2.5 Veriyi Hazırlamak

Önemli bir hatırlatmada bulunalım, verilerin kopyaları üzerinde çalışmak, orijinal veri setini sağlam tutabilmek için önemlidir.

Veri setini makine öğrenmesine hazırlamak için aşağıdakileri yapabilirsiniz:

1. Veri Temizleme
    1. Aykırı değerleri düzeltme veya kaldırma
    2. Eksik değerleri düzenleme
2. Özellik Seçimi
    1. Alakasız özellikleri silme
3. Gerektiğinde, Özellik Mühendisliği:
    1. Sürekli özellikleri ayırıklaştırma
    2. Dönüşümler ekleme
    3. Özellikleri ayrıştırma
    4. Özellikleri birleştirme
4. Özellik Ölçeklendirme
    1. Özellikleri standartlaştırma
    2. Özellikleri normalleştirme

2.5.1 Veriyi Temizlemek

Çoğu Makine öğrenme algoritması eksik özellikler ile çalışamaz. Bu nedenle eksik veriler üzerinde çeşitli düzenlemeler yapmak gerekir.

1. Eksik özellikleri içeren tüm örnekler silinebilir.
2. Eksik özellikleri içeren tüm özellikler silinebilir.
3. Eksik veriler bir değere ayarlanabilir; sıfıra, özelliklerin ortalamasına, medyan değerine gibi.

Bizim örneğimizde hatırlarsanız hiç eksik veri bulunmuyordu, bu sebeple herhangi bir eksik veri düzenlemesi yapmamıza gerek yok.

2.5.2 Metin ve Kategorik Öznitelikleri İşlemek

Çoğu Makine öğrenme algoritması sayılarla çalışmayı tercih eder, bu yüzden bu kategorik verileri ve metinleri sayılara dönüştürmek gerekir.

1. Kategorik özellikler tam sayılara eşlenebilir veya kodlanabilir.
2. Metin içeren veriler için daha karmaşık işlemler bütünü gerekebilir, özellik mühendisliği burada devreye girer.

Örneğimizdeki kategorik verilere tekrar bakalım.

```
print veri_seti.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
total_bill    244 non-null float64
tip           244 non-null float64
sex           244 non-null category
smoker        244 non-null category
day           244 non-null category
time          244 non-null category
size          244 non-null int64
dtypes: category(4), float64(2), int64(1)
memory usage: 7.2 KB
None
```

Kategorik verilerimiz; sex, smoker, day, time. Problemlerimizi de hatırlayalım.

1. Ödenen hesaba göre verilen bahşişler nasıl değişir?
2. 26.5 lira hesap ödeyen bir kadın ne kadar bahşiş bırakır?

Sorunları yanıtlayabilmek için kullanacağımız özellikler:

1. Toplam Hesap
2. Bahşiş
3. Cinsiyet

Kategorik verileri sayısallaştırmadan önce kullanmayacağımız özellikleri kaldıralım. Veriyi temizleme bölümünde eksik veriler üzerinde çalışmıştık, kullanmayacağımız özelliklerin silinmesini de özellik seçimi altında yapabiliriz.

```
del veri_seti['size']
del veri_seti['time']
del veri_seti['day']
```

```
del veri_seti['smoker']


print veri_seti.head()
```

```
 total_bill   tip     sex
0      16.99  1.01  Female
1      10.34  1.66    Male
2      21.01  3.50    Male
3      23.68  3.31    Male
4      24.59  3.61  Female
```

Cinsiyet kategorik özelliğimizi sayılara eşlemeden önce cinsiyet özniteliğinin içerisinde kaç kategori olduğuna bakalım.

```
print veri_seti['sex'].unique()
```

```
[Female, Male]
Categories (2, object): [Female, Male]
```

Female ve Male olmak üzere iki kategori olduğunu görüyoruz. Şimdi bu kategorileri sayısal değerlere dönüştürelim.

```
from sklearn.preprocessing import LabelEncoder


label_encoder = LabelEncoder()
veri_seti['sex'] = label_encoder.fit_transform(veri_seti['sex'])


print veri_seti.head()
```

```
 total_bill   tip  sex
0      16.99  1.01    0
1      10.34  1.66    1
2      21.01  3.50    1
3      23.68  3.31    1
4      24.59  3.61    0
```

Gördüğünüz gibi cinsiyetleri 0 ve 1 olarak kodlandı. Bu kodlama işlemi, kategorileri alfabetik sıralayıp, sonra 0'dan başlayarak etiketliyor. Bu durumda Female=0, Male=1 oluyor.

2.5.3 Özellik Ölçeklendirmek

Verilerinizi makine öğrenme algoritmalarına hazırlamak için birçok yöntem var. En temel olanlarına örnekler verdik. Bunların dışında verilerinize başka birçok dönüşüm de uygulayabilirsiniz.

Verilerinize uygulamak zorunda olduğunuz en önemli dönüşümlerden biri özellik ölçekleme işlemidir. Birkaç istisna dışında, makine öğrenme algoritmaları, girdilerin sayısal değerleri çok farklı ölçeklere sahip olduğunda iyi performans göstermez.

Tüm özniteliklerin aynı ölçeğe sahip olmasının iki yaygın yolu vardır:

1. Normalleştirme
2. Standartlaştırma

Normalleştirme (Min-max ölçekleme) yaparken değerler 0'dan 1'e kadar değişene kadar kayar ve yeniden ölçeklendirilir.

Standartlaştırma oldukça farklıdır, önce ortalama değeri çıkarır ve daha sonra varyansa göre bölünür, sonuçta oluşan dağılımın birim varyansı olur, aykırı değerlerden çok daha az etkilenmektedir.

$$x' = \frac{x - \bar{x}}{x_{max} - x_{min}} \qquad x_{new} = \frac{x - \mu}{\sigma}$$

Örneğimize geri dönecek olursak şimdilik bir özellik ölçeklendirmesi yapmayacağız. Eğer veriniz üzerinde özellik ölçeklendirmesi yapmak konusunda karar veremediyseniz, özelliklerinizi ölçeklendirmeniz büyük olasılıkla bir sorun yaratmayacaktır.