Veri Önişleme – dvm.

Değişken dönüşümleri (Veri Dönüşümleri)
One-Hot dönüşümü
Veri Dönüşümleri

Değişken dönüşümü VS Değişken Standardizasyonu

Herbir standarditazyon işlemi bir değişken dönüşümüdür. (Dönüşüm>standadizasyon)

- Standardizasyonda, değişkenin içeridiği varyans yapısı, bilgi yapısı bozulmaz. Sadece bir standarta getirilir.
- Değişkenin içeriği değişir. Ancak veri'nin dağılım,yayılım bilgisinin özütü bozulmaz.
- Örneğin *x* değişkeni standardizasyon işlmeninden önce , veri sıralandığında 43. sırada ise, işlemden sonra elde edilelen *x_standart* değişkeni de 43. sırada yer almaya devam eder.
- Böylece yapılar halen birbiri ile kıyaslanabiliyor.

Değişken standardizasyon İşlemleri

- **Standardizasyon:**
  - from sklearn import preprocessing
  - preprocessing.scale(df)
- **Normalizasyon:**
  - preprocessing.normalize(df)
- **Min-Max Dönüşümü:**
  - scaler = preprocessing.MinMaxScaler(feature_range = (min_deger,max_deger))
  - scaler.fit_transform(df)
- **Bu fonksiyonların çıktısı numpy.array tipinde.**
- **df değişkeni korunuyor. Üzerine yazılmıyor.**

Değişken Dönüşümleri

- Her standadizasyon işlemi, bir veri dönüşümüdür.
- Bazı durumlarda, kategorik değişkenlerin değerlerinin numerik değişkenlere dönüştürülmesi gerekir.
- Örneğin bir değişken iki değer alabiliyor: Hasta ve saglıklı

  Fonksiyona verirken string bilgi kullanamıyoruz. O yüzden örneğin Hasta'yı 0 ve sağlıklıyı de 1 değerine dönüştürmemiz gerekir.
- Dönüşüm yaparken, kategorik değişkenin hangi ölçek türünde olduğuna dikkat etmemiz gerekir.

Ölçek (Scale) Türleri

- Sınıflandırma (adlandırma) ölçeği (nominal scale)
- Sıralama ölçeği (ordinal scale)
- Eşit aralıklar ölçeği (interval scale)
- Oran ölçeği (ratio scale)

Sınıflandırma ölçeği (nominal scale)

- Sınıflandırma ölçeği, araştırmacıya nesneleri belirli kategorilere ve gruplara atamasına izin veren ölçektir.
  - Sağlık durumuna göre araştırmaya katılanlar, sağlıklı ve hasta olarak sınıflandırılır.
  - Milliyet kategorileri: Türk, Alman, Rus vb.

Fabrikada imal edilen ve paketlenmek üzere yürüyen bir şerit üzerinde taşınan mamülleri hatalı veya hatasız diye nominal ölçekle ölçerek sınıflandırmak mümkündür.

Örnek

- Sınıflandırma ölçeği, ankete katılan kişinin **cinsiyeti, çalıştığı bölüm, uyruğu vb.**

Cinsiyet: Erkek__ Kadın__

Çalıştığı bölüm:

___ Üretim
___ Satış
___ Muhasebe
___ Finans
___ Personel
___ ARGE
___ Diğerleri

Sıralama ölçeği (ordinal scale)

- Nesnelerin veya alternatiflerin kendi grupları içinde sıralarını belirten ama aralarındaki farkın miktarını belirtmeyen ölçeklerdir.
- Verilen değerleri kademeli bir şekilde "azdan çoğa", "iyiden kötüye", "kısadan uzuna" doğru sıralanabilir.
- Belirli bir hastalığın en hafiften en şiddetliye doğru sıralanması da mümkündür.

Örnek

- Aşağıdaki bilgisayar sistemlerinin ofisinizde kullanım sıklıklarına göre sıralayınız. En çok kullanılana 1, en çok kullanılan 2. sisteme 2 vb. kullanarak sıralayınız. Kullanılmayan sisteme 0 giriniz.

__ Apple	__ HP
__ Compaq	__ IBM
__ Comp USA	__ Packard Bell
__ Dell	__ Sony
__ Gateway	__ Toshiba
__ Diğerleri	

Eşit Aralıklar Ölçeği (interval scale)

- Mutlak sıfırı olmayan **(arbitrary zero point)** ancak ölçülen değerler arasındaki mesafeler belli olan ölçeklerdir.
- Veri üzerinde belirli aritmetik işlemler yapmamıza izin verir.
- Belirli bir değişkenin ortalaması, standart sapması vb. hesaplanabilir.

