleri arasında sıralanmış verileri içerir.



2) Encoding Nedir?

Veri bilimi ve makine öğrenimi süreçlerinde, veri setlerindeki kategorik değişkenleri makinelerin anlayabileceği bir formatta yeniden yazmaya denir. Bu değişkenler, algoritmalar tarafından işlenebilir bir hale getirmek için, sayısal temsili değerlere dönüştürülürler.



3) Sık Kullanılan Encoding Türleri

- Ordinal Encoding
- Label Encoding
- One-Hot Encoding
- Rare Encoding
- Target Encoding

4) Ordinal Encoding

Kategorik değişkenlerin sıralı bir şekilde sayısal bir formata dönüştürülmesini sağlar. Bu dönüşümde, kategorik değerler belirli bir sıralamaya göre sayılarla eşlenir.

EĞİTİM	Etiket
İLKOKUL	0
ORTAOKUL	1
LİSE	2
LİSANS	3
YÜKSEK_LİSANS	4
DOKTORA	5

4) Ordinal Encoding

```
from sklearn.preprocessing import OrdinalEncoder

ordinalEncoder = OrdinalEncoder(categories=sıralı_kategoriler)

df['Etiket'] = ordinalEncoder.fit_transform(df['EĞİTİM'])
```

4) Ordinal Encoding

AVANTAJLARI	DEZAVANTAJLARI
Doğrusal İlişkilerin Korunması	Sıralama Hataları
Bilgi Kaybının Azaltılması	Aykırı Değerlere Karşı Hassasiyet
Düşük Boyutluluk	Modelin Doğrusallık Varsayımı

5) Label Encoding

Her bir kategoriye benzersiz bir sayı atanır. Kategorik değişkenin iki sınıfı olması durumunda yapılacak işleme özel olarak binary encoding de denebilir.

CİNSİYET	Etiket
KADIN	1
ERKEK	0

5) Label Encoding

```
from sklearn.preprocessing import LabelEncoder

labelEncoder = LabelEncoder()

df['Etiket'] = labelEncoder.fit_transform(df['CİNSİYET'])
```

5) Label Encoding

AVANTAJLARI	DEZAVANTAJLARI
Basit ve Etkili	Anlamsal İlişkilerin Yanlış Yorumlanması
Düşük Boyutluluk	Model Yanıltılabilirliği
Genel Amaçlı Uygulanabilirlik	Kategorik Değerlerin Farklı Ağırlıklara Sahip Olması

5) Label Encoding

Doğru bir kullanım değildir!

Kategoriler arasında ordinal bir sıra belirtme durumu bulunmuyorsa, değişkenin sayısal bir büyüklük olarak tanımaması adına Label Encoding kullanılması uygun olmaz!

TAKIM	Etiket
Detroit Pistons	0
Chicago Bulls	1
Miami Heat	2
Brooklyn Nets	3

6) One-Hot Encoding

Nominal değişkenleri bir ölçüm problemine sebep olmadan oluşturmak adına one-hot encoding kullanmak gerekir. Bu teknikle nominal değişkenin sınıfların her biri yeni bir değişkenlere dönüştürülür.

TAKIM	Pistons	Bulls	Heat	Nets
Detroit Pistons	1	0	0	0
Chicago Bulls	0	1	0	0
Miami Heat	0	0	1	0
Brooklyn Nets	0	0	0	1

6) One-Hot Encoding

```
import pandas as pd

df = pd.get_dummies(df, columns=['TAKIM'])
```

6) One-Hot Encoding

AVANTAJLARI	DEZAVANTAJLARI
Basit ve Etkili	Yüksek Boyutluluk
Genel Amaçlı Uygulanabilirlik	Sparse (Seyrek) Matrisler
Bağımsızlık	Kategoriler Arasındaki İlişkilerin Yakalanmaması

6) One-Hot Encoding

One-hot encoding tekniği kullanılırken, oluşturulan değişkenler arasında çok yüksek ilişkinin olması durumuna çoklu doğrusal bağlılık (multicollinearity) denir. Bu durumda değişkenlerden birinin atılması uygun olacaktır. Bu değişkene dummy değişken denir.