Örnek

Bu ölçeğe en iyi verebilecek örnek Fahrenheit derece ve Celcius derece cinsinden sıcaklık ölçümleridir. Fahrenheit ve Celcius türü termometrelerde sıfır derece sıcaklığın olmadığı anlamına gelmez. Bununla birlikte sıfır derece kendisinden daha yüksek ve düşük sıcaklığın olduğunu belirtmektedir.

Oran Ölçeği (ratio scale)

- Bu ölçek sıfırı mutlak, değerlerin arası ise eşit ölçeklerdir.
- Voltaj, gelir (para birimi cinsinden), uzunluk, ağırlık, yaş vb. ölçmeler oranlı ölçeklerdir.
- Ölçülen değer yoksa sıfır elde edilir.
- ***Her türlü istatistik analizine uygun en gelişmiş ölçektir.***
- Oran ölçeği ilk üç ölçekten daha güçlü bir ölçektir.
- Daha çok fiziki bilimlerde kullanılır.

Oran Ölçeği (ratio scale)

- Oran ölçeği üzerinde ölçülmüş noktalar veya sayılar birbirinin katı olarak ifade edilebilirler.
- Bu ölçek üzerinde ölçülmüş verilere tüm aritmetik işlemler uygulanabilir. Ölçek üzerinde her nokta birbirinin katı olarak ifade edilebilir.
- **Örneğin,** "90 kg gelen bir insanın ağırlığı 45 kg gelen bir insanın ağırlığının 2 katıdır" denilebilir.

- **0-1 Dönüşümü:**
  - from sklearn.preprocessing import LabelEncoder
  - lbe = LabelEncoder()
  - lbe.fit_transform(df["gender"])
- **"1 ve Diğerleri (0) " Dönüşümü:**
  - df["day"].str.contains("Sun")
  - import numpy as np
  - df["yeni_day"] = np.where(df["day"].str.contains("Sun"), 1, 0)
- **Çok Sınıflı Dönüşüm:**
  - lbe.fit_transform(df["day"])

Neden One-Hot Encoding'e ihtiyaç duyulmuş?

Nominal ölçek türü ile etiketlenmiş kategorik değişkenlerin, birbirlerine karşı durumları arasında yorum yapamayız.

Örneğin çiçeğinin türlerini ele alalım.

İris setosa->0

İris Versicolar->1

İris Virginica->2 dönüşütü. Numeriğe dönüşünce sanki virginica, setosa arasında 2 birimlik bir fark vardır gibi duruyor. Bu algoritmaların yanılmasına sebep oluyor

ÇÖZÜM : ONE-HOT Encoding

One-Hot Encoding Dönüşümü ve Dummy Değişken

- Orijinal veri datasetinden çıkarılmalıdır.
- Kategorik değişkenin Sınıf sayısı bir azaltımalıdır.

]:

	total_bill	tip	gender	smoker	day	time	size	yeni_gender	yeni_day
0	16.99	1.01	Female	No	Sun	Dinner	2	0	1
1	10.34	1.66	Male	No	Sun	Dinner	3	1	1
2	21.01	3.50	Male	No	Sun	Dinner	3	1	1
3	23.68	3.31	Male	No	Sun	Dinner	2	1	1
4	24.59	3.61	Female	No	Sun	Dinner	4	0	1

]:

```
]: df_one_hot = pd.get_dummies(df, columns = ["gender"], prefix = ["gender"])
```

```
]: df_one_hot.head()
```

]:

	total_bill	tip	smoker	day	time	size	yeni_gender	yeni_day	gender_Male	gender_Female
0	16.99	1.01	No	Sun	Dinner	2	0	1	0	1
1	10.34	1.66	No	Sun	Dinner	3	1	1	1	0
2	21.01	3.50	No	Sun	Dinner	3	1	1	1	0
3	23.68	3.31	No	Sun	Dinner	2	1	1	1	0
4	24.59	3.61	No	Sun	Dinner	4	0	1	0	1

Sürekli Verileri Kategorik Değişkenlere Bölme

sklearn.preprocessing.KBinsDiscretizer

class sklearn.preprocessing.KBinsDiscretizer(n_bins=5, *, encode='onehot', strategy='quantile')