TAKIM	Bulls	Heat	Nets
Detroit Pistons	0	0	0
Chicago Bulls	1	0	0
Miami Heat	0	1	0
Brooklyn Nets	0	0	1

6) One-Hot Encoding

```
import pandas as pd

df = pd.get_dummies(df, columns=['TEAMS'], drop_first=True)
```

7) Rare Encoding

Bir veri setindeki kategorik değişkenlerin bazı gözlem sınıflarının frekanslarının çok düşük olduğu durumda, Rare encoding, bu nadir kategorilere ayrı bir kategoride birleştirerek bu sorunu çözmeye çalışır.

MARKA	Toplam
Renault	96
Hyundai	64
Rolls Royce	9
BMW	32
Audi	53
Bentley	6

MARKA	Toplam
Renault	95
Hyundai	64
BMW	32
Audi	53
Diğer	15

7) Rare Encoding

```
import pandas as pd

counts = df['MARKA'].value_counts(normalize=True)
rare_categories = counts[counts < 10].index
df['MARKA'] = df['MARKA'].apply(lambda x: 'Diğer' if x in
rare_categories else x)
```

7) Rare Encoding

AVANTAJLARI	DEZAVANTAJLARI
Daha İyi Genelleme	Bilgi Kaybı
Daha Az Parametre	Genel Model Performansının Azalması
Daha Az Boyutluluk	Kategorik Değerlerin Anlamının Değişmesi

8) Target Encoding

Her bir kategori için, bir hedef değişken üzerindeki ortalama veya diğer bir istatistiksel özelliğinin kullanılmasıyla gerçekleştirilir. Hesaplanan değerler her bir kategori için kullanılan sayısal temsilciler olur.

ÜRÜN	Top_Alma	Top_Almama	Alma_Oranı
Elektronik	10	20	0,333
Giyim	0	30	0,0
Ev ve Bahçe	30	0	1,0
Spor	15	15	0,5

8) Target Encoding

```
import pandas as pd

target_means = df.groupby('ÜRÜN')['Satin_Alma'].mean()
df['Alma_Oranı'] = df['ÜRÜN'].map(target_means)
```

8) Target Encoding

AVANTAJLARI	DEZAVANTAJLARI
Bilgi Kullanımı	Veri Sızıntısı
Model Performansını Artırma	Aykırı Değerlere Karşı Hassasiyet
Daha Az Boyutluluk	Kategori Dengeleme Sorunu

9) Kaynakça

- https://towardsdatascience.com/encoding-categorical-variables-a-deep-dive-into-target-encoding-2862217c2753
- https://medium.com/datarunner/veri-%C3%B6n-i%CC%87%C5%9Fleme-kategorik-verilerin-say%C4%B1salla%C5%9Ft%C4%B1r%C4%B1lmas%C4%B1-6ba7e78c1be1
- https://yogeshchauhan09.medium.com/categorical-data-encoding-in-machine-learning-8c5e30a19585
- https://michael-fuchs-python.netlify.app/2019/06/16/types-of-encoder/
- https://medium.com/@hhuseyincosgun/%C3%B6zellik-m%C3%BChendisli%C4%9Fi-encoding-i%CC%87%C5%9Flemleri-8918f97bc8d8

9) Kaynakça

- https://miuul.com/not-defteri/encoding-nedir-turleri-nelerdir
- https://medium.com/swlh/an-introduction-to-categorical-feature-encoding-in-machine-learning-cd0ca08c8232
- https://scikit-learn.org/stable/modules/preprocessing.html
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
- https://towardsdatascience.com/one-hot-encoding-scikit-vs-pandas-2133775567b8
- ChatGPT 3.5

10) Soru – Cevap

11) Teşekkürler